Estimation de la taille de la population dans les expériences de capture-recapture

Thèse

Mamadou Yauck

Doctorat en statistique

Philosophiæ doctor (Ph. D.)

Québec, Canada

Résumé

La thèse présentée ici traite du problème de l'estimation de la taille de la population dans les modèles de capture-recapture. Elle s'intéresse, en particulier, à la question de l'estimation de la taille de la population dans le cadre d'une expérience de capture-recapture à structure d'échantillonnage imbriquée, qui combine les méthodes de population fermée à l'intérieur des périodes primaires (PP) et de population ouverte d'une PP à une autre: le design robuste. Cette thèse propose une méthodologie d'estimation de la taille de la population et de l'incertitude associée aux estimateurs obtenus dans le contexte du design robuste.

Dans un premier temps, on aborde le problème de l'estimation des paramètres du design robuste dans le cas d'un nombre suffisamment élevé d'occasions de capture. On généralise le papier fondamental de Jolly (1965) au design robuste en proposant une procédure séquentielle d'estimation des paramètres pour la classe des modèles de design robuste présentés dans Rivest and Daigle (2004) et un estimateur de la variance des paramètres par bootstrap paramétrique. Ces résultats théoriques ont été appliqués à des données d'activation d'applications sur les téléphones intelligents. Les données sont recueillies sur une période d'un an et démi et concernent des utilisateurs de téléphones intelligents qui ont visité un grand concessionnaire automobile basé aux États-Unis.

Dans un deuxième temps, on s'intéresse à l'estimation de la taille de la population à partir de deux sources d'information du design robuste: les données à l'intérieur d'une PP (ou intra-période) et les données d'une PP à une autre (ou inter-période). On démontre que les estimateurs de la taille de la population obtenus avec les informations intra-période et inter-période sont asymptotiquement indépendants pour une large classe de modèles de population fermée à l'intérieur des PP. Ainsi, l'estimateur du maximum de vraisemblance pour la taille de la population dans le cas du design robuste est asymptotiquement équivalent à un estimateur pondéré pour le modèle de population ouverte et le modèle de population fermée. On montre que l'estimateur pondéré diffère de celui donné dans Kendall *et al.* (1995); on démontre que leur estimateur n'est pas efficace, puis on donne une formule explicite pour son efficacité comparativement à l'estimateur pondéré. La perte d'efficacité est ensuite évaluée dans une étude de simulation, puis à travers un exemple tiré de Santostasi *et al.* (2016) et qui traite de l'estimation de la taille de la population d'une espèce de dauphins vivant dans le Golfe de

Corinthe (Grèce).

Enfin, on se propose d'étendre les résultats du problème précédent aux modèles de design robuste présentés dans Kendall *et al.* (1995) et implémentés dans MARK (White and Burnham, 1999). Dans le contexte du design robuste, on dérive l'estimateur du maximum de vraisemblance pour la taille de la population; on propose également trois méthodes d'estimation de la variance de l'erreur associée à l'estimateur. On démontre ensuite que l'estimateur du maximum de vraisemblance pour la taille de la population est plus efficace que l'estimateur des moments proposé par Kendall *et al.* (1995); la perte d'efficacité de l'estimateur de Kendall ainsi que la performance des trois méthodes d'estimation de la variance de l'erreur associée à l'estimateur du maximum de vraisemblance sont évaluées via une étude de simulation.

Abstract

This thesis deals with the capture-recapture estimation of population sizes under a hierarchical study design where a capture-recapture experiment, involving secondary capture occasions, is carried out within each sampling period (SP) of an open population model: the robust design. This thesis proposes a methodology for the estimation of population sizes under the robust design and the uncertainty associated with the estimators.

The first problem deals with the estimation of the parameters of a robust design with an arbitrary large number of capture occasions. To do so, we generalize the seminal paper of Jolly (1965) to the robust design and propose a sequential estimation procedure for the class of robust design models presented in Rivest and Daigle (2004). A simple parametric bootstrap variance estimator for the model parameters is also proposed. These results are used to analyze a data set about the mobile devices that visited the auto-dealerships of a major auto brand in a US metropolitan area over a period of one year and a half.

The second problem deals with the estimation of population sizes using two sources of information for the robust design: the within and the between primary period data. We prove that the population size estimators derived from the two sources are asymptotically independent for a large class of closed population models. In this context, the robust design maximum likelihood estimator of population size is shown to be asymptotically equivalent to a weighted sum of the estimators for the open population Jolly-Seber model (Jolly 1965; Seber 1965) and for the closed population model. This article shows that the weighted estimator is more efficient than the moment estimator of Kendall *et al.* (1995). A closed form expression for the efficiency associated with this estimator is given and the loss of precision is evaluated in a Monte Carlo study and in a numerical example about the estimation of the size of dolphin populations living in the Gulf of Corinth (Greece) and discussed by Santostasi *et al.* (2016).

The third problem deals with the estimation of population sizes under the robust design models presented in Kendall *et al.* (1995) and implemented in MARK (White and Burnham, 1999). We derive the maximum likelihood estimator for the population size and propose three methods of estimation for its uncertainty. We prove that the derived maximum likelihood estimator is more efficient than the moment estimator provided in Kendall *et al.* (1995). The loss of precision associated with the Kendall estimator and the perfomance of the three

methods of estimation for the variance of the maximum likelihood estimator are evaluated in a Monte Carlo study.

Table des matières

Résumé ii

Abstract iv

Table des matières vi

Liste des tableaux viii

Liste des figures x

Remerciements xv

Avant-propos xvi

Introduction 1

1 Les méthodes de capture-recapture 5
1.1 Les modèles de population fermée 5
1.2 Les modèles de population ouverte 18
1.3 Le design robuste 26

2 Capture-recapture Methods for Data on the Activation of Applications on Mobile Phones 30
Résumé 30
Abstract 31
2.1 Introduction 31
2.2 Capture-recapture modelling for activation data 33
2.3 Modelling of the robust-design data and parameter estimation 34
2.4 The robust-design with model M_0 within primary sessions 38
2.5 Model M_h within primary sessions 42
2.6 Case study: Clientele Estimation at Auto Dealerships 45
2.7 Discussion 48

3 On the Estimation of Population Sizes in Capture-recapture Experiments 50
Résumé 50
Abstract 51
3.1 Introduction 51
3.2 Model building 52
3.3 Estimating population sizes in capture-recapture experiments: A review . . 56

3.4 The estimation of population sizes in a robust design . . . 58
3.5 Comparisons between the maximum likelihood estimator and the moment estimator of Kendall et al. (1995) for N_i . . . 61
3.6 Numerical investigations . . . 61
3.7 Discussion . . . 66

4 Capture-recapture Estimation of Population Sizes Under the Robust Design **68**
Résumé . . . 68
Abstract . . . 69
4.1 Introduction . . . 69
4.2 Notation and assumptions . . . 70
4.3 Population size estimators for the robust design . . . 71
4.4 Numerical investigations . . . 75
4.5 Case Study . . . 79
4.6 Discussion . . . 80

Conclusion **81**

A Arguments techniques et matériel supplémentaire du chapitre 2 **83**
A.1 Arguments techniques . . . 83
A.2 Matériel supplémentaire . . . 95

B Arguments techniques et matériel supplémentaire du chapitre 3 **105**
B.1 Arguments techniques . . . 105
B.2 Matériel supplémentaire . . . 107

C Arguments techniques et matériel supplémentaire du chapitre 4 **110**
C.1 Derivation of Equation (4.7) . . . 110
C.2 Derivation of Equation (4.10) . . . 111
C.3 Equation (C.5) in the special case of model M_t^t with constant survival . . . 112
C.4 Equation (C.5) in the special case of model M_b^0 with constant survival . . . 113

D Aspects computationnels **114**
D.1 Programmes et données du chapitre 2 . . . 114
D.2 Données des chapitres 3 & 4 . . . 114

Bibliographie **115**

Liste des tableaux

1.1 Données issues d'une expérience de capture-recapture à deux occasions de capture. 10
1.2 Fréquences prédites des données de l'expérience de Lincoln-Petersen. 10
1.3 Matrice des statistiques sur la recapture des unités relâchées à chacune des $I-1$ périodes d'échantillonnage du modèle de Cormack-Jolly-Seber. 20
1.4 Fréquences observées et prédites pour un modèle de Cormack-Jolly-Seber avec 2 périodes de recapture et une survie dépendant du temps. 20
1.5 Fréquentes prédites pour un modèle de Jolly-Seber avec $I = 3$ périodes de capture et une survie dépendant du temps. 22

3.1 Simulation results for the estimation of N_2 under a stationary robust design model for M_t^t. 64
3.2 Simulation results for the estimation of $var(\hat{N}_2)$ under a stationary robust design model for M_t^t. 65
3.3 Abundance estimates for striped and common dolphins, under robust design model M_t^t. 66

4.1 Simulation results for the estimation of N_3 under the robust design model M_t^t. 77
4.2 Simulation results for the estimation of N_3 under the robust design model M_b^0. 77
4.3 Simulation results for the estimation of the MSE of $\hat{N}_3$ under the robust design model M_t^t. 78
4.4 Simulation results for the estimation of the MSE of $\hat{N}_3$ under the robust design model M_b^0. 79
4.5 Abundance estimates for striped and common dolphins, under robust design model $\{\phi(.)\, p(t.t)\}$. 80

A.1 Simulation results for the estimation of N_5 under the robust design model for M_{Dh}^t with $p^* = 0.3,\ 0.5$, $\phi = 0.6,\ 0.8$ and three scenarios for the entry process. All the results are presented in percentages 96
A.2 The bias, with its Monte Carlo standard error in parenthesis, the root mean squared error and the 95% coverage of the estimator of *inc*. The results are expressed in percentages. 97
A.3 Clientele size estimates and their coefficients of variation for the first 20 weeks of the experiment, under models M_h for a closed population, Jolly-Seber and M_{Dh}^t. 100
A.4 Survival probability estimates and their coefficients of variation for the first 20 weeks of the experiment, under Jolly-Seber and M_{Dh}^t models. 101

A.5 Capture probability estimates and their coefficients of variation for the first 20 weeks of the experiment, under models M_h for a closed population, Jolly-Seber and M^t_{Dh}. 102
A.6 Estimates of new arrivals and their coefficients of variation for the first 20 weeks of the experiment, under Jolly-Seber and M^t_{Dh} models. 103

Liste des figures

1.1 Représentation schématique d'une expérience de capture-recapture sur une population de poissons. Une marque noire (resp. rouge) indique une capture au jour 1 (resp. 2). Tiré de Rivest (2013). 9

1.2 La relation entre les modèles de population fermée M_{tbh}, M_{tb}, M_{th}, M_{hb}, M_h, M_b, M_t, M_0, présentés dans l'article de Otis *et al.* (1978). 17

1.3 La procédure d'échantillonnage décrivant un modèle de Cormack-Jolly-Seber avec une survie dépendant du temps. Les unités présentes dans la population au début de l'expérience (t_1) sont suivies jusqu'au temps de censure (t_I) correspondant au dernier temps d'observation. 19

1.4 La procédure d'échantillonnage décrivant le modèle de Jolly-Seber avec une survie dépendant du temps. Les unités présentes dans la population au début de l'expérience (t_1) sont suivies jusqu'au temps de censure (t_I) correspondant au dernier temps d'observation. 22

1.5 La procédure d'échantillonnage décrivant la paramétrisation de Schwarz and Arnason (1996) avec une survie dépendant du temps. Les unités présentes dans la population au début de l'expérience (t_1) sont suivies jusqu'au temps de censure (t_I) correspondant au dernier temps d'observation. 26

1.6 La procédure d'échantillonnage décrivant un modèle de design robust à I périodes primaires (PP) et ℓ_i $(i = 1, 2, \ldots, I)$ périodes secondaires (PS) d'échantillonnage. Les tailles de la population $\{N_i\}$ sont estimées avec l'information intra période primaire. Les estimations des probabilités de survie $\{\phi_i\}$ et des naissances $\{B_i\}$ sont obtenues avec l'information inter période primaire. 27

2.1 The sampling process describing the sequential procedure in which the Poisson random variables $\tilde{U}_i$ and $\tilde{M}_i$ $(i = 1, 2, ..., I)$ are simulated. 43

2.2 Differences between the estimated capture probabilities under the Jolly-Seber model and the following closed population models: M_0, M_h Chao (LB), M_h Darroch, M_h Gamma for positive values $a = 3.5, 4.5, 5.5, 6.5$, and M_h Poisson for $a = 2$. For each of the closed population models 76 capture probability estimates are obtained and their respective differences with the corresponding Jolly-Seber estimates are plotted. 46

2.3 Evolution of the estimated clientele size over the 76 weeks under models M_h for a closed population, Jolly-Seber and M_{Dh}^t, with a 95% confidence band for M_{Dh}^t. 47

3.1 The sampling scheme for the robust design with I sampling periods (SP) and ℓ_i $(i = 1, 2, \ldots, I)$ secondary occasions within each SP. 53

A.1 Data simulation process 96

A.2 Evolution of the detection probability for 76 weeks and for Metropolitan Area (MA) 1 and 2. 98
A.3 Evolution of the population size for 76 weeks and for three starting days (Sunday, Monday, Tuesday) for the PSPs. 99
A.4 Efficiency comparison (Squared Coefficient of variation) between the robust design estimate of N_i and those obtained under models M_h closed population and Jolly-Seber. The relative efficiencies are calculated with 1000 bootstrap samples and their values from the 76 PSP plotted. 104

À mon cher Sénégal, pays de la Téranga.

La recherche est le démiurge qui remodèle sans cesse la face du monde. L'histoire des techniques montre qu'à chaque découverte d'une nouvelle source d'énergie correspond un bond prodigieux de la civilisation matérielle. Du moulin à eau de l'antiquité, à la future fusée photonique de demain, la physionomie du monde est en perpétuelle transformation. Chaque nouvelle théorie modifie notre vision de l'Univers ; la mécanique classique avec NEWTON et LAPLACE aux XVIII^{e} et XIX^{e} siècles, la relativité générale au début du XX^{e} siècle avec EINSTEIN et la mécanique quantique, en plein développement depuis les travaux de Louis de BROGLIE, HEISENBERG et SCHRÖDINGER, nous ont présenté successivement un univers aussi aberrant pour le sens commun que fascinant pour l'intellect.

Cheikh Anta Diop, extrait du discours prononcé lors de la conférence d'ouverture de la 9e biennale de l'Association Scientifique ouest-africaine (ASOA, West African Science Association : WASA) tenue à la Faculté des Sciences de l'Université de Dakar du 27 mars au 1er avril 1974.

Remerciements

Je tiens à témoigner mon profond respect et ma reconnaissance à M. Louis-Paul Rivest pour sa disponibilité, son ouverture d'esprit, son écoute attentive et son soutien aussi bien financier que moral tout au long de cette aventure. Sa grande sagesse, sa générosité et son oeil critique ont permis d'améliorer la qualité de ce travail et d'assurer la réussite de tous les projets entrepris. Son sens de la pédagogie, son vaste savoir-faire et ses judicieuses orientations ont contribué à faire de cette thèse une expérience unique.

Je tiens également à remercier tous les membres du Département de Mathématiques et de Statistique de l'Université Laval pour tous les efforts consentis tout au long de ma formation. J'exprime ma gratitude à Lajmi, Frédéric, Emmanuelle, Gaétan et Hélène, dont la présence constante et l'aide m'ont permises de mener à bien les tâches d'enseignement et de consultation qui m'ont été confiées.

J'exprime ma reconnaissance au Département de Mathématiques et de Statistique de l'Université Laval, au Conseil de Recherches en Sciences Naturelles et en Génie du Canada et à l'Institut des Sciences Mathématiques du Québec pour le soutien matériel et financier sans lequel ce travail n'aurait pu être mené dans les conditions idoines.

Je tiens à exprimer ma reconnaissance et ma gratitude à toutes les personnes qui m'ont entourées, soutenues et encouragées durant cette expérience. Papa et maman, je vous remercie pour m'avoir inculqué la discipline, l'amour du travail et la culture de l'excellence. Votre sagesse, votre affection et votre rigueur ont fait de moi la personne que vous aimez si tendrement. Ma réussite est la vôtre. Je remercie mes soeurs Mbayang, Atta, Niassa, Bébé Fa, et mes frères Cheikh, Alsine, Mara pour leur présence, leur soutien inconditionnel et l'ambiance familiale qu'ils ont toujours cultivée. Je remercie Samba pour l'amitié grandissante et inconditionnelle qu'il a toujours soutenue et entretenue. Awa, ton amitié, ta sagesse et tes conseils ont largement contribué à forger ma personnalité. Alima, je te remercie pour ton amour sincère et ton soutien inconditionnel.

À ceux et celles qui ont contribué de près ou de loin à la réalisation de ce travail, cette thèse est la vôtre.

Avant-propos

La thèse de doctorat que nous présentons ici contient trois articles dont je suis le premier auteur. Le premier article, présenté au chapitre 2, s'intitule *Capture-recapture Methods for Data on the Activation of Applications on Mobile Phones*. Écrit en collaboration avec Louis-Paul Rivest, mon directeur de recherche, et Greg Rothman, professionnel de recherche dans une agence de marketing basée aux États-Unis, cet article a été publié en ligne dans le *Journal of the American Statistical Association* le 3 mai 2018. J'ai participé à l'élaboration de l'idée et proposé une première version du manuscrit qui a été progressivement et significativement améliorée par Louis-Paul Rivest. J'ai également effectué la majeure partie des calculs théoriques et toute la programmation informatique. Le deuxième article, présenté au chapitre 3, s'intitule *On the Estimation of Population Sizes in Capture-recapture Experiments*. Cet article, écrit en collaboration avec Louis-Paul Rivest, est soumis au *Journal of Multivariate Analysis* le 17 novembre 2018. Pour celui-ci, j'ai participé à l'élaboration de l'idée, que j'ai ensuite validée à travers les calculs théoriques et computationnels. J'ai proposé une première version du manuscrit qui a été améliorée et validée par Louis-Paul Rivest. Le troisième article, présenté au chapitre 4, s'intitule *Capture-recapture Estimation of Population Sizes Under the Robust Design*. J'ai proposé l'idée de cet article, écrit la première version du manuscrit et effectué tous les calculs théoriques ainsi que la programmation. Cet article, soumis au journal *Journal of Agricultural, Biological and Environmental Statistics* le 12 décembre 2018, est écrit en collaboration avec Louis-Paul Rivest qui a revisé le travail.

Introduction

Le problème de l'estimation des paramètres démographiques de populations d'animaux, tels que la taille de la population, la survie, l'émigration et l'immigration, remonte à 1894 avec les travaux de Petersen (Lecren, 1965) sur des populations de poissons. Depuis, les applications se sont multipliées dans des domaines tels que la biologie (Seber 1982; Borchers *et al.* 2002), la démographie et les sciences sociales (Sekar and Deming 1949; IWGDMF (International Working Group for Disease Monitoring and Forecasting) 1995a,b), les sciences médicales (Hook and Regal 1993; Bishop *et al.* 1975) et, récemment, la technologie du téléphone mobile.

Les méthodes de capture-recapture sont utilisées dans les problèmes d'estimation des paramètres démographiques de populations (souvent) difficiles à rejoindre. Habituellement, on organise une expérience de capture-recapture qui consiste à un échantillonnage répété des unités qui vivent dans un territoire bien délimité. L'expérience peut être menée sur plusieurs jours ou semaines consécutifs appelés occasions de capture. Au tout début de l'expérience, on capture un premier échantillon d'unités dans la population ; on pose une marque sur chaque unité capturée, puis on les remet dans la population. À la deuxième occasion, et 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 pose une marque sur celles qui sont vues pour la première fois avant de les remettre dans la population. Au terme de l'expérience, conduite sur un nombre fini d'occasions de capture, on dispose de l'historique des unités capturées au moins une fois : il consiste en une séquence de 0 (manqué) et 1 (marqué) résumée dans un tableau des historiques de capture individuels. Les données de capture-recapture représentent les fréquences empiriques de l'ensemble des historiques de capture observables. Le but de l'expérience est d'estimer la taille de la population, la survie et la migration entre autres paramètres en ajustant des modèles aux données de capture-recapture.

Les modèles de capture-recapture peuvent être classés 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. Le modèle le plus simple possède deux occasions de capture supposées indépendantes. L'estimateur de la taille de la population qui en résulte est communément appelé estimateur de Petersen-Lincoln, du nom des auteurs Petersen (Lecren, 1965) et Lincoln (Lincoln, 1930) qui

ont introduit pour la première fois cette méthodologie en écologie. Un exemple sur le calcul de l'estimateur de la taille d'une population de poissons est présenté au chapitre 1. L'estimateur de Lincoln-Petersen ne présente pas des propriétés asymptotiques satisfaisantes, notamment en termes de biais, si le nombre d'unités capturées à nouveau à la seconde occasion est faible ; voir Chapman (1951) pour un estimateur corrigé et Seber (1982) pour une présentation succincte sur la procédure d'estimation et les propriétés asymptotiques de l'estimateur.

L'expérience de capture-recapture la plus générale, souvent appelée "Schnabel Census" (Schnabel, 1938), est conduite sur trois ou plusieurs occasions durant lesquelles les unités sont captuées, marquées puis relâchées. Pour traiter des données de capture-recapture issues de l'expérience de Schnabel, Otis *et al.* (1978) ont introduit un ensemble de modèles de population fermée qui supposent trois types de variation dans la capture des unités : (1) variation temporelle (ou entre les occasions de capture), (2) variation comportementale après la première capture et (3) variation hétérogène d'une unité à l'autre. Otis *et al.* (1978) ont proposé des méthodes d'estimation des paramètres démographiques par maximum de vraisemblance, qui s'appuient sur l'hypothèse d'une distribution multinomiale des fréquences empiriques associées aux historiques de capture. Une alternative log-linéaire, décrite dans Bishop *et al.* (1975) et développée dans Cormack (1989), suppose que les fréquences empiriques des historiques observables suivent des lois de Poisson indépendantes. Fewster and Jupp (2009) ont montré que l'estimateur du maximum de vraisemblance Poisson pour la taille de la population admet les mêmes propriétés asymptotiques que l'estimateur obtenue avec l'approche multinomiale. Par ailleurs, Sandland and Cormack (1984) ont montré que la variance multinomiale asymptotique de l'estimateur de la taille de la population est égale à la variance Poisson diminuée de la valeur de l'estimateur.

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 (PC). À chaque PC, les unités sont capturées, marquées et relâchées sur une seule occasion de capture. On considère généralement que les PC sont assez espacées dans le temps, rendant plausible l'hypothèse d'ouverture de la population. Le modèle le plus adapté à ce scénario est le modèle de Jolly-Seber (Jolly 1965; Seber 1965). Jolly (1965) et Seber (1965) ont proposé une méthode d'estimation par maximum de vraisemblance en supposant que le nombre d'unités marquées avant chaque PC est un paramètre fixe ; ils ont déduit des estimateurs du maximum de vraisemblance pour les paramètres démographiques ainsi que des formules explicites pour leurs variances asymptotiques. Cormack (1989) et Schwarz and Arnason (1996) ont proposé des méthodes d'estimation par maximum de vraisemblance pour les approches Poisson et multinomiale respectivement.

Dans le modèle de population ouverte de Jolly-Seber, certains paramètres associés à la première et à la dernière PC ne sont pas estimables (Pollock *et al.*, 1990). Par ailleurs, s'il existe une hétérogénéité persistante dans la capture des unités, celles qui sont marquées auront

tendance à avoir des probabilités de capture plus élevés que celles qui ne sont pas marquées. De ce fait, la proportion des unités marquées dans l'échantillon aura tendance à surestimer la vraie proportion des unités marquées dans la population, entrainant une sous-estimation de la taille de la population. Le même phénomène de surestimation (sous-estimation) de la taille de la population peut être observé lorsque les unités sont moins sujettes (plus sujettes) à une capture subséquente après le premier marquage, voir Pollock (1975) et Nichols *et al.* (1984). Pollock (1982) a introduit une approche d'échantillonnage qui permet d'obtenir des estimateurs "robustes" à l'hétérogénéité et au changement comportemental des unités : le design robuste. Il s'agit d'une expérience de capture-recapture constituée d'un nombre fini de périodes de capture primaires (PP), qui peuvent être des mois ou des années. À l'intérieur de chaque PP, des sessions de capture sont menées sur un nombre fini de périodes secondaires (PS), par exemple des journées ou des semaines. On suppose que les PS sont assez rapprochées pour que l'hypothèse de fermeture de la population soit plausible. On suppose que d'une PP à l'autre, la population est sujette à des entrées et à des sorties d'unités, rendant plausible l'hypothèse d'ouverture. Pollock (1982) a proposé d'appliquer les modèles de population fermée présentés dans Otis *et al.* (1978) aux données des PS (ou information intra-période) pour estimer la taille de la population à chaque PP. Le modèle de Jolly-Seber est ensuite appliqué aux données des PP (ou information inter-période) pour estimer la survie.

Kendall *et al.* (1995) ont proposé une approche d'estimation des paramètres du design robust par maximum de vraisemblance. La fonction vraisemblance qu'ils ont construite a principalement deux composantes : la vraisemblance du modèle de Jolly-Seber pour l'information inter-période et la vraisemblance du modèle de population fermée pour l'information intra-période. Kendall *et al.* (1995) ont obtenu les estimateurs du maximum de vraisemblance pour les probabilités de capture et de survie. Pour la taille de la population, un estimateur des moments a été proposé du fait de la non apparition de ce paramètre dans la vraisemblance. Par ailleurs, les modèles de population fermée qui prennent en compte l'hétérogénéité dans la capture des unités n'étaient pas considérées dans leur article du fait de la difficulté à obtenir des estimateurs du maximum de vraisemblance. Rivest and Daigle (2004) ont proposé une approche d'estimation des paramètres du design robust dans un cadre log-linéaire. Les modèles présentés dans leur article permettent de modéliser les trois types de variation dans la capture des unités (Otis *et al.*, 1978) à l'intérieur des PP, puis d'obtenir des estimateurs du maximum de vraisemblance pour les probabilités de survie, les probabilités de capture et la taille de la population. Cependant, leur méthode ne marche pas pour des problèmes mettant en jeu plus de 20 occasions de capture : le vecteur dépendant ainsi que la matrice de design associée au modèle linéaire généralisé dépassent la mémoire d'exécution dans les logiciels standard. Cette thèse traite, entre autres problèmes, de l'estimation de la taille de la population dans le contexte du design robuste et en présence d'un nombre suffisamment élevé d'occasions de capture.

Les objectifs de cette thèse sont :

(i) Développer des procédures d'estimation des paramètres du design robuste, en particulier la taille de la population, lorsque le modèle contient un nombre suffisamment élevé d'occasions de capture ;

(ii) Évaluer l'incertitude associée à l'estimateur de la taille de la population ;

(iii) Étudier les propriétés asymptotiques de l'estimateur de la taille de la population.

Cette thèse est organisée comme suit. Le chapitre 1 traite de la théorie sur les modèles de capture-recapture ; un accent particulier est mis sur l'estimation de la taille de la population dans les modèles de population fermée et les modèles de population ouverte. Les procédures d'estimation des paramètres associés aux modèles étudiés, le calcul de l'incertitude associée aux estimateurs ainsi que leurs propriétés asymptotiques sont étudiées. Dans le chapitre 2, on s'intéresse à l'estimation des paramètres du design robuste dans le cas d'un nombre suffisamment élevé d'occasions de capture. On généralise le papier fondamental de Jolly (1965) au design robuste (1) en proposant une procédure séquentielle d'estimation des paramètres pour la classe des modèles de design robuste présentés dans Rivest and Daigle (2004) et (2) en proposant un estimateur de la variance des paramètres par bootstrap paramétrique. Le chapitre 3 traite de l'estimation de la taille de la populationà partir de deux sources d'information du design robuste : les données inter-période et intra-période. On montre que les estimateurs de la taille de la population obtenus avec ces deux sources d'information sont asymptotiquement indépendants pour une large classe de modèles de population fermée à l'intérieur des PP. Ainsi, l'estimateur du maximum de vraisemblance pour la taille de la population dans le cas du design robuste est asymptotiquement équivalent à un estimateur pondéré pour le modèle de Jolly-Seber et le modèle de population fermée. Dans le chapitre 4, on se propose d'étendre les résultats du chapitre 3 aux modèles de design robuste présentés dans Kendall *et al.* (1995) et implémentés dans MARK (White and Burnham, 1999). Dans le contexte du design robuste, on dérive l'estimateur du maximum de vraisemblance pour la taille de la population ; on propose également trois méthodes d'estimation de la variance de l'erreur associée à l'estimateur. Les détails techniques ainsi que le matériel supplémentaire en lien avec les chapitres 2, 3 et 4 sont présentés dans les annexes A, B et C respectivement. Les aspects computationnels liés aux trois chapitres de cette thèse, incluant les programmes ayant servi à construire les figures, à mener les études de simulation et à analyser les données, sont fournis dans l'annexe D.

Chapitre 1

Les méthodes de capture-recapture

Ce chapitre traite de la théorie des méthodes de capture-recapture. Il s'intéresse, en particulier, à l'estimation de la taille de la population dans les modèles de populations fermée et ouverte.

1.1 Les modèles de population fermée

Cette section traite des modèles de capture-recapture dans le cas d'une population fermée : spécification du modèle et hypothèses, procédure d'estimation des paramètres, calcul de la variance de l'estimateur de la taille de la population.

1.1.1 Données et notation

On considère une expérience de capture-recapture conduite sur une population fermée constituée de N unités. L'historique de capture d'une unité i est une séquence de 0 (manqué) et de 1 (capturé) résumé dans le vecteur $\omega_i = (\omega_{i1}, \omega_{i2}, \ldots, \omega_{i\ell})$, où ℓ est le nombre d'occasions de capture. Le nombre d'unités distinctes capturées au moins une fois au cours de l'expérience est $n = \sum_{\omega=1}^{s} n_\omega$, où $s = 2^\ell - 1$ est le nombre d'historiques de capture observables. Les historiques de capture individuels sont souvent représentés dans un tableau $n \times \ell$:

$$H = \begin{pmatrix} \omega_{11} & \omega_{12} & \ldots & \omega_{1\ell} \\ \omega_{21} & \omega_{22} & \ldots & \omega_{2\ell} \\ \ldots & \ldots & \ldots & \ldots \\ \omega_{n1} & \omega_{n2} & \ldots & \omega_{n\ell} \end{pmatrix}, \tag{1.1}$$

À partir de $H = (\omega_{ij})_{i,j}$, on définit les statistiques suivantes :

- $u_j = \sum_{i=1}^{n} \prod_{k=1}^{j-1} (1 - \omega_{ik})\omega_{ij}$ 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 total d'unités capturées à l'occasion j ;
- f_j est le nombre d'unités capturés exactement j fois, $j = 1, 2, \ldots, \ell$.

On définit Ω comme étant l'ensemble des historiques de capture observables. Dans la suite de ce chapitre, on fera usage de la représentation des données par les fréquences $\{n_\omega\}$ associées aux historiques de capture observables $\omega \in \Omega$. On désigne respectivement par μ_ω et $P_\omega(\boldsymbol{\theta})$ la fréquence prédite et la probabilité de réalisation de l'historique ω, où $\boldsymbol{\theta}$ est un vecteur de paramètres de dimension au plus s ; $P_0(\boldsymbol{\theta}) = 1 - \sum_\omega P_\omega(\boldsymbol{\theta})$ représente la fréquence du seul historique non observable $\omega = (0, 0, \ldots, 0)$. On écrit alors

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

où $\Omega^0 = \Omega \bigcup (0, 0, \ldots, 0)$. On émet les deux hypothèses suivantes :

(i) La probabilité qu'une unité choisie aléatoirement dans la population ait l'historique ω, $P_\omega(\boldsymbol{\theta})$, est la même pour chaque unité ;

(ii) Les unités agissent de façon indépendante.

Partant de ces hypothèses, on présente la procédure d'estimation des paramètres selon trois approches.

1.1.2 Procédure d'estimation : approche multinomiale

Des deux hypothèses émises dans la section précédente, on déduit que les s variables aléatoires $(N - n, \{n_\omega\})$ sont indépendantes et suivent une loi multinomiale (Darroch, 1958) :

$$(N - n, \{n_\omega\}) \sim M(N; P_0(\boldsymbol{\theta}), \{P_\omega(\boldsymbol{\theta})\}). \tag{1.2}$$

La fonction de vraisemblance s'écrit alors

$$L_M(N, \boldsymbol{\theta}\}) = \frac{N!}{\prod_{\omega\in\Omega} n_\omega!\,(N-n)!} P_0(\boldsymbol{\theta})^{N-n} \prod_{\omega\in\Omega} P_\omega(\boldsymbol{\theta})^{n_\omega}. \tag{1.3}$$

L'approche de maximisation simultanée de (1.3) est dite non conditionnelle (ou "Unconditional") ; on dénote par $\hat{N}_U$ et $\hat{\boldsymbol{\theta}_U}$ les estimateurs du maximum de vraisemblance de N et $\boldsymbol{\theta}$ obtenus. Une approche de maximisation dite conditionnelle, basée sur la décomposition de la vraisemblance (1.3) est présentée à la section suivante.

1.1.3 Procédure d'estimation : approche conditionnelle

La vraisemblance multinomiale donnée en (1.3) peut être réécrite comme le produit de deux termes :

$$L_M(N; \boldsymbol{\theta}) = L_1(N; P_0(\boldsymbol{\theta})) L_2(\boldsymbol{\theta}), \tag{1.4}$$

avec

$$L_1(N; P_0(\boldsymbol{\theta})) = \frac{N!}{n!(N-n)!} \{P_0(\boldsymbol{\theta})\}^{N-n} \{1 - P_0(\boldsymbol{\theta})\}^n \tag{1.5}$$

et

$$L_2(\boldsymbol{\theta}) = \frac{n!}{\prod_{\omega\in\Omega} n_\omega!} \{Q_\omega(\boldsymbol{\theta})\}^{n_\omega}, \tag{1.6}$$

où

$$Q_\omega(\boldsymbol{\theta}) = \frac{P_\omega(\boldsymbol{\theta})}{1 - P_0(\boldsymbol{\theta})}$$

est la probabilité qu'une unité ait l'historique de capture ω sachant qu'elle a été capturée au moins une fois au cours de l'expérience. La quantité L_1 est une fonction de vraisemblance binomiale basée sur la probabilité de n, alors que L_2 représente la vraisemblance multinomiale du vecteur $(n_1, n_2, \ldots, n_{\ell-1})$ contionnelle au nombre total de captures n.

La fonction de vraisemblance (1.3) peut être maximisée selon une approche dite conditionnelle, faite en deux étapes. On estime d'abord $\boldsymbol{\theta}$ en se servant de la vraisemblance conditionnelle L_2 donnée en (1.6) ; on obtient l'estimateur du maximum de vraisemblance de $\boldsymbol{\theta}$, $\hat{\boldsymbol{\theta}}_{\boldsymbol{C}}$. On estime ensuite N en maximisant la vraisemblance $L_1(N; P_0(\hat{\boldsymbol{\theta}}_{\boldsymbol{C}}))$; on obtient alors l'estimateur

$$\hat{N}_C = \left[\frac{n}{1 - P_0(\hat{\boldsymbol{\theta}}_{\boldsymbol{C}})} \right], \tag{1.7}$$

où [ ] est la fonction partie entière. Les estimateurs de N obtenus à partir des approches conditionnelle et non conditionnelle ne sont généralement pas égaux. Sanathanan (1972) a montré que

$$\hat{N}_U \leq \hat{N}_C, \tag{1.8}$$

puis Cormack and Jupp (1991) ont montré que cette différence est d'ordre 1. Sanathanan (1972) a montré par ailleurs que, sous certaines conditions de régularité, $(\hat{N}_C, \hat{\boldsymbol{\theta}}_{\boldsymbol{C}})$ et $((\hat{N}_U, \hat{\boldsymbol{\theta}}_{\boldsymbol{U}}))$ sont des estimateurs consistants de $(N, \boldsymbol{\theta})$. Pour plus de détails, voir Sanathanan (1972), Fienberg (1972), Chapman (1951) et Bishop *et al.* (1975).

1.1.4 Procédure d'estimation : vraisemblance Poisson

Cette approche est basée sur l'hypothèse que les fréquences observées $\{n_\omega\}$ suivent des lois de Poisson indépendantes (Cormack, 1989). On écrit

$$n_\omega \sim \text{Poisson}\left(\mu_\omega = NP_\omega(\boldsymbol{\theta})\right), \tag{1.9}$$

où $\boldsymbol{\theta}$ est le vecteur des paramètres du modèle. La vraisemblance Poisson s'écrit :

$$L_P\left(N; \boldsymbol{\theta}\right) = \prod_{\omega \in \Omega} \frac{e^{-\mu_\omega} \mu_\omega^{n_\omega}}{n_\omega!}. \tag{1.10}$$

Cette vraisemblance peut être réécrite sous la forme

$$L_P\left(N; \boldsymbol{\theta}\right) = L_P^1\left(P_0\left(\boldsymbol{\theta}\right)\right) L_P^2\left(\boldsymbol{\theta}\right)$$

avec

$$L_P^1\left(P_0\left(\boldsymbol{\theta}\right)\right) = \frac{e^{-NP^*(\boldsymbol{\theta})} \{NP^*\left(\boldsymbol{\theta}\right)\}^n}{n!} \tag{1.11}$$

et

$$L_P^2(\boldsymbol{\theta}) = \frac{n!}{\prod_{\omega\in\Omega} n_\omega!}\{Q_\omega(\boldsymbol{\theta})\}^{n_\omega}, \tag{1.12}$$

où $P^*(\boldsymbol{\theta}) = 1 - P_0(\boldsymbol{\theta})$. Les fonctions de vraisemblance (1.12) et (1.6) sont identiques : les vraisemblances multinomiale et Poisson partagent la même vraisemblance conditionnelle $n!/\prod_{\omega\in\Omega} n_\omega!\{Q_\omega(\boldsymbol{\theta})\}^{n_\omega}$. Ce résultat est discuté dans Sandland and Cormack (1984, page 3). Cormack and Jupp (1991) ont montré que l'estimateur Poisson $\hat{\boldsymbol{\theta}}$ est identique à l'estimateur conditionnel $\hat{\boldsymbol{\theta}}_{\boldsymbol{C}}$ obtenu à la Section 1.1.3.

Pour la plupart des modèles étudiés, la fréquence prédite μ_ω est log-linéaire :

$$\log \mu_\omega = \gamma + \boldsymbol{X_\omega}^\top \boldsymbol{\beta}, \tag{1.13}$$

où γ représente le logarithme de la fréquence prédite pour les unités non capturées, $\boldsymbol{X_\omega}$ est le vecteur de dimension $d \times 1$ des variables explicatives dont la spécification dépend du modèle considéré, et $\boldsymbol{\beta}$ est le vecteur $d \times 1$ des paramètres inconnus. Les équations (1.10) et (1.13) définissent un modèle de régression Poisson. Dans l'équation (1.13), $\boldsymbol{\theta} = (\gamma, \boldsymbol{\beta}^\top)^\top$ représente le vecteur des paramètres. On remarque que $\boldsymbol{X_0}$, la valeur de $\boldsymbol{X}$ 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 $\boldsymbol{\omega}$ correspond au vecteur $\boldsymbol{X_\omega}$. On définit le vecteur $\boldsymbol{y}$ des fréquences observées pour les $2^\ell - 1$ historiques de capture et $\boldsymbol{\mu}$ son vecteur des fréquences prédites. Les statistiques exhaustives minimales du modèle (1.13) peuvent être résumées 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 comme la probabilité de $\boldsymbol{\omega}$. On associe donc à la matrice $\boldsymbol{X}$ un vecteur aléatoire de dimension d ; on pose, pour ce vecteur aléatoire, $\boldsymbol{\mu_X}$ et $\boldsymbol{\Sigma}$ comme le vecteur de dimension $d \times 1$ des espérances et la matrice de variance-covariance $d \times d$ respectivement. Les paramètres de (1.13) peuvent être estimés en utilisant un modèle linéaire généralisé avec une distribution de Poisson (1.9) et une fonction de lien logarithmique (Sandland and Cormack 1984; Cormack and Jupp 1991). L'estimateur de N pour le modèle (1.13) s'écrit

$$\hat{N} = n + e^{\hat{\gamma}}, \tag{1.14}$$

où $e^{\hat{\gamma}}$ est l'estimateur du nombre d'unités manquées ; $\hat{\gamma}$ est une estimation de l'ordonnée à l'origine de la régression de Poisson. De (1.10) et (1.13), on tire la log-vraisemblance :

$$LL(\boldsymbol{\theta}) = \boldsymbol{y}^\top \boldsymbol{X\theta} - \boldsymbol{\mu}. \tag{1.15}$$

La fonction de score Poisson de dimension $d \times 1$ et la matrice d'information de Fisher $d \times d$ sont données respectivement par

$$s_p(\boldsymbol{\theta}) = \boldsymbol{X}^\top (\boldsymbol{y} - \boldsymbol{\mu}) \tag{1.16}$$

et

$$I_p(\boldsymbol{\theta}) = \boldsymbol{X}^\top \boldsymbol{D}(\boldsymbol{\mu}) \boldsymbol{X}, \tag{1.17}$$

où $\boldsymbol{D}(\boldsymbol{\mu})$ est la matrice diagonale $s \times s$ de $\boldsymbol{\mu}$. L'estimateur du maximum de vraisemblance de N obtenu en maximisant la vraisemblance Poisson définie en (1.10), ou en résolvant l'équation (1.16) égale à 0, a les mêmes propriétés asymptotiques que l'estimateur obtenu en maximisant la vraisemblance multinomiale donnée en (1.3), voir Fewster and Jupp (2009). Sandland and Cormack (1984) ont montré que la variance asymptotique multinomiale, $\mathrm{Var}_M(\hat{N})$, est égale à la variance Poisson, $\mathrm{Var}_P(\hat{N})$, moins N. Pour les modèles (1.13), la variance multinomiale asymptotique de l'estimateur $\hat{N}$ est

$$\mathrm{Var}_M\left(\hat{N}\right) = \frac{N}{(1-p^*)^{-1} - 1 - \boldsymbol{\mu}_{\boldsymbol{X}}^{\top}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_{\boldsymbol{X}}}, \tag{1.18}$$

où $p^* = 1 - e^{\gamma}/N$ est la probabilité d'être capturée au moins une fois au cours de l'expérience. Une estimation de la variance de (1.18), $v(\hat{N})$, est donnée par (Rivest and Lévesque, 2001)

$$v\left(\hat{N}\right) = e^{\hat{\gamma}} + e^{2\hat{\gamma}} v\left(\hat{\gamma}\right), \tag{1.19}$$

où $v\left(\hat{\gamma}\right)$ est une estimation de la variance de $\hat{\gamma}$, obtenue en ajustant le modèle linéaire généralisé (1.13) ; elle représente le premier élément de l'inverse de la matrice d'information de Fisher Poisson observée $I_p^{-1}(\boldsymbol{\theta})$ calculée à partir de (1.17), voir Annexe B.1.

Exemple 1 ***L'estimateur de Lincoln-Petersen***

Le cas de $\ell = 2$ occasions de capture est à la base du développement théorique et méthodologique des méthodes de capture-recapture. Supposons qu'on s'intéresse à l'estimation de la taille N d'une population de poissons. Une expérience de capture-recapture est menée sur deux journées consécutives (voir figure 1.1).

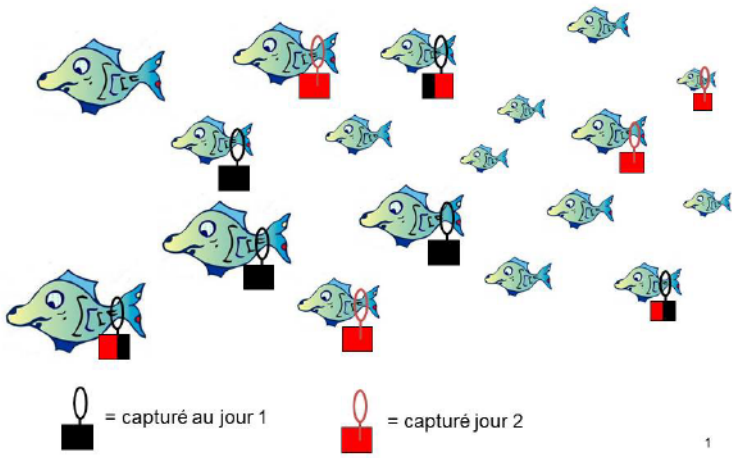

FIGURE 1.1 – Représentation schématique d'une expérience de capture-recapture sur une population de poissons. Une marque noire (resp. rouge) indique une capture au jour 1 (resp. 2). Tiré de Rivest (2013).

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'une journée à l'autre. Après deux jours, il y'a trois historiques de capture observables $(1,0),(1,1),(0,1)$ avec des fréquences respectives n_{10}, n_{11}, n_{01}. Ces données sont représentées dans le tableau de contingence 1.1.

TABLE 1.1 – Données issues d'une expérience de capture-recapture à deux occasions de capture.

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

Les fréquences prédites correspondant aux données du tableau de contingence 1.1 sont représentées dans le tableau suivant.

TABLE 1.2 – Fréquences prédites des données de l'expérience de Lincoln-Petersen.

	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	$\mu_{01} = N(1-p_1)p_2$	$\mu_{11} = Np_1p_2$

Ces fréquences prédites peuvent être réécrites sous la forme (1.13) en fonction des paramètres log-linéaires $\gamma = \log\{N(1-p_1)(1-p_2)\}$ et $\beta_j = \log\{p_j/(1-p_j)\}\,(j=1,2)$. Le paramètre γ est estimé en résolvant l'équation (1.16) égale à 0 ; on obtient alors l'estimateur du nombre de poissons manqués :

$$e^{\hat{\gamma}} = \frac{n_{01}n_{10}}{n_{11}}. \tag{1.20}$$

En substituant (1.20) dans (1.14) (avec $n = n_{10}+n_{11}+n_{01}$), on obtient l'estimateur du nombre de poissons

$$\hat{N} = (n_{10}+n_{11}) \,/\, \left(\frac{n_{11}}{n_{01}+n_{11}}\right). \tag{1.21}$$

Cet estimateur est communément appelé l'estimateur de Lincoln-Petersen, en référence à Petersen (Lecren, 1965), qui s'est intéressé à l'estimation de la taille d'une population de poissons, et à (Lincoln, 1930), qui s'est intéressé aux retours de bagues de sauvagines, voir Seber (1982, Ch. 3). La variance multinomiale de $\hat{N}$ est obtenue en substituant $\boldsymbol{\mu_X} = (p_1, p_2)$ et

$$\boldsymbol{\Sigma}^{-1} = \begin{pmatrix} \{p_1(1-p_1)\}^{-1} & 0 \\ 0 & \{p_2(1-p_2)\}^{-1} \end{pmatrix}$$

dans (1.18). On obtient

$$Var_M(\hat{N}) = \frac{N(1-p_1)(1-p_2)}{p_1p_2}. \tag{1.22}$$

Une estimation de la variance multinomiale (1.22), $v(\hat{N})$, est calculée en substituant la quantité $v(\hat{\gamma}) = 1/n_{10} + 1/n_{01} + 1/n_{11}$ dans (1.19) ; on obtient

$$v\left(\hat{N}\right) = \hat{N} \times \frac{n_{01} n_{10}}{n_{11}^2}. \tag{1.23}$$

L'estimateur de Petersen (1.21) est également donné dans Seber (1982, page 131). La même formule de la variance (1.23) est donnée dans Seber (1982, page 60). L'estimateur (1.21) peut comporter un biais souvent dû à un faible nombre d'unités recapturées (n_{11}). Des versions non biaisées de l'estimateur de Lincoln-Petersen sont discutées dans Chapman (1951) et dans Seber (1982, page 61).

1.1.5 Modélisation de la capture des unités

On définit le paramètre p_{ij} comme étant la probabillité que l'unité i ($i = 1, 2, \ldots, N$) soit capturée à l'occasion j ($j = 1, 2, \ldots, \ell$). Dans leur monographe *Statistical Inference from Capture Data for Closed Animal Populations*, Otis *et al.* (1978) ont présenté trois types de variations qui peuvent affecter la probabilité p_{ij} : (1) temporelle (t), (2) comportemental (b) et (3) hétérogène (h). Plusieurs modèles en découlent : M_0 (l'indice 0 signifie qu'il n'y a aucune variation dans la probabilité de capture des unités), M_t, M_b et M_h, et les modèles qui combinent jusqu'à deux types d'effets (M_{bh}, M_{th}). La spécification de $P_\omega(\boldsymbol{\theta})$ dans (1.9) permet de modéliser différents scénarios pour la capture des unités dans la population. Ces modèles sont traités dans cette section.

Le modèle M_0

Le modèle M_0 suppose que $p_{ij} = p$: la probabilité de capture est constante d'une à l'autre et d'une occasion de capture à l'autre. En termes de paramétrisation log-linéaire, $d = 1$, $X_\omega = \sum_{j=1}^{\ell} \omega_j$ est le nombre de fois qu'une unité est marquée et $\boldsymbol{X}^\top \boldsymbol{y}$ est le nombre total de captures. Le modèle (1.13) devient $\log \mu_\omega = \gamma + \sum_{j=1}^{\ell} \omega_j \beta$, où $\gamma = \log\left\{N\,(1-p)^\ell\right\}$ et $\beta = \log\{p/(1-p)\}$. Le vecteur de paramètres est $\boldsymbol{\theta} = (\gamma, \beta)^\top$ et la fréquence prédite des unités manquées est $\mu_0 = N\,(1-p)^\ell$. La variable aléatoire X_ω suit une binomiale de paramètres ℓ et p, donc $\mu_X = \ell p$ et $\Sigma = \ell p\,(1-p)$. En substituant ces deux quantités dans (1.19), on obtient la variance multinomiale asymptotique de $\hat{N}$

$$\mathrm{Var}_M\left(\hat{N}\right) = \frac{N}{(1-p)^{-\ell} + (\ell - 1) - \ell\,(1-p)^{-1}}. \tag{1.24}$$

La vraisemblance multinomiale pour ce modèle est obtenue en remplaçant $P_\omega(\boldsymbol{\theta})$ par p dans (1.3). La maximisation de cette vraisemblance n'aboutit pas à une solution explicite pour l'estimateur de la taille de population $\hat{N}$. Otis *et al.* (1978, page 105) ont proposé un algorithme de calcul des estimateurs du maximum de vraisemblance pour N et p.

Le modèle M_t

Comme généralisation du modèle M_0 par l'inclusion d'une variation temporelle aux probabilités de capture, le modèle M_t (l'indice t signifie une variation de la probabilité de capture dans le temps) suppose que $p_{ij} = p_j$, $j = 1, 2, \ldots, \ell$. La matrice de design de ce modèle compte $d = \ell$ colonnes, $\boldsymbol{X}_{\boldsymbol{\omega}} = (\omega_1, \omega_2, \ldots, \omega_\ell)$. Le modèle (1.13) s'écrit $\log \mu_\omega = \gamma + \sum_{j=1}^{\ell} \omega_j \beta_j$, où $\gamma = \log \left\{ N \prod_{j=1}^{\ell} (1 - p_j) \right\}$ et $\beta_j = \log \{p_j / (1 - p_j)\}$, $j = 1, 2, \ldots, \ell$. Le vecteur de paramètres est $\boldsymbol{\theta} = (\gamma, \beta_1, \beta_2, \ldots, \beta_\ell)^\top$ et la fréquence prédite des unités manquées est $\mu_0 = N \prod_{j=1}^{\ell} (1 - p_j)$. Les variables aléatoires $\omega_1, \omega_2, \ldots, \omega_\ell$ sont indépendantes et suivent des lois de Bernoulli, donc $\boldsymbol{\mu}_{\boldsymbol{X}} = (p_1, p_2, \ldots, p_\ell)^\top$ et $\boldsymbol{\Sigma}$ est une matrice diagonale d'éléments $p_j (1 - p_j)$, $j = 1, 2, \ldots, \ell$. Darroch (1958) a montré que les estimateurs du maximum de vraisemblance pour p et N satisfont

$$\hat{p}_j = \frac{n_j}{\hat{N}}, \quad j = 1, 2, \ldots, \ell,$$

et

$$1 - \frac{n}{\hat{N}} = \prod_{j=1}^{\ell} \left(1 - \frac{n_j}{\hat{N}} \right).$$

Darroch (1958) a également montré que la variance multinomiale asymptotique de $\hat{N}$ est

$$\mathrm{Var}_M \left(\hat{N} \right) = \frac{N}{\prod_{j=1}^{\ell} (1 - p_j)^{-1} + (\ell - 1) - \sum_{j=1}^{\ell} (1 - p_j)^{-1}}. \tag{1.25}$$

L'équation (1.25) peut être retrouvée en substituant $\boldsymbol{\mu}_{\boldsymbol{X}}$ et $\boldsymbol{\Sigma}$ dans (1.19). Le modèle M_t constitue une généralisation du modèle à deux occasions de capture présenté dans l'exemple 1.

Le modèle M_b

Le relâchement de l'hypothèse d'indépendance entre les occasions de capture est à la base du modèle M_b, vu comme une généralisation du modèle M_0 ; l'indice b fait référence à un changement comportemental (ou "behavioral") de l'unité après la première capture. Le modèle suppose que $p_{ij} = p$ jusqu'à la toute première capture de l'unité. Puis, pour toute recapture de l'unité déjà marquée, $p_{ij} = c$. Ainsi, toutes les unités de la population exhibent le même comportement face à une éventuelle recapture. Si $p > c$, on se retouve dans la situation "trap-shy" : l'unité est plus difficile à revoir après la première capture. Si $p < c$, on se retrouve dans la situation "trap-happy" : l'unité devient plus facile à appréhender après la première capture. Dans le cas du modèle M_b, seules les premières captures contribuent à l'estimation de N : le paramètre c est un paramètre de nuisance. Le vecteur des statistiques exhaustives pour ce modèle est $(u_1, u_2, \ldots, u_\ell)$; leurs fréquences prédites sont

$$\mathrm{E}(u_j) = \mu_j = Np(1 - p)^{j-1}, \quad j = 1, 2, \ldots, \ell.$$

Le modèle log-linéaire s'écrit

$$\log \mu_j = \gamma + (j - 1)\beta, \tag{1.26}$$

où $\gamma = \log(Np)$ et $\beta = \log(1-p)$. L'estimateur de la taille de la population peut ainsi s'exprimer en fonction des paramètres du modèle, $\hat{N} = e^{\hat{\gamma}} / \left(1 - e^{\hat{\beta}}\right)$; la probabilité qu'une soit capturée au moins une fois est $p^* = 1 - e^{\beta\ell}$ et l'espérance du nombre d'unités manquées est $\mu_0 = e^{\gamma} e^{\beta\ell} / \left(1 - e^{\beta}\right)$. La variance multinomiale asymptotique de l'estimateur $\hat{N}$ est

$$\mathrm{Var}_M\left(\hat{N}\right) = \frac{N(1-p)^{\ell}\left\{1-(1-p)^{\ell}\right\}}{\left\{1-(1-p)^{\ell}-\ell^2 p^2 (1-p)^{\ell-1}\right\}^2}. \tag{1.27}$$

Pour plus de détails sur le modèle (1.26) et sur le calcul de (1.27), voir le chapitre 3. Otis *et al.* (1978, page 107) ont proposé une méthode itérative pour trouver les estimateurs du maximum de vraisemblance dans le cas multinomial.

D'un point de vue statistique, le modèle M_b est équivalent au modèle d'élimination (ou "removal model"), voir Moran (1951) et Zippin (1956). L'existence des estimateurs du maximum de vraisemblance dans les modèles d'élimination est sujette à certaines conditions sur le nombre d'unités capturées pour la première fois. Par exemple, pour deux occasions de capture, $u_1 > u_2$: le nombre d'unités capturées pour la première fois doit généralement décroître d'une occasion à l'autre (Otis *et al.*, 1978).

Modèle M_h

Le relâchement de l'hypothèse d'homogénéité des probabilités de capture entre les unités de la population, postulée par les modèles dits homogènes (M_0, M_t et M_b), est à la base du modèle M_h ; l'indice h signifie qu'il y'a une hétérogénéité dans la capture des unités. Le modèle postule que chaque unité a sa propre probabilité de capture : $p_{ij} = p_i$. Ainsi, les unités qui ont une probabilité de capture élevée auront tendance à apparaitre dans l'échantillon des unités marquées à une proportion beaucoup plus élevée que dans la population, entrainant une sous-estimation de N.

Il y a principalement deux approches de modélisation de l'hétérogénéité, lesquelles dépendent de l'information disponible sur les unités capturées. La première approche considère que la source d'hétérogénéité entre les unités est observable, et peut-être modélisée à l'aide de covariables individuelles mesurées sur les unités capturées (Huggins, 1989). La seconde approche considère que l'hétérogénéité est non observable, et peut-être modélisée par un effet aléatoire individuel. Cette approche suppose que les probabilités de capture $\{p_i\}$ constituent un échantillon de taille N issu d'une distribution de probabilité. Deux grandes classes de modèles d'hétérogénéité ont été développés, selon que la distribution des probabilités de capture $\{p_i\}$ est discrète ou continue. Norris and Pollock (1996) et Pledger (2000) ont proposé des distributions de mélange pour des modèles de classes latentes. Ces méthodes permettent de partitionner les unités en des groupes relativement homogènes en termes de probabilité de capture. Cette approche est implémentée dans le logiciel MARK (White and Burnham, 1999).

Coull and Agresti (1999) et Dorazio and Royle (2003) ont proposé de modéliser l'hétérogénéité avec des variables latentes continues. Dans la suite de cette section, l'accent sera mis sur cette approche.

De façon formelle, la probabilité p_i qu'une unité i soit capturée à l'occasion j satisfait la relation

$$\log\left(\frac{p_i}{1-p_i}\right) = \alpha_i + \beta, \quad i = 1, 2, \ldots, N, \tag{1.28}$$

où α_i, $i = 1, 2, \ldots, N$ sont des variables aléatoires indépendantes de distribution $F(x)$ (de type normale, beta ou gamma) et β représente un effet fixe. En supposant des distributions relativement complexe pour α_i, Rivest and Baillargeon (2007) ont obtenu des modèles log-linéaires relativement simples pour μ_ω,

$$\log \mu_\omega = \gamma + \beta \sum_{j=1}^{\ell} \omega_j + \tau\psi\left(\sum_{j=1}^{\ell} \omega_j\right), \tag{1.29}$$

où τ est un paramètre d'hétérogénéité, $\psi(\cdot)$ est une fonction convexe telle que $\psi(0) = 0$. Par exemple, en supposant que α_i suit un mélange particulier de lois normales, on obtient le modèle M_h de Darroch *et al.* (1993) avec $\psi(t) = t^2/2$; $\psi(t) = \exp(at) - 1$ $(a > 1)$ correspond au modèle de mélange Poisson avec paramètre de forme τ ; $\psi(t) = -\log(\lambda + t) + \log(\lambda)$ $(\lambda > 1)$ correspond a un modèle de mélange gamma négatif (Lindsay 1986; Rivest and Baillargeon 2007). L'estimateur de N obtenue avec les trois fonctions ψ sont ainsi ordonnés : Poisson $<$ normal $<$ Gamma (Rivest and Baillargeon, 2007). La matrice design du modèle M_h compte $d = 2$ colonnes, $\boldsymbol{X}_{\boldsymbol{\omega}} = (\sum \omega_i, \psi(\sum \omega_i))^\top$. On peut retrouver le modèle M_0 en posant $\tau = 0$. La probabilité qu'une unité soit capturée au moins une fois est

$$p^* = \frac{\sum_{\omega\in\Omega} \exp\left\{\beta \sum_{j=1}^{\ell} \omega_j + \tau\psi\left(\sum_{j=1}^{\ell} \omega_j\right)\right\}}{1 + \sum_{\omega\in\Omega} \exp\left\{\beta \sum_{j=1}^{\ell} \omega_j + \tau\psi\left(\sum_{j=1}^{\ell} \omega_j\right)\right\}}. \tag{1.30}$$

Pour le modèle M_0, (1.30) devient $p^* = 1 - (1+\beta)^{-\ell} = 1 - (1-p)^{\ell}$. Le vecteur des paramètres du modèle M_h est $\boldsymbol{\theta} = (\gamma, \beta, \tau)^\top$ et la fréquence prédite des unités manquées est $\mu_0 = N(1-p^*)$. Le vecteur des espérances est $\boldsymbol{\mu_X} = \left(\sum_{\boldsymbol{j=1}}^{\boldsymbol{\ell}} \boldsymbol{jP_j}, \sum_{\boldsymbol{j=1}}^{\boldsymbol{\ell}} \boldsymbol{\psi(j) P_j}\right)$ et la matrice de variance-covariance est

$$\boldsymbol{\Sigma} = \begin{pmatrix} \sum_{j=1}^{\ell} j^2 P_j - \left\{\sum_{j=1}^{\ell} jP_j\right\}^2 & \sum_{j=1}^{\ell} j\psi(j) P_j - \sum_{j=1}^{\ell} jP_j \sum_{j=1}^{\ell} \psi(j) P_j \\ & \sum_{j=1}^{\ell} \psi^2(j) P_j - \left\{\sum_{j=1}^{\ell} \psi(j) P_j\right\}^2 \end{pmatrix},$$

où

$$P_j = \Pr\left(\sum_{j=1}^{\ell} \omega_j = j\right) = \binom{\ell}{j} \exp\{\beta j + \tau\psi(j)\} \Big/ \sum_{j=0}^{\ell} \binom{\ell}{j} \exp\{\beta j + \tau\psi(j)\}$$

est la probabilité qu'une unité soit capturée j fois. La variance multinomiale asymptotique de $\hat{N}$ est obtenue en remplaçant ces deux quantités dans (1.18).

Burnham and Overton (1978) et Pollock and Otto (1983) ont proposé une approche non paramètrique basée sur l'application de la méthode du Jackniffe, proposée par Quenouille (1949), aux fréquences de capture $(f_1, f_2, \ldots, f_\ell)$ pour estimer N. L'idée de la méthode est de partir de la statistique n comme estimateur de N, puis de procéder à la réduction de son biais. Chao *et al.* (1992) ont proposé un estimateur basé sur la mesure d'une couverture d'échantillon ; voir Amstrup *et al.* (2005) pour plus de détails.

Huggins (2001) et Link (2003) ont montré qu'il y'a un problème d'estimabilité de la taille de la population en présence d'hétérogénéité : deux modèles qui ajustent bien aux données peuvent donner des estimations complètement différentes de N. En effet, les données ne permettent pas d'identifier clairement la distribution $F(x)$ dont les estimations de N dépendent ; voir Rivest and Baillargeon (2013) pour des exemples d'applications à des données sur de petits mammifères de l'espèce *Micromys minutus*.

Chao (1989) a proposé un estimateur pour la borne inférieure de N comme alternative aux approches existantes. Dans le cadre de la modélisation log-linéaire, l'idée est d'ajouter dans la matrice $\boldsymbol{X}$ du modèle M_h un paramètre à chaque historique $\boldsymbol{\omega}$ comportant plus de deux captures. Le modèle qui en résulte est dénoté M_{Ch}. Rivest and Baillargeon (2007) ont proposé des estimateurs de borne inférieure pour différents modèles M_h dans un cadre log-linéaire, puis ont étudié les biais de ces estimateurs. L'estimateur $\hat{N}$ obtenu s'exprime en fonction de f_1 et f_2,

$$\hat{N} = n + \frac{(\ell - 1) f_1^2}{2\ell f_2}. \tag{1.31}$$

Quand ℓ tend vers l'infini, (1.31) correspond à l'estimateur de borne inférieure obtenu dans le cas où les occasions de capture sont continues ; cet estimateur est présenté dans Chao (1987). L'estimateur de borne inférieure (1.31) peut être obtenu en utilisant (1.16) ; le vecteur de paramètres $\boldsymbol{\theta} = (\gamma, \beta)^\top$ est alors estimé en résolvant les équations qui égalisent les statistiques observées f_1 et f_2 à leurs valeurs prédites sous le modèle homogène M_0. La non prise en compte de l'information $f_k\,(k \geq 3)$ sur les unités capturées trois fois ou plus, quoique pertinente dans un modèle hétérogène, cause une augmentation de la variance de l'estimateur $\hat{N}$ (Rivest and Baillargeon, 2013).

Le modèle M_{th}

L'ajout d'un effet temporel à la capture des unités dans le modèle M_h donne le modèle M_{th}. Le modèle postule que chaque unité a sa propre probabilité de capture, qui est variable d'une occasion à l'autre. Le modèle M_{th} obéit à une écriture log-linéaire, sous la forme (1.29), avec la définition d'un effet fixe $\beta_j\,(j = 1, 2, \ldots, \ell)$ spécifique à chaque occasion de capture. La matrice design du modèle compte $d = \ell + 1$ colonnes, $\boldsymbol{X}_{\boldsymbol{\omega}} = (\omega_1, \omega_2, \ldots, \omega_\ell, \psi(\sum \omega_i))^\top$. Pour plus de détails, voir Pledger (2000) , Dorazio and Royle (2003) et Rivest and Baillargeon (2007) . Darroch *et al.* (1993) et Agresti (1994) ont proposé de modéliser l'hétérogénéité en ajoutant

des interactions à toutes les paires d'occasions de capture. Formellement, le modèle s'écrit

$$\log \mu_\omega = \gamma + \sum_{j=1}^{\ell} \beta_j \omega_j + \tau \sum_{j>k} \omega_j \times \omega_k. \tag{1.32}$$

Coull and Agresti (1999) ont montré que le modèle (1.32) s'ajuste bien à des données comportant des captures hétérogènes.

Rivest (2008) a étudié l'impact de l'inclusion d'effets temporel et hétérogène sur l'estimateur de N. Il a montré que, pour le modèle M_{th}, le paramètre d'hétérogénéité diminue si l'effet temporel est omis : le modèle M_h prend en compte, de façon indirecte, une variation temporelle dans la capture des unités. Ainsi, en présence d'hétérogénéité, un effet temporel a très peu d'impact sur l'estimateur de N.

Le modèle M_{bh}

Le modèle M_{bh} peut être vu comme une généralisation du modèle M_h pour permettre aux probabilités de capture des unités de dépendre des captures précédentes. Le modèle suppose que $p_{ij} = p_i$ jusqu'à la toute première capture de l'unité. Puis, pour toute capture subséquente, $p_{ij} = c_i$. Un des attraits de ce modèle est qu'il permet, contrairement au modèle M_b, une variation comportementale propre à chaque unité de la population. Ce modèle compte $2N+1$ paramètres $(N, \{p_i\}, \{c_i\})$. Le temps écoulé jusqu'à la première capture suit une loi mélange géométrique. Une approche de modélisation consiste donc à ajouter un terme d'hétérogénéité au modèle M_b donné en (1.26). On peut également supposer que les unités capturées avant $\ell_0 \geq 0$ occasions de capture ne sont pas représentatives des unités non marquées. De manière formelle, on considère que les ℓ_0 premières valeurs de $\{u_j\}$ ne suivent pas le modèle (1.26). Le modèle s'écrit

$$\log \mu_{\ell_0+j} = \gamma + (j-1)\,\beta, \quad j = 1, 2, \ldots, j - \ell_0, \tag{1.33}$$

où $\gamma = \log(N_1 p)$, $\beta = \log(1-p)$; N_1 représente les unités non marquées après t_0 occasions, p est la probabilité qu'une unité soit capturée aux occasions $\ell_0 + 1, \ldots, \ell$. Avec cette paramétrisation, $\hat{N}_1 = e^{\hat{\gamma}} / \left(1 - e^{\hat{\beta}}\right)$ et l'estimateur de la taille de la population est $\hat{N} = \sum_{j=1}^{\ell_0} u_j + \hat{N}_1$. Pour plus de détails sur le calcul de la variance multinomiale asymptotique de l'estimateur $\hat{N}$, voir Rivest and Lévesque (2001) .

La relation entre les modèles de population fermée, discutés Otis *et al.* (1978) et présentés dans cette section, est illustrée à la figure 1.2.

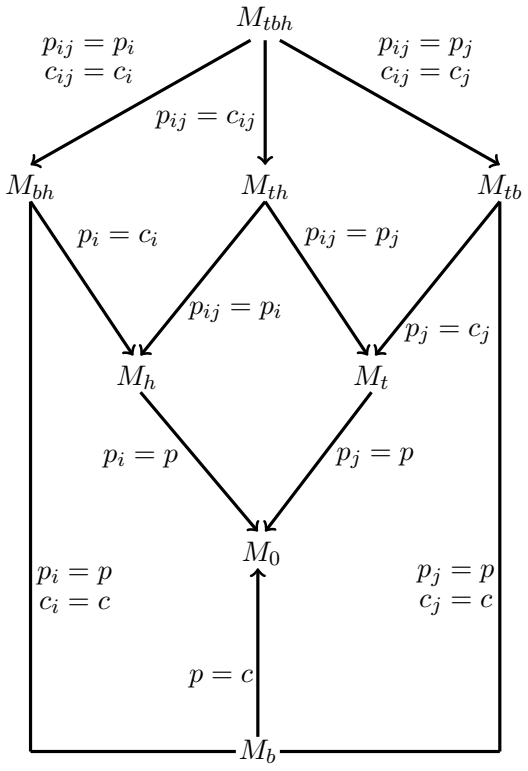

FIGURE 1.2 – La relation entre les modèles de population fermée M_{tbh}, M_{tb}, M_{th}, M_{hb}, M_h, M_b, M_t, M_0, présentés dans l'article de Otis *et al.* (1978).

1.1.6 Développements récents

Le problème de l'estimation de N a connu des développements théoriques importants au cours de la dernière décennie. Des avancées méthodologiques considérables, supportées par l'exposition des méthodes de capture-recapture à de nouveaux domaines d'applications, ont été achevées. Cette section relate les développements récents liés aux méthodes de capture-recapture pour populations fermées.

Modélisation bayesienne

King *et al.* (2008) et King *et al.* (2010) ont développé une méthodologie pour l'ajustement des modèles de population fermée M_0, M_t, M_b, M_{tb}, M_h, M_{th}, , M_{bh}, M_{tbh} dans un cadre bayésien. Ils ont proposé le modèle saturé M_{tbh}, qui exprime le logit de la probabilité de capture p_{ij} en fonction des paramètres log-linéaires. Le modèle spécifie des distributions a priori pour tous les paramètres. Par exemple, la distribution a priori choisie pour la taille de la population N peut être une distribution (non informative) uniforme ou de Jeffrey. Pour une présentation détaillée sur la construction de la vraisemblance ainsi que l'estimation des paramètres dans un cadre bayésien, voir King *et al.* (2010, pages 345-360), Link and Barker (2009) et McCrea and Morgan (2015, page 45).

Modèles spatiaux de capture-recapture

Dans la conduite d'une expérience de capture-recapture, l'installation de trappes dans la zone d'habitat de la population est une pratique assez commune. Les méthodes d'échantillonnage sont généralement de deux types : (1) des trappes qui retiennent l'unité pour la durée de l'occasion de capture et (2) des outils de détection qui ne nécessitent pas la capture de l'unité (détection visuelle, acoustique ou par caméra, ...). Ces méthodes requièrent la connaissance des coordonnées géographiques des trappes ainsi que la possibilité d'identifier clairement les unités détectées à chaque occasion de capture, voir Borchers *et al.* (2002, pages 2-3). Les modèles spatiaux incorporent la dimension spatiale des données recueillies dans la modélisation de la capture des unités.

Borchers *et al.* (2002) ont proposé de modéliser les données spatiales de capture-recapture en utilisant deux composantes : (1) une composante "état" (ou "state model") qui modélise la position des unités et (2) une composante "observation" (ou "observation model") qui décrit le processus d'observation des unités conditionnellement à leurs positions géographiques. Le "state model" suppose une distribution de probabilité conjointe pour le vecteur des coordonnées géographiques des unités observées ; il peut s'agir d'une distribution uniforme (Illian *et al.*, 2008) avec N comme paramètre ou d'un processus de Poisson homogène avec la densité des unités comme paramètre (N est considéré comme aléatoire). La probabilité de capture est modélisée comme étant fonction de la distance entre la position de la trappe et le "centre" de la position de l'unité (Borchers and Efford 2008; Borchers 2012).

1.2 Les modèles de population ouverte

Cette section traite des modèles de population ouverte : spécification du modèle et hypothèses, procédure d'estimation, calcul de la variance de l'estimateur de la taille de la population. Deux classes de modèles ont été développées pour traiter des données issues d'une expérience de capture-recapture pour population ouverte. La première, développée par Jolly (1965) et Seber (1965), s'intéresse particulièrement à l'estimation de la taille de la population. L'approche Jolly-Seber (JS) modélise les recaptures des unités déjà marquées ainsi que les premières captures des unités non marquées, permettant l'estimation de la survie, de la probabilité de capture et de la taille de la population. La seconde classe de modèles, discutée dans Cormack (1964), Jolly (1965) et Seber (1965) puis vulgarisée dans un cadre unificateur par Lebreton *et al.* (1992), s'intéresse à l'estimation de la survie des unités dans la population. L'approche Cormack-Jolly-Seber (CJS) se base sur la construction d'une vraisemblance conditionnelle au nombre d'unités relâchées à chaque période d'échantillonnage, permettant l'estimation de la survie et de la probabilité de capture. On suppose, sans perte de généralité, qu'il n'y a pas de décès en capture. Cette section présente les modèles de CJS et de JS.

1.2.1 Le modèle de Cormack-Jolly-Seber

On considère une expérience de capture-recapture dans laquelle le marquage des unités s'effectue avant le début de l'étude ; par exemple, des sessions annuelles de marquage sont organisées aux périodes de reproduction des unités. On suppose qu'au cours de l'étude, les marques posées sur les unités sont clairement visibles : une recapture correspond donc à l'observation d'un tag posé sur une unité ; voir Pollock *et al.* (1990, page 51) et Cormack (1964) pour des exemples typiques d'études. On suppose que les unités sont observées sur I périodes discrètes ou périodes de capture (PC). On définit également les paramètres suivants :

— ϕ_i : la probabilité conditionnelle qu'une unité soit en vie à la PC $i+1$ sachant qu'elle était présente et en vie à la PC i ;

— p_i : la probabilité conditionnelle qu'une unité déjà marquée soit capturée à la PC i sachant qu'elle est en vie à cette période ;

— χ_i est la probabilité qu'une unité ne soit plus revue après la PC i ; elle satisfait la relation récursive $\chi_i = (1-\phi_i) + \phi_i(1-p_{i+1})\chi_{i+1}$; $\bar{\chi}_i = 1 - \chi_i$.

La procédure d'échantillonnage qui sous-tend le modèle de CJS est illustrée sur la figure 1.3.

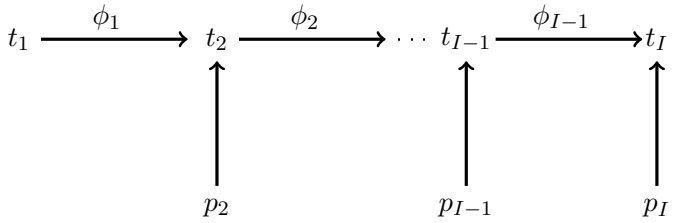

FIGURE 1.3 – La procédure d'échantillonnage décrivant un modèle de Cormack-Jolly-Seber avec une survie dépendant du temps. Les unités présentes dans la population au début de l'expérience (t_1) sont suivies jusqu'au temps de censure (t_I) correspondant au dernier temps d'observation.

On définit les statistiques suivantes :

— R_i : le nombre d'unités relâchées à la PC i, $i = 1, 2, \ldots, I$;

— m_{ij} : le nombre d'unités relàchées à la PC i et observées à nouveau à la PC j, $j = i+1, i+2, \ldots, I$; $m_{i,\infty}$ représente le nombre d'unités relâchées à la PC i et qui ne sont jamais revues.

L'information sur la recapture des unités est habituellement résumée dans un m-array, comme illustrée dans le tableau 1.3.

TABLE 1.3 – Matrice des statistiques sur la recapture des unités relâchées à chacune des $I-1$ périodes d'échantillonnage du modèle de Cormack-Jolly-Seber.

	Période de recapture						
Période de relâche	2	3	...	$I-1$	I	Jamais	Total
1	m_{12}	m_{13}	...	$m_{1,I-1}$	m_{1I}	$m_{1\infty}$	R_1
2		m_{23}	...	$m_{2,I-1}$	m_{2I}	$m_{2\infty}$	R_2
$\vdots$			...	...	...	...	...
$I-2$				$m_{I-2,I-1}$	$m_{I-2,I}$	$m_{I-2,\infty}$	R_{I-2}
$I-1$					$m_{I-1,I}$	$m_{I-1,\infty}$	R_{I-1}

Le tableau 1.3 est construit en se basant sur la recapture des unités relâchées à chaque période. Pour une unité recapturée k ($k \leq I-1$) fois, son historique de capture est découpée en k parties dont les fréquences observées correspondent aux statistiques $\{m_{ij}\}$. Les fréquences prédites correspondant aux données du tableau 1.3 sont représentées dans le tableau 1.4

TABLE 1.4 – Fréquences observées et prédites pour un modèle de Cormack-Jolly-Seber avec 2 périodes de recapture et une survie dépendant du temps.

	Période de recapture		
Période de relâche	2	3	Jamais
1	$R_1\phi_1 p_2$	$R_1\phi_1(1-p_2)\phi_2 p_3$	$R_1\chi_1$
2		$R_2\phi_2 p_3$	$R_2(1-\phi_2 p_3)$

On définit $r = \phi_2 p_3$. Les paramètres ϕ_2 et p_3 ne peuvent pas être estimés séparément ; de façon générale, seul $\phi_{I-1}p_I$ est estimable. Le paramètre p_1 n'apparait pas dans l'écriture des fréquences prédites, le rendant non estimable. Cormack (1964, page 3) a dérivé une vraisemblance conditionnelle au nombre d'unités relâchées à chaque PC. À partir du tableau 1.4 des fréquences prédites, on écrit la vraisemblance conditionnelle au nombre d'unités relâchées à chacune des périodes 1 et 2 :

$$L(\phi_1, p_2, r) \propto (\phi_1 p_2)^{m_{12}}\{\phi_1(1-p_2)r\}^{m_{13}}(\chi_1)^{m_{1\infty}} r^{m_{23}}(1-r)^{m_{2\infty}}. \tag{1.34}$$

La maximisation de cette vraisemblance fournit des estimateurs qui sont asymptotiquement non biaisés, normalement distribués et de variance minimale parmi tous les autres estimateurs de cette classe, voir Cormack (1964, pages 4-6) et Lebreton *et al.* (1992, pages 6-8).

Modèles à covariables individuelle et temporelle

Les probabilités de survie et de capture peuvent varier en fonction des attributs des unités observées ou de covariables temporelles. Beaucoup de travaux se sont focalisés sur la modélisa-

tion de la survie et de la capture comme fonction de variables discrètes caractérisant l'état des unités observées à différentes périodes (Lebreton *et al.*, 1992). Par exemple, les probabilités de survie et de capture peuvent être modélisées comme fonctions de variables temporelles environnementales ou liées à l'effort de capture. Cette modélisation incorpore les covariables dans la construction de la vraisemblance dont la maximisation permet d'estimer les paramètres de la relation fonctionnelle. Lebreton *et al.* (1992) ont introduit une fonction de lien linéaire entre les paramètres et les covariables et qui peut être de type inverse, identité, logarithme ou logit, voir Amstrup *et al.* (2005, page 9). Bonner and Schwarz (2006) ont proposé une approche de modélisation des probabilités de survie et de capture en présence de covariables continues. Ils ont utilisé une fonction de lien logit puis une approche bayesienne d'estimation des paramètres du modèle.

1.2.2 Le modèle de Jolly-Seber

On considère une expérience de capture-recapture dans laquelle les unités sont capturées, marquées puis relâchées sur I PCs. À chaque PC, les unités vues pour la première fois sont marquées et relâchées ; les tags posés sur les unités précédemment marquées sont notés avant qu'elles ne soient relâchées. Ce modèle suppose que toutes les unités, marquées ou non, ont la même probabilité d'être capturées. Cette hypothèse permet l'estimation de paramètres démographiques tels que la taille de la population, les naissances et le taux de reproduction.

On définit les paramètres suivants :

- B_i est l'espérance du nombre d'unités qui entrent dans la population entre les PC i et $i+1$, pour $i = 1, 2, \ldots, I-1$, si on suppose qu'il n'y a pas de mortalité liée à la capture des unités ;
- N_i est la taille espérée de la population à la PC i ; elle satisfait la relation $N_i = N_{i-1}\phi_{i-1} + B_{i-1}$;
- U_i est l'espérance du nombre d'unités non marquées juste avant la PC i ; elle satisfait la relation $U_i = U_{i-1}\left(1 - p_i\right)\phi_{i-1} + B_{i-1}$ pour $i = 1, ..., I-1$; $U_1 = N_1$;
- M_i est l'espérance du nombre d'unités marquées dans la population avant la PC i ; elle satisfait la relation $M_i = (M_{i-1} + U_{i-1}p_{i-1})\phi_{i-1}$; $M_1 = 0$.

La procédure d'échantillonnage qui sous-tend le modèle de JS est illustrée sur la figure 1.4.

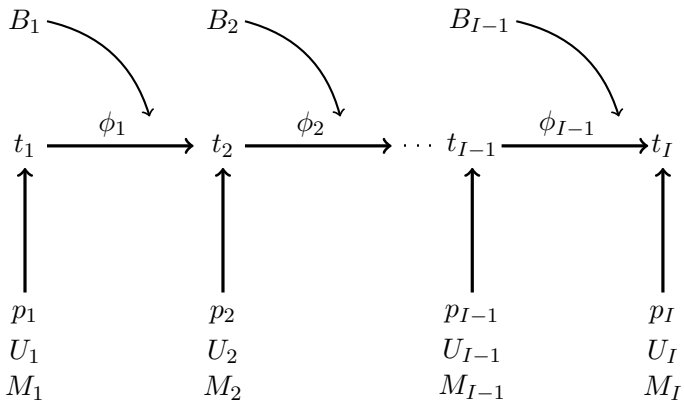

FIGURE 1.4 – La procédure d'échantillonnage décrivant le modèle de Jolly-Seber avec une survie dépendant du temps. Les unités présentes dans la population au début de l'expérience (t_1) sont suivies jusqu'au temps de censure (t_I) correspondant au dernier temps d'observation.

Un historique de capture pour le modèle de Jolly-Seber est une séquence de 0 et de 1 résumée dans le vecteur $\delta = (\delta_1, \delta_2, \ldots, \delta_I)$, avec $\delta_i = 1$ si l'unité est capturée à la PC i et 0 sinon. Reprenons l'exemple de l'expérience à $I = 3$ occasions de capture décrite à la section précédente. Le nombre d'historiques de capture observables est $\text{card}(\Omega) = 2^3 - 1 = 7$; les fréquences espérées correspondantes μ_δ sont présentées à la table 1.5.

TABLE 1.5 – Fréquentes prédites pour un modèle de Jolly-Seber avec $I = 3$ périodes de capture et une survie dépendant du temps.

δ	μ_δ
$(1,1,1)$	$U_1 p_1 \phi_1 p_2 \phi_2 p_3$
$(1,1,0)$	$U_1 p_1 \phi_1 p_2 \chi_2$
$(1,0,1)$	$U_1 p_1 \phi_1 (1-p_2) \phi_2 p_3$
$(1,0,0)$	$U_1 p_1 \chi_1$
$(0,1,1)$	$U_2 p_2 \phi_2 p_3$
$(0,1,0)$	$U_2 p_2 \chi_2$
$(0,0,1)$	$U_3 p_3$

Les fréquences prédites μ_δ du tableau 1.5 s'expriment en fonction de paramètres associés au modèle de CJS $(p_1, p_2, p_3, \phi_1, \phi_2)$ et de paramètres propres au modèle de JS (U_1, U_2, U_3). Les estimateurs associés à ces paramètres sont obtenus par la maximisation de la vraisemblance du modèle original de JS, qui est le produit de deux composantes : (1) L_1 qui concerne la population des unités non marquées à chaque PC et (2) L_2 qui contient l'information sur la recapture des unités relâchées à chaque PC, et qui correspond à la vraisemblance du modèle de CJS ; voir Seber (1982, page 196), Pollock *et al.* (1990), Kendall *et al.* (1995, page 4), Amstrup *et al.* (2005, page 44).

Procédure d'estimation : approche multinomiale

Sous l'hypothèse d'indépendance entre les unités et entre les PCs, les 7 historiques de capture suivent une loi multinomiale de paramètres n et de vecteur de probabilités défini par les fréquences prédites $\{\mu_\delta\}$ présentées à la table 1.5. La composante $L_1\left(\{u_i\}|\{U_i\},\{p_i\}\right)$ de la vraisemblance modélise le nombre d'unités capturées pour la première fois à chaque PC i, u_i, comme une fonction binomiale du nombre d'unités non marquées dans la population à chaque PC i : $u_i \sim \text{Bin}(U_i, p_i)$. Dans l'exemple à $I = 3$ PCs, L_1 s'écrit

$$L_1\left(\{u_i\}|\{U_i\},\{p_i\}\right) = \prod_{i=1}^{3} \binom{U_i}{u_i} p_i^{u_i} \left(1-p_i\right)^{U_i - u_i}. \tag{1.35}$$

La composante $L_2\left(\{n_\delta\}|\{u_i\},\{p_i\},\{\phi_i\}\right)$ modélise les recaptures à chaque PC i ; elle correspond à la fonction de vraisemblance de CJS, $L(\phi_1, p_2, r)$, donnée à l'équation (1.34) dans l'exemple avec $I = 3$ PCs. La procédure d'estimation des paramètres du modèle de JS peut être décrite en quelques étapes : (1) maximiser L_2 pour obtenir $\hat{p}_2, \hat{\phi}_1$, (2) estimer l'espérance du nombre d'unités non marquées dans la population à la PC 2, U_2, à partir de L_1, $\hat{U}_2 = u_2/\hat{p}_2$, (3) calculer un estimateur des moments du nombre d'unités marquées dans la population à la PC 2, M_2, $\hat{M}_2 = m_2/\hat{p}_2$, et (4) calculer l'estimateur de la taille de la population à la PC 2, N_2, $\hat{N}_2 = \hat{M}_2 + \hat{U}_2 = n_2/\hat{p}_2$. Les paramètres N_1, N_3 et ϕ_2 ne sont pas estimables ; seuls $N_1 p_1$, $N_3 p_3$ et $r = \phi_2 p_3$ le sont (Pollock *et al.*, 1990).

Procédure d'estimation : vraisemblance Poisson

Cormack (1985) et Cormack (1989) ont présenté une approche log-linéaire pour le modèle de JS. Cette méthode considère toutes les historiques de capture, observables ou non. La fréquence prédite μ_δ pour l'historique de capture δ s'exprime, sous forme log-linéaire, en fonction des paramètres N_i, ϕ_i et p_i,

$$\log \mu_\delta = \boldsymbol{Z_\delta}\gamma + \boldsymbol{X_\delta}\beta, \tag{1.36}$$

où $\boldsymbol{X_\delta}$ est l'élément ligne, de dimension $I \times 1$, de la matrice $\boldsymbol{X}$ pour l'historique $\boldsymbol{\delta}$; $\boldsymbol{X_\delta} = \boldsymbol{\delta}$ et $\beta_i = \log\left\{p_i/\left(1-p_i\right)\right\}$, $i = 1, 2, \ldots, I$. Le vecteur $\boldsymbol{Z_\delta}$ est l'élément ligne, de dimension $(2I-1) \times 1$, de la matrice $\boldsymbol{Z}$ pour l'historique δ. Il s'écrit

$$\boldsymbol{Z}_\delta = \left(\underbrace{1, \bar{\delta}_1, \bar{\delta}_1\bar{\delta}_2, ..., \prod_{j=1}^{I-1}\bar{\delta}_j}_{\boldsymbol{Z_{1\delta}}}, \underbrace{\bar{\delta}_I, \bar{\delta}_I\bar{\delta}_{I-1}, ..., \prod_{j=0}^{I-2}\bar{\delta}_{I-j}}_{\boldsymbol{Z_{2\delta}}} \right), \tag{1.37}$$

où $\bar{\delta}_i = 1 - \delta_i$. On a $\boldsymbol{Z}_{1\delta} = (1, 1, .., 1, 0, ..., 0)$ et la transition $1-0$ à la première PC, disons j, à laquelle l'unité est capturée ; $\boldsymbol{Z}_{2\delta} = (0, 0, .., 0, 1, ..., 1)$, avec la transition $0-1$ s'opérant à la dernière PC, disons k, à laquelle l'unité est capturée. Le paramètre $\boldsymbol{\gamma}$ s'exprime en fonction des paramètres ϕ_i, p_i, N_i and B_i définis à la Section 1.2.2. Par exemple, si la première et

la dernière capture pour l'historique ω se produisent aux PC j and k respectivement, alors $\boldsymbol{Z}_{\delta(\omega)}\boldsymbol{\gamma} = U_j \left\{\prod_{i=j}^{k-1}(1-p_i)\phi_i\right\}(1-p_k)\chi_k$, voir Rivest and Daigle (2004). Dans l'exemple à $I = 3$ occasions de capture, (1.36) devient

$$\log \mu_\delta = \gamma_0 + (1-\delta_1)\gamma_1 + (1-\delta_1)(1-\delta_2)\gamma_2 + (1-\delta_3)\gamma_3 + (1-\delta_2)(1-\delta_3)\gamma_4 + \sum_{j=1}^{3}\delta_j\beta_j,$$

où $\gamma_0 = U_1(1-p_1)\phi_1(1-p_2)\phi_2(1-p_3)$, $\log\{p_j/(1-p_j)\}$, $j = 1,2,3$, $\gamma_i = U_{i+1}/\{U_i(1-p_i)\phi_i\}$, $i = 1,2$ et $\gamma_i = \chi_{i-2}/\{(1-p_2)\phi_2\chi_{i-1}\}$, $i = 3,4$. La vraisemblance Poisson, à l'image de la vraisemblance multinomiale de JS, a deux composantes L_1 et L_2. La première composante est la version Poisson de (1.35) ; elle s'écrit

$$L_1\left(\{u_i\}|\{U_i\},\{p_i\}\right) \propto \prod_{i=1}^{3}\exp(-U_i p_i)(U_i p_i)^{u_i}. \tag{1.38}$$

La seconde composante $L_2\left(\{n_\delta\}|\{u_i\},\{p_i\},\{\phi_i\}\right)$ est identique à la vraisemblance du modèle de CJS donnée dans (1.34). La maximisation de la vraisemblance $L_1 \times L_2$, donnée par (1.38) et (1.34), aboutit à des formules explicites pour $\hat{p}_2, \hat{\phi}_1, \hat{r}$ (Rivest and Baillargeon, 2013) qui sont identiques aux estimateurs du maximum de vraisemblance obtenus par la maximisation de (1.34) ; l'estimateur de N_2 est identique à celui obtenu en utilisant l'approche multinomiale, $\hat{N}_2 = n_2/\hat{p}_2$. Le modèle log-linéaire souffre de problèmes d'identifiabilité : les matrices $\boldsymbol{X}$ et $\boldsymbol{Z}$ ne sont pas de plein rang. En effet, les paramètres $(\gamma_0, \gamma_1, \gamma_3, \beta_1, \beta_3)$ ne sont pas estimables ; seuls $\gamma_1 - \beta_1$ (N_1p_1), $\gamma_3 - \beta_3$ (N_3p_3) et $\gamma_0 + \beta_1 + \beta_3$ (ϕ_2p_3) le sont, voir Rivest and Daigle (2004). Pour plus de détails sur le calcul des estimateurs dans un cadre plus général, voir Pollock *et al.* (1990, Ch. 4).

Dans le cas $I = 10$, il y a $2^{10}-1 = 1023$ historiques de capture observables. L'écriture de (1.36) ainsi que la maximisation de la vraisemblance dans le cadre d'une régression de Poisson peut devenir difficile à effectuer avec les programmes informatiques standard. La décomposition des historiques de capture dans le cadre du modèle de CJS offre souvent une alternative intéressante : il y a $9 \times 10/2 = 45$ statistiques de recapture $\{m_{ij}\}$ à traiter dans l'écriture de la vraisemblance, comparativement aux 1023 historiques de capture dans la régression de Poisson.

Modèles à survie ou/et capture constantes

Le modèle classique de Jolly-Seber suppose que les paramètres d'intérêt (survie, probabilité de capture, taille de population, nouveaux arrivants) varient d'une PC à l'autre, impliquant un nombre important de paramètres à estimer. À moins d'un nombre assez conséquent de recaptures au cours de l'expérience, la précision autour des estimations peut être pauvre. Des restrictions, si appliquées sur la version originale du modèle, ont l'avantage de réduire le nombre de paramètres à estimer et améliorer la précision des estimations. Jolly (1982) a présenté trois modèles, B, C et D définis comme suit :

— Modèle B : ce modèle suppose que les survies sont constantes tout au long de l'expérience, $\phi_1 = \phi_2 = \cdots = \phi_{\ell-1}$. Pour plus de détails sur la procédure d'estimation, voir Jolly (1982, pages 4-6) ;

— Modèle C : ce modèle suppose des probabilités de capture constantes tout au long de l'expérience, $p_1 = p_2 = \cdots = p_\ell$. Jolly a présenté les équations d'estimation ainsi que des formules pour la variance asymptotique des estimateurs, voir Jolly (1982, pages 6-8) ;

— Modèle D : ce modèle suppose que les probabilités de survie et de capture sont constantes tout au long de l'expérience, voir Jolly (1982, pages 8-9) pour plus de détails.

L'obtention des estimateurs du maximum de vraisemblance pour ces trois modèles requiert des algorithmes itératifs décrits dans Jolly and Dickson (1980). Les modèles B et D sont implémentés dans le programme JOLLY.

Modélisation du processus des naissances

Schwarz and Arnason (1996) ont argumenté que la fonction de vraisemblance du modèle de population ouverte de Jolly-Seber n'est pas bien définie pour (principalement) deux raisons :

— Les naissances B_i ne sont pas directement inclus dans la fonction de vraisemblance. Puisque le nombre de nouveaux arrivants dans la population doit être toujours positif, la contrainte $\hat{B}_i \geq 0$ n'est pas facilement imposable ;

— Les contraintes sur les modèles à décès seulement (tous les $B_i = 0$) et à naissances seulement (tous les $\phi_i = 1$) ne sont pas facilement imposables.

Pour contourner ces difficultés, ils ont proposé la paramétrisation des naissances $B_0, B_1, \ldots, B_{I-1}$ par N, le nombre total d'unités distinctes disponibles pour l'expérience de capture-recapture, et $b_0, b_1, \ldots, b_{I-1}$, la proportion des unités qui entrent dans la population entre deux PCs et qui survivent jusqu'à la prochaine période ; B_0 représente le nombre d'unités dans la population juste avant la PC 1. Ils définissent ainsi une super-population constituée de N unités qui entrent dans l'expérience selon une distribution multinomiale,

$$(B_1, B_2, \ldots, B_{I-1}) \sim \mathrm{M}\left(N; b_1, \ldots, b_{I-1}\right).$$

La procédure d'échantillonnage est illustrée sur la figure (1.5).

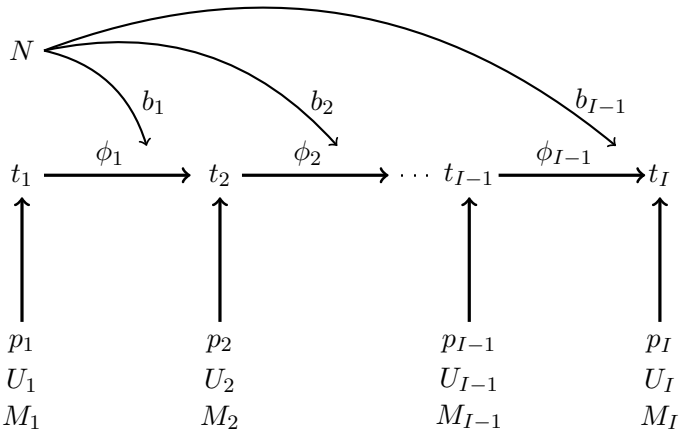

FIGURE 1.5 – La procédure d'échantillonnage décrivant la paramétrisation de Schwarz and Arnason (1996) avec une survie dépendant du temps. Les unités présentes dans la population au début de l'expérience (t_1) sont suivies jusqu'au temps de censure (t_I) correspondant au dernier temps d'observation.

Cette paramétrisation aboutit à la réécriture de la vraisemblance du modèle de JS Schwarz and Arnason (1996, pages 4-5). La maximisation de cette vraisemblance conduit aux estimateurs usuels de $\{\phi_i\}$, $\{N_i\}$, $\{B_i\}$.

1.3 Le design robuste

Le design robuste est une stratégie d'échantillonnage de long-terme constituée de I périodes primaires (PP), par exemple des années. À l'intérieur de chaque PP $i\,(i = 1, 2, \ldots, I)$, des sessions d'échantillonnage répétées sont menées sur ℓ_i périodes secondaires (PS), par exemple des journées consécutives. Les ℓ_i PS d'échantillonnage sont assez rapprochées dans le temps pour que l'hypothèse de fermeture de la population soit tenable. D'une PP à l'autre, on suppose que la population subit des changements (entrées et sorties d'unités) qui rendent vraisemblable l'hypothèse que la population est ouverte. La procédure d'échantillonnage est illustrée sur la figure (1.6).

Pollock (1982) a proposé de combiner les modèles de population fermée et de population ouverte pour estimer les paramètres démographiques. En particulier, il recommande d'utiliser l'information des PS pour estimer la taille de la population à chaque PP, $N_1, N_2, \ldots, N_I$. Le modèle de population ouverte de Jolly-Seber peut être ajusté aux données entre les PP pour estimer les probabilités de survie $\phi_1, \phi_2, \ldots, \phi_{I-1}$.

Cette section traite de l'estimation des paramètres démographiques dans le cas d'un design robuste : modèle et hypothèses, procédure d'estimation, calcul de la variance des estimateurs. L'accent est mis sur l'approche log-linéaire d'estimation par maximum de vraisemblance pré-

senté dans Rivest and Daigle (2004) .

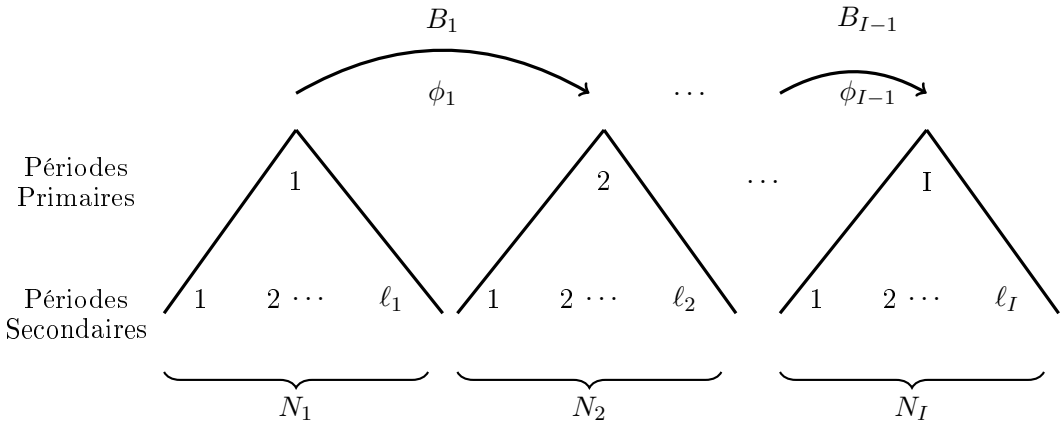

FIGURE 1.6 – La procédure d'échantillonnage décrivant un modèle de design robust à I périodes primaires (PP) et ℓ_i $(i = 1, 2, \ldots, I)$ périodes secondaires (PS) d'échantillonnage. Les tailles de la population $\{N_i\}$ sont estimées avec l'information intra période primaire. Les estimations des probabilités de survie $\{\phi_i\}$ et des naissances $\{B_i\}$ sont obtenues avec l'information inter période primaire.

1.3.1 Données et notation

Un historique de capture peut être résumé dans le vecteur $\boldsymbol{\omega}$ de dimension $(\sum_i \ell_i) \times 1$ contenant les éléments $\omega_{ij} = 1$ si l'unité est capturée à la PS j de la PP i, ℓ_i étant le nombre de PS à l'intérieur de la PP i. Ces éléments constituent l'information intra-période primaire ou information secondaire. L'information inter-période primaire est obtenue en "poolant" les ℓ_i occasions de capture relatives à la PP i ; on obtient l'historique de capture $\boldsymbol{\delta}(\boldsymbol{\omega}) = (\delta_1, \delta_2, \ldots, \delta_I)$, avec $\delta_i = 1$ si $\sum_i \omega_{ij} > 0$ i.e. l'unité a été capturée au moins une fois à l'intérieur de la PP i ; $p_i^* = \Pr(\delta_i = 1)$ est la probabilité qu'une unité soit capturée à l'intérieur de la PP i. Les modèles de design robuste sont désignés par M_α^β, où α représente la variation dans la capture des unités à l'intérieur d'une PP tandis que β représente la source de variation entre les PPs. Les sources de variation dans la capture des unités correspondent à celles présentées dans Otis *et al.* (1978) et discutées à la Section 1.1.5 ; par exemple, M_0^t signifie que les probabilités de capture sont homogènes à l'intérieur d'une PP mais varient dans le temps entre les PPs.

1.3.2 Procédure d'estimation : approche conditionnelle

Kendall *et al.* (1995) ont développé un cadre formel d'estimation des paramètres du design robuste. Ils ont développé des fonctions de vraisemblance pour une variété de modèles de population fermée à l'intérieur d'une PP. La vraisemblance pour les données complètes a trois composantes :

1. les composantes L_1 et L_2 relatives à l'information inter PP et présentées à la Section 1.2.2 dans le cadre du modèle de population ouverte de JS ;

2. une composante L_3 relative à l'information intra PP et qui modélise la capture des unités d'une PS à une autre ; elle est donnée dans l'équation (1) de Kendall *et al.* (1995).

Prenons l'exemple d'une expérience à $I = 3$ PP et à $\ell_i = 2$ PS à l'intérieur de chaque PP. La maximisation de la vraisemblance complète $L_1 \times L_2 \times L_3$ rend possible l'estimation des paramètres p_1^*, p_3^*, ϕ_2,N_1, N_3 grâce à l'information (supplémentaire) intra PP apportée par L_3. Kendall *et al.* (1995) ont proposé une méthode d'estimation en trois étapes : (1) maximiser $L_2 \times L_3$ pour obtenir les estimateurs du maximum de vraisemblance $\{\hat{p}_i^*; i = 1, 2, 3\}$ et $\{\hat{\phi}_i; i = 1, 2\}$, (2) estimer $\{U_i\}$ à partir de L_1 : $\hat{U}_i = u_i \hat{p}_i^*$ $(i = 1, 2, 3)$, (3) calculer un estimateur des moments pour $\{N_i\}$: $\hat{N}_i = \hat{U}_i + \hat{M}_i = u_i/\hat{p}_i^* + m_i/\hat{p}_i^*$ $(i = 1, 2, 3)$ et un estimateur des naissances $\hat{B}_i = \hat{N}_{i+1} - \hat{N}_i \hat{\phi}_i$ $(i = 1, 2)$. Les estimations des variances multinomiales asymptotiques des estimateurs $\{\hat{p}_i^*; i = 1, 2, 3\}$ et $\{\hat{\phi}_i; i = 1, 2\}$ sont obtenues avec le programme informatique SURVIV (Kendall *et al.*, 1995) ; pour les estimateurs $\hat{N}_i$ $(i = 1, 2, 3)$, les estimations des variances sont obtenues par bootstrap paramétrique.

1.3.3 Procédure d'estimation : vraisemblance Poisson

Rivest and Daigle (2004) ont proposé des modèles de population fermée à l'intérieur des PP pour lesquels la fréquence prédite μ_ω pour l'historique ω s'exprime sous forme log-linéaire. Les modèles M_0, M_t, M_h and M_{th}, présentés à la Section 1.1.4, répondent à ce critère. La matrice de design pour le modèle de design robuste a deux composantes : une composante $\boldsymbol{Z}$ qui modélise l'information inter-période primaire et une composante $\boldsymbol{X} = [\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_I]$, composée de I matrices, qui modélise l'information intra-période primaire. Ces matrices ont $2^{\sum_i \ell_i} - 1$ lignes chacune. Le modèle s'écrit

$$\log \mu_\omega = \boldsymbol{Z}_{\delta(\omega)} \boldsymbol{\gamma} + \boldsymbol{X}_{\boldsymbol{\omega}} \boldsymbol{\beta} = \boldsymbol{Z}_\delta \boldsymbol{\gamma} + \sum_{i=1}^{I} \boldsymbol{X}_{i,\omega} \beta_i, \tag{1.39}$$

où $\boldsymbol{X}_{i,\omega}$ représente l'élément ligne de $\boldsymbol{X}_i$ pour l'historique de capture ω ; la matrice de design $\boldsymbol{X}_i$ dépend du modèle choisi à l'intérieur de la PP i, soit M_0, M_t, M_h, ou M_{th}, et β_i contient les paramètres log-linéaires pour le modèle choisi à l'intérieur de la PP i. L'élément ligne de $\boldsymbol{Z} = [\boldsymbol{Z}_1, \boldsymbol{Z}_2]$ pour l'historique $\boldsymbol{\omega}$ dépend de $\delta(\omega) = (\delta_1, \ldots, \delta_I)$; il correspond à celui donné à la Section 1.2.2 sur le modèle de population ouverte de Jolly-Seber.

La procédure d'estimation des paramètres du modèle (1.39) et de leurs variances est décrite dans Rivest and Daigle (2004, page 4). Cependant, leur méthode d'estimation ne peut pas être utilisée pour des problèmes impliquant $I \geq 20$ occasions de capture puisque, dans le cadre d'une régression de Poisson, le vecteur qui sous-tend le modèle linéaire généralisé devient trop large pour être facilement manipulé avec les programmes informatiques standard. Yauck *et al.* (2018) ont proposé une procédure séquentielle d'estimation des paramètres du design robuste pour des modèles de population fermée à l'intérieur des PP appartenant à la famille des modèles présentés dans Rivest and Lévesque (2001). Cette question sera traitée au chapitre suivant.

Transition

Dans le chapitre suivant, on se place dans le contexte du design robuste, une méthode d'échantillonnage de capture-recapture présentée à la Section 1.3. On aborde le problème de l'estimation des paramètres du design robuste dans le cas d'un nombre suffisamment élevé d'occasions de capture. On propose une procédure séquentielle d'estimation des paramètres pour la classe des modèles de design robuste présentés à la Section 1.3.3 ; on propose également un estimateur de la variance des paramètres par bootstrap paramétrique. Ces résultats ont été appliqués à des données relatives aux activités des utilisateurs de téléphones intelligents qui ont visité un grand concessionnaire automobile basé aux États-Unis, et recueillies sur une période d'un an et démi.

Chapitre 2

Capture-recapture Methods for Data on the Activation of Applications on Mobile Phones

Résumé

Ce travail concerne l'analyse des données marketing sur l'activation d'applications sur les appareils mobiles. Chaque application a un numéro d'identification haché spécifique à l'appareil sur lequel elle a été installée. Ce numéro peut être enregistré par une plateforme à chaque activation de l'application. Les activations sur le même appareil sont liées ensemble en utilisant le numéro d'identification. En se focalisant sur les activations qui ont eu lieu dans une entreprise, on peut créer un ensemble de données de capture-recapture sur les appareils, c'est-à-dire les utilisateurs qui visitent l'entreprise : les unités sont propriétaires d'appareils mobiles et les occasions de capture sont par exemple les jours. Les techniques de capture-recapture peuvent être appliquées aux données d'activations pour estimer le nombre total d'utilisateurs qui ont visité l'entreprise sur une durée déterminée, donnant ainsi une estimation indirecte de sa fréquentation. Ce travail soutient que le design robuste, une méthode pour traiter des données à structure imbriquée, peut être utilisé dans ce contexte. Un nouvel algorithme d'estimation des paramètres du design robuste, avec un nombre assez important d'occasions de capture, ainsi qu'un bootstrap paramétrique pour l'estimation de leurs variances sont présentés. En sus, des résultats théoriques et méthodologiques sont introduits pour une plus large application du design robuste. Ces résultats sont mis en oeuvre à travers une étude sur des données d'activations relatives à des concessionnaires d'une grande entreprise automobile établies dans la zone métropolitaine des États-Unis sur une période d'un an et démi. Les développements techniques sont présentés dans le matériel supplémentaire disponible en ligne.

*M*ots-clé : Hétérogénéité, Modèle de Jolly-Seber, Régression de Poisson, Design robuste.

Abstract

This work is concerned with the analysis of marketing data on the activation of applications (apps) on mobile devices. Each application has a hashed identification number that is specific to the device on which it has been installed. This number can be registered by a platform at each activation of the application. Activations on the same device are linked together using the identification number. By focusing on activations that took place at a business location one can create a capture-recapture data set about devices, that is users, that "visited" the business: the units are owners of mobile devices and the capture occasions are time intervals such as days. A unit is captured when she activates an application, provided that this activation is recorded by the platform providing the data. Statistical capture-recapture techniques can be applied to the app data to estimate the total number of users that visited the business over a time period, thereby providing an indirect estimate of foot traffic. This article argues that the robust design, a method for dealing with a nested mark-recapture experiment, can be used in this context. A new algorithm for estimating the parameters of a robust design with a fairly large number of capture occasions and a simple parametric bootstrap variance estimator are proposed. Moreover, new estimation methods and new theoretical results are introduced for a wider application of the robust design. This is used to analyze a data set about the mobile devices that visited the auto-dealerships of a major auto brand in a US metropolitan area over a period of one year and a half. Technical developments are provided in the Supplementary Material available online.

*K*eywords: Heterogeneity, Jolly-Seber model, Poisson regression, Robust design.

2.1 Introduction

The mobile ecosystem is a dynamic environment involving many parties who want to profit from the actions of mobile device users. Four members of the mobile ecosystem are device users, publishers, advertisers, and supply side platforms. Users carry out daily activities such as checking the weather or reading the news by activating applications that they have downloaded on their mobile devices. An application is owned by a publisher who can charge third parties to display advertising. Publishers and advertisers meet through a supply side platform that provides the software to target a specific audience that the advertiser wants to reach. A good description is available in a recent report of the Interactive Advertising Bureau (IAB), a non-profit business organization that develops industry standards for this on-line market, see Gallo (2015).

The process of buying and selling mobile advertisements is a complex interaction between users, publishers, advertisers, and supply side platforms. Each time a user fires an application, an instantaneous auction (Real-Time Bidding) takes place for the opportunity to put advertisement on the device where it has been activated. A buyer is selected and the transaction between advertiser and publisher is recorded in a log that usually contains information about the ad impression, a hashed ID providing anonymized information on the application

and the device on which it has been fired, the location (latitude, longitude) of the device and the time when the ad was served. After the bidding process is over, the log data can be acquired by marketing platforms.

Marketing platforms thus accumulate a massive amount of impression data from across the United States on a daily basis. The re-purposing of these data creates opportunities for business and financial entities to extract information on the scale of the patterns and movement of people who attend entertainment venues, visit retail stores, or patronize restaurants. According to a recent report of Pew Research Center, about 80% of the US population owns a smartphone, see Smith (2017). Businesses are leveraging the ubiquity of smartphones to provide measurements of foot traffic visits to purchase locations in response to marketing events. Methodologies to analyze the location data resulting from mobile advertisements are discussed in Ivie (2017) and Smith (2015).

Once collected, the data undergo many processing steps, see (Ivie, 2017, ch. 5), to extract relevant fields and are stored in large scale distributed computing systems. Impressions originating from the same device and involving different apps are linked together using the IDs provided at the time of the transactions. Parts of the data are the result of fake human or robotic impressions that are removed either in the collection process or through fraud detection algorithms. Additional challenges include the measurement error of GPS location technology. By focusing on the activations that took place at a business location over a given time period, one can create a capture-recapture data set about persons that "visited" the business. The units are mobile devices, identified through applications whose advertisement opportunities are managed by supply side platforms from which the data have been bought, and the capture occasions are time intervals of fixed length. In the discussion in Section 2 and the example of Section 6, the capture occasions are calendar days.

Infering population sizes or foot traffic from data on app activations might seem risky. The identification of apps and their association with devices through ID numbers is subject to processing errors. Clearly the app data have limitations, however once the methodology is developed it can be tested on cases where the true population is known. For instance, validation can be attempted on publicly traded entertainment companies that operate amusement parks as they must publish their quarterly attendance. Alternatively, as proposed in Smith (2015, page 9), one can start with a baseline foot traffic estimate and tracks changes over time to assess the impact of marketing events; capture-recapture modelling can be used for that purpose.

Statistical techniques for capture-recapture experiments apply to app activation data. For a given capture occasion, the data consist of a partial list of the devices on which apps were fired at the location of interest. This is a new form of multi-list data. Section 6 considers data on the daily app activations at auto-dealerships of a major brand in a US metropolitan

area over more than 70 weeks in 2014 and 2015. The goal of the analysis is to measure the difference in foot traffic between 2014 and 2015, to provide weekly clientele estimates and to investigate their variation over the study period. In Section 2 we argue that, over a week, the assumption that the population is closed is tenable. However, over 70 weeks, we are facing an open population; new clients are births and decided clients are deaths. Thus Pollock's robust design should be used to analyze these data and to infer weekly population sizes. This is a fairly large model involving more than 300 parameters for the auto-dealership data and the standard methods for fitting a robust design, based on likelihood maximization, might fail. The goal of this paper is to provide a new algorithm for estimating the parameters of a robust design with an arbitrarily large number of capture occasions. To do so we generalize the seminal paper of Jolly (1965) to the robust design and propose a sequential estimation procedure. A simple parametric bootstrap variance estimator for the model parameters is also proposed. Even though this paper is motivated by the analysis of app data, it provides new estimation methods and new theoretical results that are of interest for more traditional applications of the robust design.

2.2 Capture-recapture modelling for activation data

Consider a population of N people contemplating visiting a business over a week. A person in the population is recorded in the activation data set if three conditions are met: (i) the person actually visits the business during the week, (ii) she activates an app on her mobile device while on location and (iii) this activation is recorded by the supply side platform providing the data. The probability of such an event involves the probability of visiting the business on a given day, the conditional probability of firing an app while in the business precinct and finally the conditional probability that the activation is recorded. If these three conditional probabilities are constant over time and over population units, then the number of captures for a customer could be modeled with a binomial distribution with a constant probability of success (this defines capture-recapture model M_0) .

The assumption of homogeneity is likely to be violated. One might expect the visit probability and that to activate an application to vary between persons in the population. The visit probability should depend on a person's interest for the product sold by the business while the app activation probability may depend on socio-demographic variables. Thus an heterogeneity in capture probability, leading to a binomial distribution with a random probability of success, should apply (this defines capture-recapture model M_h). A time effect with large capture probabilities on traditional shopping days such as Friday and Saturday is also likely; capture probabilities that vary from one occasion to the next and that are constant across units define model M_t. Arguing as in Rivest (2008), a time effect is indirectly accounted for in a mixed binomial model with a random probability of success and model M_h applies in this context. With a mixed binomial model, population sizes are difficult to estimate as models fitting the

data equally well can give drastically different estimates. This is well documented in the literature, see Link (2003), Hwang and R. Huggins (2005), or Rivest and Baillargeon (2013). A person not owning a mobile device cannot be captured; this is known as a hidden population in the capture-recapture literature, see Millar *et al.* (2008). These units are missed and the app activation data set is likely to underestimate the size of a business clientele.

The process of buying durable goods, such as a car, can be described by an open population model. Customers enter the population when they feel that the time has come to buy a product. Then they gather information by visiting dealerships and getting quotes from various sources. They leave the population when they make up their mind and either buy a product or elect not to buy it. A closed population is plausible at a weekly level, while at a monthly or yearly scale one is facing an open population; thus the app activation data could possibly be modeled with the so-called robust design having a hierarchical structure: the capture occasions are days that are nested within primary sampling periods (SP) that are weeks. This is a working assumption that is not completely true: customers do not necessarily enter the population at the beginning of a week or leave on the weekend. The robustness of the population size estimators to within primary period arrivals and departures is tested in simulations reported in Section 5.

A robust design model allows weekly capture probabilities to be estimated from two sources, either the closed population model for the weekly data or the open population model for the between week data. It then combines closed population and open population models to give weekly estimators of population sizes with a smaller sampling variance than closed population models. The robust design has several parameters; in this work the focus is on the expected population sizes for every SP as they provide information about foot traffic.

2.3 Modelling of the robust-design data and parameter estimation

In a capture-recapture experiment a unit is observed if it is captured at least once. Its capture history is defined as a vector $\omega = (0, 1, 1, 0, ..., 1)$, a sequence of 1's (capture) and 0's (miss) that contains the capture information for all the secondary capture occasions. This section summarizes the notation used to describe this data set and the model for its analysis.

2.3.1 Notation and assumptions

We distinguish (1) the notation used to describe the data themselves, (2) important summary statistics and (3) the robust design parameters for its analysis.

Capture-recapture data

— Subscript i denotes primary sampling period i, $i = 1, \ldots, I$;

— Subscript j denotes a secondary capture occasion within a SP, $j = 1, ..., \ell_i$; ℓ_i is the number of secondary capture occasions within SP i;

— $\boldsymbol{\omega}$ is a $(\sum_i \ell_i) \times 1$ capture history vector of secondary capture occasions; $\omega_{ij} = 1$ if the unit has been captured on occasion j in SP i and 0 otherwise;

— The between SP capture information can be summed up in a $I \times 1$ vector $\boldsymbol{\delta} = \boldsymbol{\delta}(\boldsymbol{\omega})$ which has entry $\delta_i = 1$ if the unit has been captured at least once during SP i, that is if $\sum_j \omega_{ij} > 0$, and $\delta_i = 0$ otherwise.

Statistics

— n_ω is the frequency of units with capture history $\boldsymbol{\omega}$; the frequency $n_{00...0}$ for the unobserved capture history $\omega = (0, ..., 0)$ is unknown;

— u_i is the number of unmarked units captured during SP i;

— m_i represents the number of marked units recaptured during SP i;

— $n_i = u_i + m_i$ is the number of units captured during SP i;

— v_i is the number of units captured for the last time at SP i;

— $w_i = \sum_{s=1}^{i-1}(u_s - v_s)$ is the number of units captured at least once during the first $i-1$ SPs that will be seen at least once more, either in SP i or later;

— n is the total number of units captured at least once during the whole sampling process.

Parameters

— The survival probability between primary periods i and $i+1$ is denoted $\phi_i \in (0, 1)$ for all units in the population;

— The probability of being captured during SP i is denoted $p_i^* = Pr(\delta_i = 1)$, this depends on the closed population model describing the captures within SP i;

— The probability of not being seen after SP i, χ_i, satisfies the following recursive relationship $\chi_i = (1 - \phi_i) + \phi_i(1 - p_{i+1}^*)\chi_{i+1}$ and $\bar{\chi}_i = 1 - \chi_i$;

— $N_i, i = 1, \ldots, I$ is the expected population at the start of the i^{th} SP ;

— B_i is the expected number of new units joining the population before the start of SP $i+1$ such that $N_{i+1} = N_i\phi_i + B_i$;

— U_i is the expected number of unmarked units in the population just before SP i; it satisfies $U_i = U_{i-1}\left(1 - p_{i-1}^*\right)\phi_{i-1} + B_{i-1}$ for $i = 1, ..., I-1$ and $U_1 = N_1$;

— M_i is the expected number of marked units in the population just before SP i such that $M_i = (M_{i-1} + U_{i-1}p_{i-1}^*)\phi_{i-1}$ and $M_1 = 0$;

— $\eta_i = 1 - M_i/N_i$ is the theoretical proportion of unmarked units just before SP i and $\bar{\eta}_i = 1 - \eta_i$.

Following Cormack (1989), we assume that the counts n_ω have independent Poisson distributions with mean value μ_ω,

$$n_\omega \sim \text{Poisson}\left(\mu_\omega\right).$$

We distinguish random variables and parameters ; e.g. N_i is the expected population size at the start of the i^{th} SP , a parameter to be estimated ; $\tilde{N}_i$ is the corresponding random Poisson variable. For the asymptotic variance derivations, N_i is assumed to be large for every SP i while I and $\{\ell_i\}$ are fixed.

2.3.2 Model building

Following Rivest and Daigle (2004), we focus on closed population models for the within SP data that lead to log-linear models for μ_ω. Standard closed population models, such as M_0, M_t, M_h and M_{th}, satisfy this requirement, see Rivest and Lévesque (2001). They lead to a design matrix for the robust design model that has a between SP component, $\mathbf{Z}$, and I within SP matrices $\boldsymbol{X} = [\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_I]$. These matrices have $2^{\sum_i \ell_i} - 1$ rows, one less than the actual number of capture history because the unobserved capture history $\omega = (0, ..., 0)$ is omitted. The predicted frequency μ_ω then satisfies:

$$\log\left(\mu_\omega\right) = \boldsymbol{Z}_{\delta(\omega)}\boldsymbol{\gamma} + \boldsymbol{X}_{\boldsymbol{\omega}}\boldsymbol{\beta} = \boldsymbol{Z}_{\delta}\boldsymbol{\gamma} + \sum_{i=1}^{I} \boldsymbol{X}_{i,\omega}\beta_i, \tag{2.1}$$

where $\boldsymbol{X}_{i,\omega}$ is the row of $\boldsymbol{X}_i$ for capture history ω; the design matrix $\boldsymbol{X}_i$ depends on the within SP model, either M_0, M_t, M_h, or M_{th}, selected for the captures at the ith SP while β_i contains the log-linear parameters for the ith SP model. The row of $\boldsymbol{Z} = [\boldsymbol{Z}_1, \boldsymbol{Z}_2]$ for any $\boldsymbol{\omega}$ depends only on $\delta(\omega) = (\delta_1, \ldots, \delta_I)$. It is given by

$$\boldsymbol{Z}_{\delta(\omega)} = \left(\underbrace{1, \bar{\delta}_1, \bar{\delta}_1\bar{\delta}_2, ..., \prod_{j=1}^{I-1} \bar{\delta}_j}_{\boldsymbol{Z_{1\delta}}}, \underbrace{\bar{\delta}_I, \bar{\delta}_I\bar{\delta}_{I-1}, ..., \prod_{j=0}^{I-2} \bar{\delta}_{I-j}}_{\boldsymbol{Z_{2\delta}}} \right), \tag{2.2}$$

where $\bar{\delta}_i = 1 - \delta_i$. On has $\boldsymbol{Z}_{1\delta} = (1, 1, .., 1, 0, ..., 0)$ and the switch from 1 to 0 occurs at the first SP, say j, at which the unit is captured while $\boldsymbol{Z}_{2\delta} = (0, 0, .., 0, 1, ..., 1)$, with the switch from 0 to 1 occuring at the last SP, say k, at which the unit is captured. The parameter $\boldsymbol{\gamma}$ is expressed in terms of the parameters ϕ_i, p_i^*, N_i and B_i defined in Section 3.1. If the first and last captures of ω occur at SPs j and k respectively, then $\boldsymbol{Z}_{\delta(\omega)}\boldsymbol{\gamma} = U_j \left\{\prod_{i=j}^{k-1}(1 - p_i^*)\phi_i\right\}(1 - p_k^*)\chi_k$; see Rivest and Daigle (2004) for additional details.

2.3.3 Estimating equations

Since demographic parameters are estimated using a Poisson regression, the estimators are calculated by solving the equation that sets the score function equal to 0. This gives the

following estimating equations (McCullagh and Nelder, 1989, p. 41),

$$(\boldsymbol{Z}, \boldsymbol{X})^\top (\boldsymbol{n} - \boldsymbol{\mu}) = 0, \tag{2.3}$$

where $\boldsymbol{n}$ and $\boldsymbol{\mu}$ are $\left(2^{\sum_i \ell_i} - 1\right) \times 1$ vectors with respective elements n_ω and μ_ω;$\boldsymbol{Z}^\top \boldsymbol{n}$ and $\boldsymbol{X}^\top \boldsymbol{n}$ are the sufficient statistics for this problem. Equation (2.3) leads to the following system of equations:

$$\begin{aligned}
\sum_\omega Z^i_{1\delta(\omega)} n_\omega &= \sum_\omega Z^i_{1\delta(\omega)} \mu_\omega \quad i = 1, \ldots, I \\
\sum_\omega Z^i_{2\delta(\omega)} n_\omega &= \sum_\omega Z^i_{2\delta(\omega)} \mu_\omega \quad i = 1, \ldots, I-1 \\
\sum_\omega \boldsymbol{X}^\top_{i,\omega} n_\omega &= \sum_\omega \boldsymbol{X}^\top_{i,\omega} \mu_\omega \quad i = 1, \ldots, I,
\end{aligned}$$

where $Z^i_{1\delta(\omega)}$ and $Z^i_{2\delta(\omega)}$ respectively represent the ith element of the vectors $\boldsymbol{Z}_{1\delta(\omega)}$ and $\boldsymbol{Z}_{2\delta(\omega)}$. In terms of the statistics u_i, v_i, n defined in Section 3.1, one has

$$\begin{aligned}
\sum_\omega Z^1_{1\delta(\omega)} n_\omega &= n \\
\sum_\omega Z^i_{1\delta(\omega)} n_\omega &= n - \sum_{s=1}^{i} u_s \quad i = 2, \ldots, I \\
\sum_\omega Z^i_{2\delta(\omega)} n_\omega &= n - \sum_{s=1}^{i} v_s \quad i = 1, \ldots, I-1.
\end{aligned}$$

Thus, the sufficient statistics for the between SP component of the model are $\{(u_i, v_i) : i = 1, \ldots, I\}$ with dimension $2I - 1$ as both u_i and v_i sum to n. To finalize the estimating equations used in the sequel, observe that $w_i = \sum_{s=1}^{i-1}(u_s - v_s)$ represents the units that have been captured before SP i and that will be recaptured at SP i or later. Its expectation given $u_j, j = 1, \ldots, i-1$ is equal to $M_{i|u} = \sum_{s=1}^{i-1} u_s \prod_{k=s}^{i-1} \phi_k$, the conditional expectation of the units that are marked before session i and available for capture at SP i or later, times $1-(1-p^*_i)\chi_i$, the probability that a unit bearing a mark just before SP i will be captured once again, either during SP i or later. In other words $E(w_i|u_1, \ldots, u_{i-1}) = M_{i|u}\{1 - (1 - p^*_i)\chi_i\}$. This leads to the following estimating equations:

$$\begin{aligned}
v_i &= (M_{i|u} p^*_i + u_i)\chi_i \quad i = 1, \ldots, I, && (2.4) \\
w_i &= M_{i|u}\{1 - (1 - p^*_i)\chi_i\} \quad i = 2, \ldots, I && (2.5) \\
\sum_\omega \boldsymbol{X}^\top_{i,\omega} n_\omega &= \sum_\omega \boldsymbol{X}^\top_{i,\omega} \mu_\omega, \quad i = 1, \ldots, I && (2.6)
\end{aligned}$$

In order to calculate the Poisson maximum likelihood estimators for the parameters of the robust design, we need to solve simultaneously these three equations. In Kendall *et al.* (1995) and in Rivest and Daigle (2004) all parameters are estimated simultaneously by maximizing either a combined likelihood or the full Poisson likelihood. In problems where I, the number of SPs, is large this can be problematic. An alternative sequential strategy is proposed here. It is first implemented in the next section when the model within primary session is M_0.

2.4 The robust-design with model M_0 within primary sessions

This section assumes that units are captured according to model M_0 within primary sessions. It suggests a recursive algorithm to solve the estimating equations (2.3) for the parameters of the robust design. Closed form expressions for the asymptotic variances of the maximum likelihood estimators are given and efficiency comparisons with respect to the estimators obtained with the Jolly-Seber model, that has a single secondary capture occasion within each SP, are presented.

2.4.1 A simple recursive algorithm for calculating the maximum likelihood estimates

Let $p_i = \exp(\beta_i)/\{1+\exp(\beta_i)\}$ stand for the probability of capture at one occasion within SP i; then $p_i^* = 1-(1-p_i)^{\ell_i}$ and the design matrix $\boldsymbol{X_i}$ for SP i has a single column whose entry, for capture history $\boldsymbol{\omega}$, is $\sum_j \omega_{ij}$. Estimating equation (2.6) involves $\sum_\omega \sum_j \omega_{ij} n_\omega = C_i$, the total number of captures within SP i, whose expectation is $N_i \ell_i p_i$. For $i = 1, \ldots, I$ the following results hold. Since $N_i = E(M_{i|u} + u_i/p_i^*)$, equation (2.6) is equivalent to

$$n_i^* = C_i p_i^*/(\ell_i p_i) \quad = \quad M_{i|u} p_i^* + u_i. \tag{2.7}$$

Equations (2.4) and (2.7) give $\chi_i = v_i/\tilde{n}_i$. From (2.5) and (2.7)

$$w_i = M_{i|u}\{1-(1-p_i^*)v_i/n_i^*\} = M_{i|u}\{1-v_i/n_i^*\} + (n_i^*-u_i)v_i/n_i^*.$$

This gives the following expression for $M_{i|u}$

$$M_{i|u} = n_i^* - u_i + n_i^*(w_i - n_i^* + u_i)/(n_i^* - v_i). \tag{2.8}$$

From (2.7), $n_i^* = p_i^*\{n_i^* - u_i + n_i^*(w_i - n_i^* + u_i)/(n_i^* - v_i)\} + u_i$. This estimating equation depends on a single parameter, N_i, as one can set $p_i = C_i/(N_i \ell_i)$. It is easily transformed into an estimating equation for N_i, $f_{i,N}(N_i) = 0$, where

$$f_{i,N}(N_i) = N_i\left[n_i^*(w_i - n_i^* + u_i)/\left\{n_i^*(w_i - v_i) + u_i v_i\right\} - \{1 - C_i/(\ell_i N_i)\}^{\ell_i}\right]. \tag{2.9}$$

The algorithm to estimate the parameters is as follows. For $i = 1, \ldots, I$ do the following:

1. Solve (2.9) to calculate the maximum likelihood estimator for N_i and let $\hat{p}_i$, $\hat{p}_i^*$, $\hat{M}_i$ and $\hat{n}_i$ be the estimators obtained by plugging in $\hat{N}_i$ in the expressions for p_i, p_i^*, $M_{i|u}$ and n_i^*. Note that $\hat{M}_1 = 0$ and $\hat{M}_I = (n_I - u_I)/\hat{p}_I^*$;
2. For $i > 1$, estimate the survival probability between SP $i-1$ and i, ϕ_{i-1}, by

$$\hat{\phi}_{i-1} = \hat{M}_i/\{\hat{M}_{i-1} + u_{i-1}\} \tag{2.10}$$

and set $\hat{\phi}_{i-1} = 1$ if this quantity is larger than 1.

3. The number of new recruits between SP $i-1$ and i, B_{i-1}, is estimated by $\hat{B}_i = \hat{N}_i - \hat{N}_{i-1}\hat{\phi}_{i-1}$, set the estimate to 0 if this quantity is negative.

In this algorithm, the estimator for ϕ_{i-1} is derived by noting that, from the definition of $M_{i|u}$ given in Section 3.3, $M_{i|u} = (M_{i-1|u} + u_{i-1})\phi_{i-1}$. In the above algorithm, out of range estimates for ϕ_i and B_i are brought back to the boundary of the parameter space without impacting the values of the in range estimates.

2.4.2 Variance calculations

This section presents closed form expressions for the asymptotic variances of $\hat{N}_i$ and $\hat{\phi}_i$ defined through (2.9) and (2.10). They show explicitly the impact of the secondary sampling occasions on the asymptotic variances of the estimators. They include as special cases the Jolly-Seber variances (Jolly, 1965), when $\ell_i = 1$ for all SPs, and to the closed population variances of Darroch (1958). The following notation is used throughout this section.

— $P_{2i} = 1 - (1-p_i)^{\ell_i} - \ell_i p_i (1-p_i)^{\ell_i - 1}$ is the probability of being captured at least twice at SP i ,

— $p_{1i} = \ell_i p_i (1-p_i)^{\ell_i - 1}$ is the probability of being captured once at SP i,

— $D_i = 1 - (1-p_i^*)\chi_i\eta_i - \bar{\eta}_i\bar{\chi}_i$ is the probability, for an unmarked unit, to be captured at least once at SP i or later and, for a marked unit, to be captured for the last time at SP i,

and note that $p_i^* = P_{2i} + p_{1i}$. Following Jolly (1965) the asymptotic variance of $\hat{N}_i$ is evaluated as the variance of $\hat{N}_i$ under Poisson sampling, minus N_i, its Poisson variance.

Proposition 1 *In a robust design with M_0 within SP, the asymptotic variances of $\hat{N}_i$ and $\hat{\phi}_i$ are given by*

$$var\left(\hat{N}_i\right) = \frac{N_i D_i (1-p_i^*)}{p_i^* \bar{\chi}_i \bar{\eta}_i + D_i P_{2i}} \tag{2.11}$$

$$\begin{aligned} var(\hat{\phi}_i) &= \phi_i^2 \left\{ \frac{(1-p_{i+1}^*)\chi_{i+1}\left\{\bar{\eta}_{i+1} + p_{i+1}^*\eta_{i+1}\right\}}{N_{i+1}p_{i+1}^*\bar{\eta}_{i+1}\bar{\chi}_{i+1}} + \frac{(1-p_i^*)\bar{\eta}_i\chi_i}{N_i p_i^* \bar{\chi}_i(\bar{\eta}_i + \eta_i p_i^*)} + \frac{1-\phi_i}{N_{i+1}\bar{\eta}_{i+1}} \right\} \\ &- \phi_i^2 \left\{ \frac{(1-p_{i+1}^*)(\bar{\eta}_{i+1} + p_{i+1}^*\eta_{i+1})^2 \chi_{i+1}^2 P_{2,i+1}}{N_{i+1}\bar{\eta}_{i+1}\bar{\chi}_{i+1}p_{i+1}^*(\bar{\eta}_{i+1}\bar{\chi}_{i+1}p_{i+1}^* + D_{i+1}P_{2,i+1})} + \frac{(1-p_i^*)\bar{\eta}_i\chi_i^2 P_{2,i}}{N_i\bar{\chi}_i p_i^*(\bar{\eta}_i\bar{\chi}_i p_i^* + D_i P_{2i})} \right\}. \end{aligned} \tag{2.12}$$

With a single capture occasion within each SP, $P_{2i} = 0$, and (2.11) reduces to $N_i D_i (1-p_i^*)/(p_i^*\bar{\chi}_i\bar{\eta}_i)$, the formula for the Jolly-Seber variance presented in equation (2.8) of Jolly (1965). When either η_i or χ_i is equal to 1, equation (2.11) reduces to the variance of $\hat{N}_i$ for closed population model M_0, $\mathrm{var}_{CP}(\hat{N}_i) = N_i(1-p_i^*)/P_{2i}$, see Rivest and Lévesque (2001).

Then the other SPs do not contribute to the estimation of N_i. In (2.12), the first line gives the variance under a Jolly-Seber model while the second one gives the variance reduction attributable to the robust design.

2.4.3 Derivation of Proposition 1

This section sketches the proof of Proposition 1. Additional details are given in the Supplementary Material. First we consider $\hat{N}_i$. The estimating equation for this parameter is (2.9); it involves four statistics, u_i, v_i, w_i, C_i. The standard approximation to its variance obtained by linearization is

$$\text{var}\left(\hat{N}_i\right) = \left[\text{var}\left\{f_{i,N}(N_i)\right\}/A_i^2\right] - N_i, \tag{2.13}$$

where $\text{var}\left\{f_{i,N}(N_i)\right\}$ is the approximation, obtained by linearization, to the asymptotic variance of $f_{i,N}(N_i)$ seen as a function of u_i, v_i, w_i, C_i and A_i is the limit, in probability, of the derivative of $f_{i,N}(N_i)$ with respect to N_i. In the Derivations (A.1) of the Supplementary Material it is shown that

$$A_i = -(p_i^* \bar{\chi}_i \bar{\eta}_i + D_i \times P_{2i})/(\bar{\eta}_i \bar{\chi}_i). \tag{2.14}$$

In (2.13) one has $\text{var}\left\{f_{i,N}(N_i)\right\} \approx \nabla f_{i,N}^\top \Sigma_i \nabla f_{i,N}$, where Σ_i, the covariance matrix of u_i, v_i, w_i, C_i. In the Derivations (A.2) of the Supplementary Material it is shown to be equal to

$$\Sigma_i = N_i \begin{pmatrix} \eta_i p_i^* & \eta_i p_i^* \chi_i & 0 & \eta_i \ell_i p_i \\ \eta_i p_i^* \chi_i & p_i^* \chi_i & \bar{\eta}_i p_i^* \chi_i & \ell_i p_i \chi_i \\ 0 & \bar{\eta}_i p_i^* \chi_i & \bar{\eta}_i(\bar{\chi}_i + p_i^* \chi_i) & \bar{\eta}_i \ell_i p_i \\ \eta_i \ell_i p_i & \ell_i p_i \chi_i & \bar{\eta}_i \ell_i p_i & \ell_i p_i (1 - p_i + \ell_i p_i) \end{pmatrix}. \tag{2.15}$$

Now $\nabla f_{i,N}$ is the limit of the vector of partial derivatives of $f_{i,N}(N_i)$ with respect to u_i, v_i, w_i, C_i ,

$$\nabla f_{i,N} = \left\{1/(\bar{\chi}_i \bar{\eta}_i)\right\} \times \left(\bar{\chi}_i + p_i^* \chi_i, \bar{\eta}_i(1 - p_i^*), p_i^*, -p_{i1} D_i/(p_i \ell_i)\right)^\top,$$

and (2.11) is proved by evaluating (2.13).

Now we consider $\hat{\phi}_i$. Using equation (2.10), the survival probability ϕ_i can be expressed as

$$\phi_i = N_{i+1} \bar{\eta}_{i+1} / \left\{N_i(\bar{\eta}_i + \eta_i p_i^*)\right\}. \tag{2.16}$$

The variance of $\hat{\phi}_i$ involves the computation of the variance of $\hat{M}_i$ which, in turn, depends on the variance of $\hat{n}_i$; the calculations are presented in the Derivations (B) of the Supplementary Material. The variance of $\hat{\phi}_i$ calculated by linearization is,

$$\text{var}(\hat{\phi}_i) = \nabla \hat{\phi}_i^\top \Gamma_i \nabla \hat{\phi}_i \tag{2.17}$$

where $\nabla\hat{\phi}_i$, the limit of the vector of partial derivatives of $\hat{\phi}_i$ with respect to $\hat{M}_{i+1}, \hat{M}_i, u_i$, is given by

$$\nabla\hat{\phi}_i = \lim \frac{1}{N_i(\bar{\eta}_i + \eta_i p_i^*)} \left(1, -\frac{N_{i+1}\bar{\eta}_{i+1}}{N_i(\bar{\eta}_i + \eta_i p_i^*)}, -\frac{N_{i+1}\bar{\eta}_{i+1}}{N_i(\bar{\eta}_i + \eta_i p_i^*)}\right)^\top = \frac{(1, -\phi_i, -\phi_i)^\top}{N_i(\bar{\eta}_i + \eta_i p_i^*)},$$

and Γ_i, the covariance matrix between $\hat{M}_{i+1}, \hat{M}_i, u_i$ is shown in the Supplementary Material to be

$$\Gamma_i = \begin{pmatrix} \text{var}(\hat{M}_{i+1}) + N_{i+1}\bar{\eta}_{i+1} & N_i\bar{\eta}_i\phi_i & N_i\eta_i p_i^*\phi_i \\ & \text{var}(\hat{M}_i) + N_i\bar{\eta}_i & 0 \\ & & N_i\eta_i p_i^* \end{pmatrix}, \tag{2.18}$$

where

$$\text{var}(\hat{M}_i) = N_i(1-p_i^*)\bar{\eta}_i\chi_i(\bar{\eta}_i + p_i^*\eta_i)(p_i^*\bar{\eta}_i + P_{2i}\eta_i)/\left\{p_i^*(p_i^*\bar{\chi}_i\bar{\eta}_i + D_i \times P_{2i})\right\} \tag{2.19}$$

is the variance of $\hat{M}_i$ as a predictor for the unobserved number of marked units just before SP i. It is $\text{var}(\hat{M}_i|M_i)$ in Jolly (1965) notation.

The next section generalizes the results of Section 2.4.1 to model M_t within each SP. This assumes that the capture probabilities vary across capture occasions.

2.4.4 An extension to model M_t within primary sessions

With M_t within SP, the sufficient statistics $\sum_\omega \boldsymbol{X}_{i,\omega}^\top n_\omega$ in (2.6) are $\{n_{ij} : \ j = 1, \ldots, \ell_i\}$, where n_{ij} is the number of units caught during occasion j of SP i. An estimating function for p_{ij}, the probability of being captured at occasion j, is $p_{ij} = n_{ij}/N_i$. For this model $p_i^* = 1 - \prod_{j=1}^{\ell_i}(1-p_{ij})$ with corresponding estimating function $p_i^* = 1 - \prod_{j=1}^{\ell_i}(1 - n_{ij}/N_i)$.

For M_t, equation (2.7) becomes

$$\tilde{n}_i = C_i \frac{1 - \prod_{j=1}^{\ell_i}(1 - n_{ij}/N_i)}{\sum_{j=1}^{\ell_i} n_{ij}/N_i} = M_i p_i^* + u_i. \tag{2.20}$$

As in Section (2.4.1), the maximum likelihood estimator for N_i is now obtained by solving the equation $f_{i,N}^*(N_i) = 0$ where

$$f_{i,N}^*(N_i) = N_i \left[n_i^*(w_i - \tilde{n}_i + u_i)/\left\{n_i^*(w_i - v_i) + u_i v_i\right\} - \prod_{j=1}^{\ell_i}(1 - n_{ij}/N_i) \right]. \tag{2.21}$$

Once N_i is estimated, the recursive algorithm of Section 2.4.1 applies. The generalization of Proposition 1 to model M_t within sessions uses the following expression for the probability of being captured more than once during SP i

$$P_{2i}^* = 1 - \prod_{j=1}^{\ell_i}(1-p_{ij}) - \sum_{j=1}^{\ell_i}\left\{ p_{ij}\prod_{\substack{s=1 \\ s\neq j}}^{\ell_i}(1-p_{is}) \right\}.$$

Proposition 2 *With model M_t within primary sessions, the asymptotic variance of the robust design estimators $\hat{N}_i$ and $\hat{\phi}_i$ are given by (resp.) equations (2.11) and (2.12) with P_{2i} replaced by P_{2i}^*, the corresponding value for model M_t.*

Further details of the calculations that lead to the derivation of Proposition 2 are given in the Derivations (C and D) of Supplementary Material. The next section relaxes the equal catchability between units assumption and allows the modelling of heterogeneity in capture probability through a within SP model denoted M_h.

2.5 Model M_h within primary sessions

This section assumes heterogeneity in capture probability within each SP that calls for model M_h. A recursive algorithm is used to solve the estimating equations (2.3) and a parametric bootstrap is used to compute the variance of the estimated parameters.

2.5.1 Estimation procedure

When the number Y of captures follows a binomial mixture distribution, we propose using the log-linear models in Rivest and Baillargeon (2007). The probability of having k captures in SP i is, for $i = 0, \ldots, \ell_i$,

$$p_{ik} = \binom{\ell_i}{k} \exp\{\beta_i k + \tau_i \psi(k)\} \Big/ \sum_{j=0}^{\ell_i} \binom{\ell_i}{j} \exp\{\beta_i j + \tau_i \psi(j)\}, \tag{2.22}$$

where $\psi(j)$ is a non-negative convex function satisfying $\psi(0) = 0$ and (β_i, τ_i) are unknown log-linear parameters. Then the probability of at least one capture is

$$\begin{aligned} p_i^* &= 1 - p_{i0} \\ &= \sum_{j=1}^{\ell_i} \binom{\ell_i}{j} \exp\{\beta_i j + \tau_i \psi(j)\} \Big/ \sum_{j=0}^{\ell_i} \binom{\ell_i}{j} \exp\{\beta_i j + \tau_i \psi(j)\}. \end{aligned} \tag{2.23}$$

The design matrix $\boldsymbol{X}_i$ for SP i has two column vectors whose entries, for capture history ω, are $(\sum_j \omega_{ij}, \psi(\sum_j \omega_{ij}))$, with corresponding sufficients statistics $C_i = \sum_j i f_{ij}$ and $C_{\psi_i} = \sum_{j=1}^{\ell_i} \psi(j) f_{ij}$, where f_{ij} is the number of units caught j times during SP i.

For M_h, (2.6) has 2 estimating equations. Parameters (β_i, τ_i) are estimated by solving two equations. The first one is (2.9) with p_i^* given by (2.23) and

$$n_i^* = C_i \sum_{j=1}^{\ell_i} \binom{\ell_i}{j} \exp\{\beta_i j + \tau_i \psi(j)\} \Big/ \sum_{j=0}^{\ell_i} j \binom{\ell_i}{j} \exp\{\beta_i j + \tau_i \psi(j)\}.$$

The second one is simply

$$\sum_{j=1}^{\ell_i} j f_{ij} \Big/ \sum_{j=1}^{\ell_i} \psi(j) f_{ij} = \sum_{j=0}^{\ell_i} j \binom{\ell_i}{j} \exp\{\beta_i j + \tau_i \psi(j)\} \Big/ \sum_{j=0}^{\ell_i} \psi(j) \binom{\ell_i}{j} \exp\{\beta_i j + \tau_i \psi(j)\}.$$

Once the estimates $(\hat{\beta}_i, \hat{\tau}_i)$ are available one calculates

$$\hat{N}_i = C_i \frac{\sum_{j=0}^{\ell_i} \binom{\ell_i}{j} \exp\left\{\hat{\beta}_i j + \hat{\tau}_i \psi(j)\right\}}{\sum_{j=0}^{\ell_i} j \binom{\ell_i}{j} \exp\left\{\hat{\beta}_i j + \hat{\tau}_i \psi(j)\right\}}.$$

The other parameters are then estimated as in Section 2.4.1. In Section 2.6, we use the heterogeneity function $\psi(j) = j^2/2$ proposed in Darroch *et al.* (1993). This gives the robust design model M_{Dh}^t using the notation of Rivest and Daigle (2004). The next section introduces a parametric bootstrap to calculate the variances of the estimated parameters.

2.5.2 A parametric bootstrap for variance calculations

Rather than simulating the whole data set $\{n_\omega\}$, we only get bootstrap repetitions for the sufficient statistics u_i, v_i, C_i and C_{ψ_i}, $i = 1, \ldots, I$. This is all what is needed to run the estimation algorithm of the previous section. The sufficient statistics are obtained through a sequential procedure in which $\tilde{U}_i$ and $\tilde{M}_i$ $(i = 1, 2, ..., I)$, are simulated as Poisson random variables.

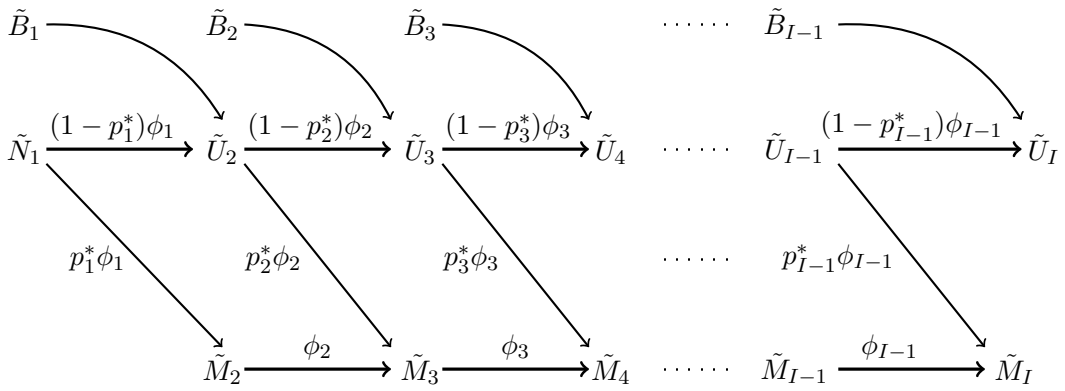

Figure 2.1 – The sampling process describing the sequential procedure in which the Poisson random variables $\tilde{U}_i$ and $\tilde{M}_i$ $(i = 1, 2, ..., I)$ are simulated.

The simulation algorithm is summarized in Figure 2.1. It starts with N_1 and B_1 equal to the estimates obtained when fitting the model. For the 1^{st} SP, the population size $\tilde{N}_1 \sim Poisson(N_1)$. Just before the 2^{nd} SP, the population has two components $(\tilde{M}_2, \tilde{U}_2)$. Given $\tilde{N}_1$, $(\tilde{M}_2, Z_2^{(m)})$ are simulated as a multinomial vector with $\tilde{N}_1$ trials and probabilities $\{p_1^*\phi_1, (1 - p_1^*)\phi_1\}$ and $\tilde{U}_2 = Z_2^{(m)} + \tilde{B}_1$, where $\tilde{B}_1 \sim Poisson(B_1)$. For the third SP, given $\tilde{U}_2$, $(Z_2^{(u)}, Z_2^{(m)})$ is simulated as a multinomial vector with $\tilde{U}_2$ trials and probabilities $\{(1-p_2^*)\phi_2, p_2^*\phi_2\}$. Given $(\tilde{U}_2, \tilde{M}_2)$, one has $\tilde{U}_3 = Z_3^{(u)} + \tilde{B}_2$ and $\tilde{M}_3 = Z_3^{(m)} + W_3$ where W_3 is a binomial random variable involving $\tilde{M}_2$ trials with probability ϕ_2 and where $\tilde{B}_2 \sim Poisson(B_2)$. The procedure for the other SPs is similar.

Once $\tilde{U}_i$ and $\tilde{M}_i$ $(i = 1, 2, ..., I)$ are simulated, we generate u_i, v_i, C_i and C_{ψ_i} as follows:

— Generate u_i from a $Binomial(\tilde{U}_i, p_i^*)$;

— Generate m_i from a $Binomial(\tilde{M}_i, p_i^*)$;

— Generate v_i from a $Binomial(u_i + m_i, \chi_i)$;

— Compute $w_i = \sum_{j=1}^{i-1}(u_i - v_i)$, $w_1 = 0$;

— Generate the vector $\{f_{ij}\}$ using a multinomial distribution with $u_i + m_i$ trials and probabilities p_{ik} given in (2.22);

— Compute $C_i = \sum_{j=1}^{J} j f_{ij}$ and $C_{\psi_i} = \sum_{j=1}^{\ell_i} \psi(j) f_{ij}$.

Now the variances for the estimated parameters are obtained as follows:

— Set the parameters $\{(\beta_i, \tau_i, \phi_i, N_i, B_i) : i = 1, \ldots, I\}$ equal to the maximum likelihood estimates

— Simulate L sets of sufficient statistics and estimate the demographic parameters for each one using the algorithm described in (2.5.1);

— Calculate the variances using a parametric bootstrap. For population sizes, considering (2.13), the variance estimate is the bootstrap variance minus $\hat{N}_i$. The coefficient of variation for $\hat{N}_i$ ($\hat{B}_i$) is calculated as the bootstrap variance of $\log(N_i)$ ($\log(B_i)$). For $\hat{\phi}_i$ ($\hat{p}_i^*$) the coefficient of variation is the bootstrap standard deviation of $\hat{\phi}_i$ ($\hat{p}_i^*$) over the corresponding bootstrap mean.

In each bootstrap sample, the population size for the first and the last SP is estimated using a bias corrected estimate, available in the function `closedp.bc` of `Rcapture`. The above algorithm applies to any log-linear model for the within SP captures.

2.5.3 Simulation study

This section reports the results of a simulation study that investigated the validity of the bootstrap inference procedure for the population sizes $\hat{N}_i$. The data were generated using a stationary robust design model M_{Dh}^t with $I = 9$ SPs (weeks) lasting $\ell = 7$ days. Two capture probabilities $p^* = 0.3,\ 0.5$ and two survival probabilities $\phi = 0.6,\ 0.8$ were used. The weekly arrival was set to $B = 210$ and the number of weeks that a unit stayed in the population was simulated using a geometric distribution with parameter $1 - \phi$. Three scenarios were used for the units' arrival. Under the first one (RD) , all of them arrive on the first day of every week, with the second one (ED1) one half arrives on day 1 and one half on day 2 while under the third scenario (ED2), the arrivals are constant for the seven days of a SP. Scenario ED1 represents a model where clients arrive at traditional shopping days, e.g. Friday and Saturday. ED1 and ED2 are violations to the robust design assumption that births occur prior to SPs. The goal of the simulation study is to estimate the stationary population size for the $5th$ SP, $N_5 = 210/(1 - \phi)$. Additional details and the results are provided in the Supplementary Material.

Overall the inference for $\hat{N}_5$ is reliable under the first two scenarios, RD and ED1. The relative bias of $\hat{N}_5$ is less than 5% and the coverage of the bootstrap confidence interval is within 2% of the nominal value of 95%. Under scenario ED2, with the daily arrivals of 30 persons, N_5 is underestimated by up to 20% and the coverage of the 95% confidence interval is well below the 95% level. If the process deviates too much from a standard robust design model, population estimates are biased. Can relative changes in population levels, the main goal of such analysis laid out in Section 4.1, be estimated well? This was investigated in a second simulation study under scenario ED2, where the number of daily births increased by $inc = 20\%$, 50% and 80% at week 10. The relative population increase between weeks 5 and 35, $(N_{35} - N_5)/N_5$ should then be equal to inc. Monte Carlo estimates of the bias and of the root mean square error of $\hat{inc} = (\hat{N}_{35} - \hat{N}_5)/\hat{N}_5$ are presented in the Supplementary Material. In most cases the bias of $\hat{inc}$ is less than 1% except for inc =80%, where positive biases of up to 8% have been found. The coverage for the 95% bootstrap confidence interval is within 1% of its nominal value. Even when the client arrival process deviates from a robust design, the proposed analysis tracks the relative changes in clientele levels relatively well. It is robust to violations to the robust design assumption.

2.6 Case study: Clientele Estimation at Auto Dealerships

This section analyzes data collected by Ninth Decimal, a California marketing platform, about the daily app activations at 11 auto-dealerships of a major brand in a US metropolitan area. The goals of the analysis was to estimate the change of clientele, in August, September and October, from 2014 to 2015. The parameters of interest are N_i $(i = 1, \ldots, I)$, the expected clientele sizes over the $I = 76$ weeks of the study period which lasted from June 2014 to November 2015. Day 1 is Sunday June 1^{st} 2014. A capture history is a list of days on which a device has been activated. For the 532 days of data collection, a total of 9316 individuals were captured; among those individuals 77% were captured only once. In the next section, a log-linear closed population model for the analysis of the within SP data is selected.

2.6.1 Selection of the Closed Population model in the Primary Sampling Periods

Within each of the 76 SPs, the population is assumed to be closed. As stated in Section 4.1, the decision to activate an app may vary across units leading to heterogeneous capture probabilities modelled by M_h. The fit of M_h was good for most weeks. Binomial probability plots discussed in Baillargeon and Rivest (2007) had a convex shape suggesting a heterogeneity in capture probability, and Chao's heterogeneity model, see Rivest and Baillargeon (2007), had a small deviance. To choose the best log-linear model we compared the estimates of p_i^* under eight models with the ones obtained under a Jolly-Seber model. Boxplots of the 76 differences $(\hat{p}_i^*)^{CP} - (\hat{p}_i^*)^{JS}$ are given in Figure 2.2. This led us to select Darroch model, with

$\psi(k) = k^2/2$ for the heterogeneity. The alternative gamma model, with $\psi(k) = -\log(k+a)$ for positive a values, was not chosen because it sometimes gives wild population estimates. As M_h Darroch overestimates slightly the Jolly-Seber capture probabilities it should give conservative estimates of population sizes.

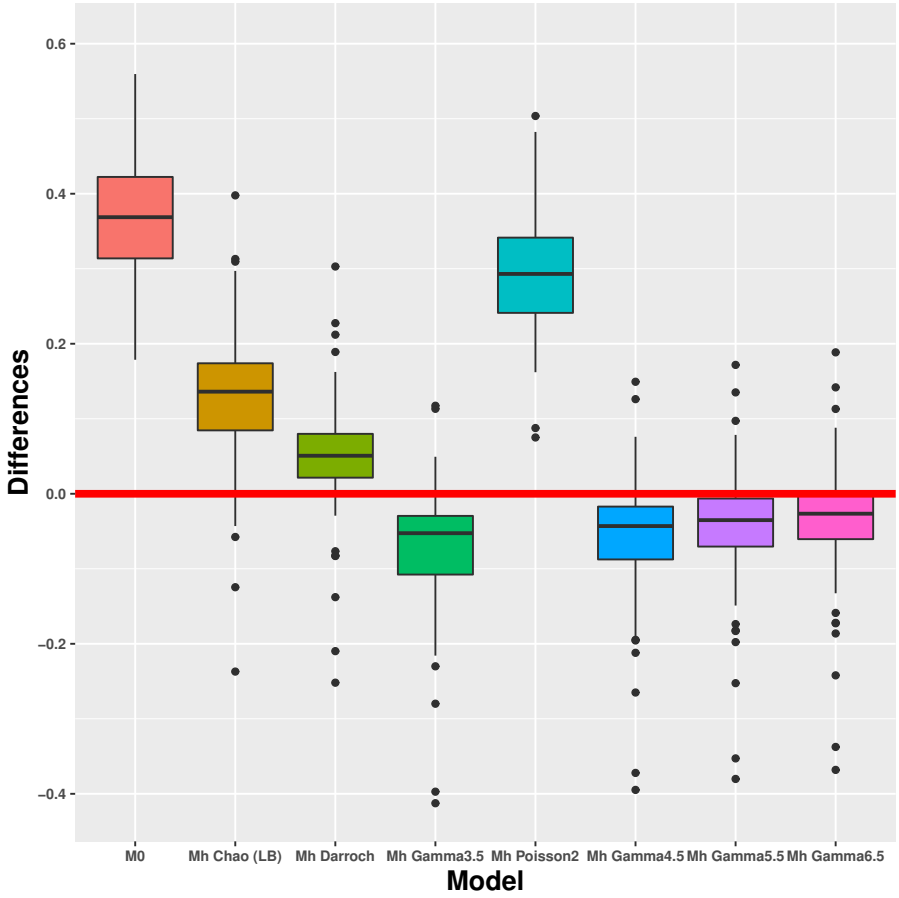

Figure 2.2 – Differences between the estimated capture probabilities under the Jolly-Seber model and the following closed population models: M_0, M_h Chao (LB), M_h Darroch, M_h Gamma for positive values $a = 3.5, 4.5, 5.5, 6.5$, and M_h Poisson for $a = 2$. For each of the closed population models 76 capture probability estimates are obtained and their respective differences with the corresponding Jolly-Seber estimates are plotted.

2.6.2 The Analysis Between Primary Sampling Periods

Robust design model M_{Dh}^t has 152 capture parameters and 151 demographic parameters for a total of 303 parameters making likelihood-based approaches impractical. It was fitted with estimating equations as presented in Section 2.5. The results were compared to estimates obtained with the Jolly-Seber model for the pooled weekly data and to a closed population model, M_h Darroch fitted, independently to the 76 SPs. The R programs implementing this methodology are available in the Supplementary Material.

Figure 2.3 shows the population size evolution over 76 weeks for the three models. The estimates $\hat{p}_i^*$ were larger at the beginning (first 18 weeks) and at the end (last 6 weeks) of the observation period. Thus, the capture-recapture approach appears to correct for a variation in the coverage of the activations by Ninth Decimal over the study period; this coverage may depend on the availability of storage for the data or on contracts with the data providers. The simple analysis that uses the weekly number of users, n_i, to track changes in foot traffic is sensitive to such a variation in coverage. Globally Figure 2.3 shows an increase in foot traffic. If $(N_{i+1}-N_i)/N_i$ denotes the growth rate between weeks i and $i+1$ then the capture-recapture estimate of the average growth rate is 5.5%, as compared with 4% calculated with the n_i's. This difference is meaningful and the robust design estimates are more reliable as they account for a variation in the coverage of the marketing platform. For August, September and October the 2015 weekly clientele is 92% larger (*s.e.* 39%) than in 2014; thus there was a significant increase in foot traffic in 2015.

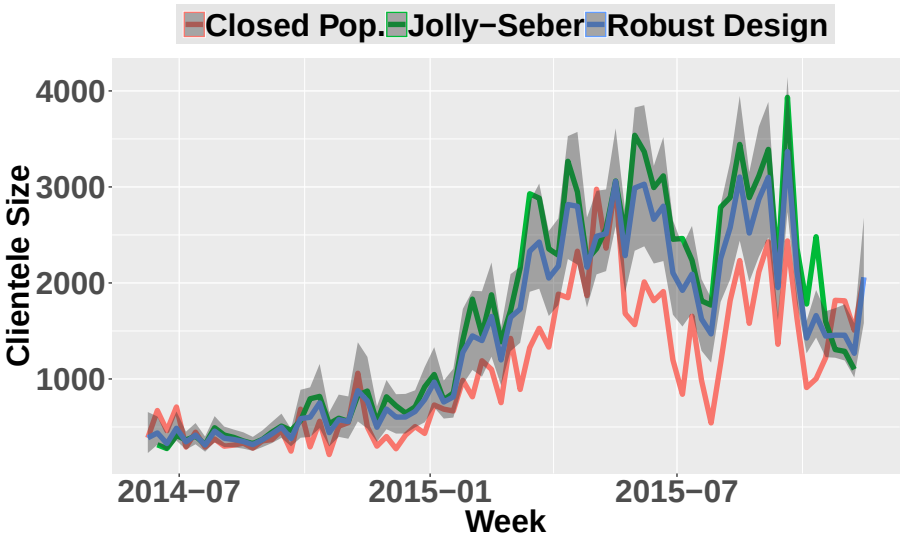

Figure 2.3 – Evolution of the estimated clientele size over the 76 weeks under models M_h for a closed population, Jolly-Seber and M_{Dh}^t, with a 95% confidence band for M_{Dh}^t.

Numerical values for the estimates $\hat{N}_i$ are provided in the Supplementary Material. The corresponding coefficients of variation, calculated with 1000 bootstrap samples, ranged between 6% and 34%, 6% and 28%, and 13% and 52% for respectively the Jolly Seber estimates, the robust design estimates and the closed population estimates. Figure 3, in the Supplementary Material, gives the efficiencies of robust design estimator with respect to the Jolly Seber and of the closed population model estimators. As laid out in Section 2.4.2 the robust design estimates of N_i are more precise than those of the closed population when survival probabilities are high; this is the case for this data set as the estimates of survival probabilities vary between 0.5 and 1. The gains in precision of the robust design estimators over the Jolly-Seber

estimators is less important with such high survival probabilities.

To complete this analysis, we report two investigations of the robustness of the results. First alternative data sets were created by removing the first few days. If days 1 and 2 are dropped then SP 1 comprises days 3 to 9 and so on. The results of these alternative analyses, indexed by a SP inital day (either Sunday, Monday or Tuesday) are provided in the Supplementary Material. The findings do not change with a redefinition of the SPs. We also fitted the M_{Dh}^t robust design to data, for the same auto brand and the same period, in a second metropolitan area. The capture probabilities should not be specific to a metropolitan area as they only depend on the data acquisition process at the marketing platform. Graphs of the two sets of weekly capture probabilities $\{\hat{p}_i^*\}$, provided in the Supplementary Material section, show a close agreement supporting the claim that the capture mechanisms are very similar in the two metropolitan areas.

2.7 Discussion

The analysis of data on the activations of applications on mobile phones raises interesting methodological issues. This paper has argued that capture-recapture techniques can be used in that context and new statistical tools have been presented for this purpose. Clearly the quality of the activation data can be questioned. A validation step is mandatory to establish the reliability of the results produced by any statistical analysis.

This work also contributes to the literature on the robust design by suggesting a new algorithm for estimating the parameters in a data set with an arbitrary large number of capture occasions. It provides, in Proposition 1, closed form expressions of the variances of estimators in a robust design that permit a simple evaluation of the gains in precision associated to the robust design.

Transition

Au chapitre précédent, nous avons abordé le problème de l'estimation des paramètres du design robuste dans le cas d'un nombre suffisamment élevé d'occasions de capture. Le chapitre suivant, quoique indépendant du chapitre précédent, se situe toujours dans le contexte du design robuste. De façon spécifique, on se démarque des approches classiques d'estimation du design robuste en abordant la question de l'estimation de la taille de la population à partir de deux sources d'information: les données à l'intérieur d'une période primaire (ou intra-période) et les données d'une période primaire à une autre (ou inter-période). On démontre que les estimateurs de la taille de la population obtenus avec les informations intra-période et inter-période sont asymptotiquement indépendants pour la classe de modèles de population fermée présentée à la Section 1.1.5. Se servant de ce résultat, on montre ensuite que l'estimateur du maximum de vraisemblance pour la taille de la population dans le cas du design robuste est asymptotiquement équivalent à un estimateur pondéré pour le modèle de population ouverte de Jolly-Seber présenté à la Section 1.2.2 et le modèle de population fermée. Ces résultats ont trouvé application dans un exemple tiré de Santostasi *et al.* (2016) et qui traite de l'estimation de la taille d'une population de dauphins vivant dans le Golfe de Corinthe (Grèce).

Chapter 3

On the Estimation of Population Sizes in Capture-recapture Experiments

Résumé

Ce travail considère une expérience de capture-recapture imbriquée avec deux niveaux d'échantillonnage : à l'intérieur de chaque période primaire d'un modèle de population ouverte, il y'a des occasions de capture secondaires pour estimer la taille de la population à la période primaire correspondante. Cette expérience est connue sous le nom de design robuste. Deux sources d'informations sont disponibles pour estimer la taille de la population à une période primaire : les données inter et intra primaire. Ce travail démontre que les estimateurs de la taille de la population obtenus à partir de ces deux sources d'informations sont asymptotiquement indépendantes pour une large classe de modèles de population fermée. Dans ce contexte, il est démontré que l'estimateur du maximum de vraisemblance pour la taille de la population obtenu sous le design robuste est asymptotiquement équivalent à une somme pondérée des estimateurs pour le modèle de population ouverte de Jolly-Seber (Jolly 1965; Seber 1965) et pour le modèle de population fermée. Cet article montre que cet estimateur pondéré est plus efficace que l'estimateur des moments de Kendall *et al.* (1995). Une formule explicite de l'efficacité associée à l'estimateur de Kendall est donnée et la perte de précision évaluée via une étude de simulation et à travers un exemple sur l'estimation de la taille d'une population de dauphins présenté dans Santostasi *et al.* (2016).

*M*ots-clé : Distributions asymptotiques, Hétérogenéité, Modèle de Jolly-Seber, Capture-recapture, Distribution multinomiale, Régression de Poisson, Design robuste.

Abstract

This work considers a nested mark-recapture experiment with two levels of sampling: within each primary sampling period of an open population model, there are secondary capture occasions to estimate the size of the population at that primary period. This scheme is known as Pollock's robust design. Two sources of information are then available to estimate the population size for a primary period: the within and the between primary period data. This work proves that the population size estimators derived from these two sources are asymptotically independent for a large class of closed population models. In this context, the robust design maximum likelihood estimator of population size is shown to be asymptotically equivalent to a weighted sum of the estimators for the open population Jolly-Seber model (Jolly 1965; Seber 1965) and for the closed population model. This article shows that the weighted estimator is more efficient than the moment estimator of Kendall *et al.* (1995). A closed form expression for the efficiency associated with this estimator is given and the loss of precision is evaluated in a Monte Carlo study and in a numerical example about the estimation of the size of dolphin populations discussed by Santostasi *et al.* (2016).

*K*eywords: Asymptotics, Heterogeneity, Jolly-Seber model, Mark-recapture study, Multinomial distribution, Poisson regression, Robust design.

3.1 Introduction

The methodology for the estimation of population sizes in capture-recapture studies is well understood for closed and open population models. For closed population models, one distinguishes the multinomial and the conditional estimator, see Fewster and Jupp (2009). Inference on population sizes can be carried out using a Poisson likelihood as pointed out in Sandland and Cormack (1984) and Cormack (1992). For open population models, Jolly (1965) and Seber (1965) built a likelihood by assuming that the number of unmarked animals prior to each capture occasion are fixed parameters; they derived maximum likelihood estimators for the so-called Jolly-Seber model. Poisson (Cormack, 1989) and multinomial (Schwarz and Arnason, 1996) likelihoods have also been considered for this problem. This work is concerned with a hierarchical study design where a capture-recapture experiment, involving *secondary capture occasions*, is carried out within each *sampling period* (SP) of an open population model. This is the robust design introduced by Pollock (1982). Rivest and Daigle (2004) obtained the maximum likelihood estimators of the population sizes through a Poisson regression. This involves a dependent vector of size $2^L - 1$, where L is the total number number of capture occasions. When $L > 20$ the dependent vector and the associated design matrix exceed the storage capacity of standard routines for generalized linear models and their estimators cannot be implemented. Kendall *et al.* (1995) proposed a moment-type estimator for abundance and, in a recent paper, Yauck *et al.* (2018) proposed a sequential estimation procedure for the parameters of the robust design when the closed population models within each primary period belong to a family considered in Rivest and Lévesque (2001).

A robust design involves information within and between the primary periods of the underlying open population experiment. These two sources of information are combined when estimating population sizes. This paper shows that for a large class of such population models, the between and the within period estimators of population sizes are asymptotically independent. In this case the robust design maximum likelihood estimator of population size is asymptotically equivalent to a weighted sum involving the estimators for the Jolly-Seber model and for the closed population model. This estimator differs from the moment estimator of Kendall *et al.* (1995). We show that the moment estimator is not efficient as it is not a maximum likelihood estimator; a formula for its efficiency is provided. The loss of precision associated with this estimator is evaluated in a Monte Carlo study and in a numerical example about the estimation of the size of dolphin populations discussed by Santostasi *et al.* (2016).

To apply the methodology proposed in this work to a robust design data set, one needs to select closed population models for the data on the secondary capture occasions within each primary period. Methods to do so are discussed in Rivest and Baillargeon (2007) and Yauck *et al.* (2018). The class of models for the within primary period data considered in this work belongs to the family of closed population models presented in Rivest and Lévesque (2001).

3.2 Model building

This section briefly presents a stochastic model for the animal population under study and the sampling process which is used to gather information on the population. It also introduces the asymptotic framework in wich estimators of population sizes will be investigated

3.2.1 A State Process for the Population

We are interested in an open animal population evolving in time, over I sampling periods (SP). These SPs could for instance be successive years. Its behavior is described by a non homogeneous stochastic process whose distribution depends on $2I-1$ parameters:

— $B_i,\ i = 0, \ldots, I-1$ where B_{i-1} is the expected number of birth in the population occurring before the ith SP and B_0 is the expected population size at the first SP;

— $\phi_i,\ i = 1, \ldots, I-1$ where ϕ_i is the probability for an animal in the population at SP i to survive to SP $i+1$.

The expected population sizes are defined, for $i = 1, \ldots, I$, by the following recurrence relationship, $N_i = N_{i-1}\phi_{i-1} + B_{i-1}$ with $N_1 = B_0$.

The stochastic process $\{\tilde{N}_i, i = 1, 2, ..., I\}$ describing the evolution of a single population over time is defined by

$$\tilde{N}_i = \sum_{j=1}^{\tilde{N}_{i-1}} \epsilon_j^{(i)} + \tilde{B}_{i-1}, \quad i = 1, \ldots, I, \tag{3.1}$$

where $\epsilon_j^{(i)}$ is a sequence of i.i.d Bernouilli random variables with probability ϕ_{i-1} and independent of $\tilde{N}_{i-1}$, such that $\epsilon_j^{(i)} = 1$ if unit j survives between $i-1$ and i and 0 otherwise and the births $\{\tilde{B}_{i-1}\}$ just before SP i are assumed to have a Poisson distribution with parameter B_{i-1}.

The stochastic process of the observed population sizes $\{\tilde{N}_i, i = 1, 2, ...\}$ is a non homogeneous autoregressive Poisson process of order 1 (AR(1)) (McKenzie, 1985). Under this model, $\{\tilde{N}_i : i = 1, \ldots, I\}$ have dependent Poisson distributions. In the capture-recapture literature, this state model is implicitly assumed in the articles of Cormack (1989) and Rivest and Daigle (2004) that use log-linear models and Poisson likelihoods to estimate the parameters of this process using capture-recapture data.

3.2.2 Capture-recapture sampling of the population

The capture-recapture experiment used to obtain information about the population is the robust design of Pollock (1982). It is illustrated in Figure 3.1. Within the ith SP a capture-recapture experiment with ℓ_i capture occasions is set up. It takes place over a short period of time so that the population is assumed to be closed, that is it experiences neither death nor birth, during the capture-recapture experiment for the ith SP.

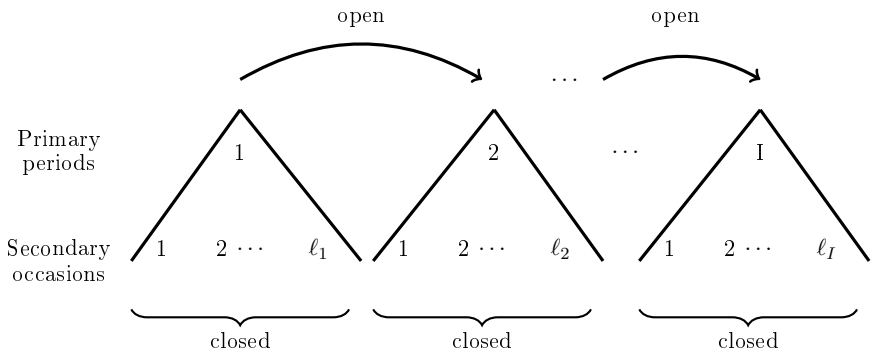

Figure 3.1 – The sampling scheme for the robust design with I sampling periods (SP) and ℓ_i $(i = 1, 2, \ldots, I)$ secondary occasions within each SP.

A population unit is given a unique tag on its first capture and it can be re-identified when it is recaptured. The data entry for this unit is a capture history, a $(\sum \ell_i) \times 1$ vector $\boldsymbol{\omega}$ with entry 1 $(\omega_{ij} = 1)$ if the unit is caught at the j^{th} $(j = 1, 2, \ldots, \ell_i)$ capture occasion of the i^{th} SP $(i = 1, 2, \ldots, I)$. The data set are the frequencies $\{n_\omega : \boldsymbol{\omega} \in \Omega\}$ where Ω is the set of the $s = 2^{\sum \ell_i} - 1$ observable capture histories. Under the Poisson process of Section 3.1, the frequencies $\{n_\omega : \boldsymbol{\omega} \in \Omega\}$ have independent Poisson distributions with parameter μ_ω that depends on the model for the capture of the units. Some of these models are now reviewed.

Suppose that $I = 1$, then dropping subscript i, the goal of the analysis is to estimate the

population size $\tilde{N}$ using the observed frequencies $\{n_\omega\}$ for the $s = 2^\ell - 1$ observable capture histories. The simplest model, called M_0, assumes that the number of captures for a unit follows a binomial distribution with parameters ℓ and p where p, the capture probability, is the same for all occasions and all units in the population. The predicted frequency for this model has the following log-linear expression

$$\mu_\omega = \exp\{\gamma + \sum_i \omega_i \beta\}, \boldsymbol{\omega} \in \Omega,$$

where $\gamma = \log\{N(1-p)^\ell\}$ is the log-predicted frequency for the number of units that are not captured and $\beta = \log\{p/(1-p)\}$ is the logit of the capture probability at a single occasion. This is easily generalized to model M_t that assumes that the capture probability varies between capture occasions. The hypothesis of independence between capture occasions can be relaxed by adding interaction terms. When there is variation in capture probabilities across the $\tilde{N}$ units, model M_h handles heterogeneity in catchability by assuming that each unit has its own capture probability p_i $(i = 1, \ldots, \tilde{N})$ sampled from a probability distribution. A behavioral change (or trap response) after the first capture leads to model M_b, a generalization of M_0 that assumes different capture probabilities, p and c, for units that are respectively marked and unmarked. Generalizations of these basic models are presented in Farcomeni (2016). Procedures for obtaining population sizes estimators in closed population experiments are presented in Section 3.3.1.

When there is a single capture occasion within each sampling period, $\ell_i = 1(\, i = 1, \ldots, I)$, then one– gets the Jolly-Seber experiment, considered by Jolly (1965) and Seber (1965). Besides the parameters $\{B_i\}$ and $\{\phi_i\}$, the model for the data involves capture probabilities $\{p_i^* : i = 1, \ldots, I\}$. The sufficient statistics for this model are

— u_i, the number of unmarked units captured during SP i;

— m_i, the number of marked units captured during SP i and $n_i = u_i + m_i$ the number of units captured during SP i;

— v_i the number of units captured for the last time at SP i;

— $w_i = \sum_{s=1}^{i-1}(u_s - v_s)$ the number of units captured before SP i that will are recaptured, either at SP i or later.

Section 3.3.2 discusses the estimation of population sizes for a Jolly-Seber open population experiment while Section 3.4 deals with general robust design experiments.

3.2.3 Asymptotic distributions of population size estimators

The next section investigates the asymptotic distribution of estimators $\hat{N}_i$ when the set of capture histories Ω is fixed and when the populations sizes N_i tend to infinity. If ν is used to index the increasing population sizes, we assume that N_i^ν / N_j^ν converges to some fixed constant bounded away from 0 and infinity for any pair (i, j), that is N_i^ν and N_j^ν are of the same order

for any pair (i, j). Under this assumption, the frequencies $\{n_\omega\}$ converge in distribution to independent normally distributed random variables:

$$\frac{n_\omega - \mu_\omega}{\sqrt{N_1}} \xrightarrow{\mathcal{L}} \mathcal{N}\left(0, \lim_{N_1 \to \infty} \frac{\mu_\omega}{N_1}\right). \tag{3.2}$$

The population estimators are expressed as $\hat{N}_i = g(n_\omega; \boldsymbol{\omega} \in \Omega)$, where the function g is assumed to satisfy

$$cN_i = g(c\mu_\omega; \boldsymbol{\omega} \in \Omega), \quad c > 0. \tag{3.3}$$

If $\hat{N}_i$ is a consistent estimator of N_i, meaning that $\hat{N}_i^\nu / N_i^\nu$ converges to 1 in probability as ν goes to infinity, then the asymptotic distribution of $\hat{N}_i - N_i$ can be derived using a standard linearization argument:

$$\sqrt{N_i}\left\{\frac{g(n_\omega; \boldsymbol{\omega} \in \Omega) - N_i}{N_i}\right\} = \sqrt{N_i}\left\{g\left(\frac{n_\omega}{N_i}; \boldsymbol{\omega} \in \Omega\right) - 1\right\}.$$

As N_i goes to infinity and using (3.2), it follows that

$$\frac{\hat{N}_i - N_i}{\sqrt{N_i}} \xrightarrow{\mathcal{L}} \mathcal{N}\left(0, \lim_{N_i \to \infty} N_i \sum_{\boldsymbol{\omega} \in \Omega} \left\{\frac{\partial}{\partial \mu_\omega} g\left(\frac{\mu_\omega}{N_i}; \boldsymbol{\omega} \in \Omega\right)\right\}^2 \mu_\omega\right). \tag{3.4}$$

Using a slight abuse of notation, we reformulate this result as meaning that the asymptotic distribution of $\hat{N}_i - N_i$ is $N\left(0, \mathrm{var}_P(\hat{N}_i)\right)$ where

$$\mathrm{var}_P(\hat{N}_i) = \sum_{\boldsymbol{\omega} \in \Omega} \left\{\frac{\partial}{\partial \mu_\omega} g\left(\mu_\omega; \boldsymbol{\omega} \in \Omega\right)\right\}^2 \mu_\omega.$$

Here $\mathrm{var}_P(\hat{N}_i)$ is an $O(N_i)$ expression such that $\mathrm{var}_P(\hat{N}_i)/N_i$ gives the asymptotic variance of $\sqrt{N_i}(\hat{N}_i/N_i - 1)$. In applications, we are interested in the properties of $\hat{N}_i$ as a predictor for the current population size $\tilde{N}_i$. The next proposition gives an expression for the asymptotic variance of the prediction error $\hat{N}_i - \tilde{N}_i$.

Proposition 3 *Let $\hat{N}_i = g(n_\omega; \boldsymbol{\omega} \in \Omega_i^*)$, where Ω_i^* be the set of capture histories with either a capture at SP i or one capture before SP i and one after, be a population size estimator satisfying (3.3). As N_i goes to ∞,*

— *$\hat{N}_i - \tilde{N}_i$ is asymptotically normal with mean 0 and asymptotic variance*

$$var(\hat{N}_i - \tilde{N}_i) = var_M(\hat{N}_i) \approx var_P(\hat{N}_i) - N_i, \tag{3.5}$$

— *If two estimators $\hat{N}_i^a$ and $\hat{N}_i^b$ only involve capture histories in Ω_i^* and satisfy (3.3), then $\left(\hat{N}_i^a - \tilde{N}_i, \hat{N}_i^b - \tilde{N}_i\right)^\top$ is asymptotically normal with mean $(0,0)^\top$ and asymptotic variance-covariance matrix*

$$\boldsymbol{\Sigma_i} = \begin{pmatrix} var_M(\hat{N}_i^a) & cov_M(\hat{N}_i^a, \hat{N}_i^b) \\ & var_M(\hat{N}_i^b) \end{pmatrix}, \tag{3.6}$$

where $cov_M(\hat{N}_i^a, \hat{N}_i^b) = cov_P(\hat{N}_i^a, \hat{N}_i^b) - N_i$ *and*

$$cov_P(\hat{N}_i^a, \hat{N}_i^b) = \sum_{\boldsymbol{\omega} \in \Omega} \left\{ \frac{\partial}{\partial \mu_\omega} g_a\left(\mu_\omega; \boldsymbol{\omega} \in \Omega_i\right) \right\} \left\{ \frac{\partial}{\partial \mu_\omega} g_b\left(\mu_\omega; \boldsymbol{\omega} \in \Omega_i\right) \right\} \mu_\omega,$$

is the asymptotic Poisson covariance between the two estimators.

Proposition 1 is proved in the Appendix. Condition (3.3) is satisfied for all the estimators considered in this work. For closed population models, one could condition on $\tilde{N}_i$ and use the multinomial distribution to investigate the asymptotic distribution of $\hat{N}_i$. Sandland and Cormack (1984) show that the prediction variance $\text{var}(\hat{N}_i - \tilde{N}_i)$ is equal to the variance under a multinomial sampling, that is why we used subscript M to represent the prediction variance in (3.5). In a similar way, Jolly (1965) suggest to use a "conditional" variance to evaluate the precision of $\hat{N}_i$.

3.3 Estimating population sizes in capture-recapture experiments: A review

Two strategies can be envisaged to estimate the size of the population at time i. One can fit a closed population model, using only the data collected at SP i, to obtain an estimator $\hat{N}_i^{CP}$. It is also possible to pool the data for all capture occasions within SP i and to use a Jolly-Seber estimator $\hat{N}_i^{JS}$. One goal of this work is to prove that the two prediction errors, $\hat{N}_i^{CP} - \tilde{N}_i$ and $\hat{N}_i^{JS} - \tilde{N}_i$, are asymptotically independent. This section derives the linearization approximation to these two estimators needed to prove the asymptotic independence using Proposition 1.

3.3.1 Closed population models

As this section considers only the capture histories within a given SP, it is convenient to drop subscript i. We investigate log-linear models that expresses μ_ω as

$$\log \mu_\omega = \gamma + {X_\omega}^\top \beta, \tag{3.7}$$

where γ is the logarithm of the predicted frequency for the number of missed units, X_ω is the $d \times 1$ vector of explanatory variables that depends on the underlying model and $\boldsymbol{\beta}$ is a $d \times 1$ vector of unknown parameters. For (3.7), the vector of unknown parameters is $\boldsymbol{\theta} = (\gamma, \beta^\top)^\top$. Note that X_0, the value of X for the null capture history $(0, 0, \ldots, 0)$ is assumed to be a vector of 0's. The matrix of explanatory variables $\boldsymbol{X}$ for this model has dimensions $s \times d$, its entries for row $\boldsymbol{\omega}$ is the vector X_ω. We define $\boldsymbol{y}$ as the vector containing the frequencies for the $s = 2^\ell - 1$ observed capture histories and $\boldsymbol{\mu}$ its corresponding vector of expectations. The sufficient statistics for (3.7) is the vector $(n, \boldsymbol{y}^\top \boldsymbol{X})^\top$ where n is the total number of units captured at least once. The probability of being captured at least once is $p^* = 1 - e^\gamma / N$. To

linearize $\hat{N}^{CP}$, it is convenient to define a probability distribution over the 2^ℓ possible capture histories by letting μ_ω/N be the probability of $\boldsymbol{\omega}$. This associates to the matrix $\boldsymbol{X}$ a d-variate random vector; let μ_X and Σ be respectively the $d \times 1$ vector of expectations and the $d \times d$ variance-covariance matrix for this random vector. For model M_0, $d = 1$, $X_\omega = \sum \omega_i$ is the total number of captures and has a binomial distribution with parameters ℓ and p; in this case $\mu_X = \ell p$ and $\Sigma = \ell p(1-p)$. A proposal for M_h has $d = 2$ and $X_\omega = (\sum \omega_i, \psi(\sum \omega_i))^\top$ where $\psi(\cdot)$ is a convex function; $\psi(t) = t^2/2$ gives a model proposed in Darroch *et al.* (1993). Other capture-recapture models, such as the proposal of Yang and Chao (2005), have the form (3.7).

For the log-linear models presented in Equation (3.7), the population size estimator for N is

$$\hat{N}^{CP} = n + e^{\hat{\gamma}}, \tag{3.8}$$

where $\hat{\gamma}$ is the estimator of the intercept obtained when fitting a Poisson regression to the capture-recapture data vector $\boldsymbol{y}$. As shown in the Appendix, an asymptotic expansion for $\hat{N}^{CP}$ is

$$\hat{N}^{CP} - N \approx \nabla \hat{N}^{CP} \begin{pmatrix} n - Np^* \\ \boldsymbol{X}^\top \boldsymbol{y} - \boldsymbol{X}^\top \boldsymbol{\mu} \end{pmatrix}, \tag{3.9}$$

where

$$\nabla \hat{N}^{CP} = \frac{1}{(1-p^*)^{-1} - 1 - \mu_X^\top \Sigma^{-1} \mu_X} \left((1-p^*)^{-1} \quad , \quad -\mu_X^\top \Sigma^{-1} \right)^\top$$

is the limit of the vector of partial derivatives of $\hat{N}^{CP}$ with respect to the sufficient statistics. The asymptotic multinomial variance of $\hat{N}^{CP}$, as given in Equation (6) of Rivest and Lévesque (2001), is

$$\begin{aligned} \mathrm{var}_M \left(\hat{N}^{CP} \right) &= (\nabla \hat{N}^{CP})^\top \mathrm{cov} \begin{pmatrix} n \\ \boldsymbol{X}^\top \boldsymbol{y} \end{pmatrix} \nabla \hat{N}^{CP} - N \\ &= \frac{N}{(1-p^*)^{-1} - 1 - \mu_X^\top \Sigma^{-1} \mu_X}, \end{aligned} \tag{3.10}$$

where cov stands for the Poisson covariance matrix. For M_0, $\mu_X = \ell p$ and $\Sigma = \ell p(1-p)$ and Equation (3.10) reduces to $N(1-p^*)/P_2$, the multinomial variance of $\hat{N}$ for closed population model M_0 presented in Rivest and Lévesque (2001), where $P_2 = 1 - (1-p)^\ell - \ell p (1-p)^{\ell-1}$ is the probability of being captured at least twice.

3.3.2 The Jolly-Seber open population model

In a Jolly-Seber model, the estimator for N_i is given in Equation (23) of Jolly (1965). It can be expressed in terms of the sufficient statistics defined in Section 2.2 as

$$\hat{N}_i^{JS} = \frac{n_i \left\{ (n_i - u_i)(n_i - v_i) + n_i(w_i - n_i + u_i) \right\}}{(n_i - u_i)(n_i - v_i)}, \; i = 2, \ldots, I-1. \tag{3.11}$$

Note that the population sizes for the first and the last sampling periods, N_1 and N_I respectively, are not estimable. An asymptotic expansion for $\hat{N}_i^{JS}$ is

$$\hat{N}_i^{JS} - N_i \approx \nabla \hat{N}_i^{JS} \begin{pmatrix} n_i - N_i p_i^* \\ u_i - N_i \eta_i p_i^* \\ v_i - N_i p_i^* \chi_i \\ w_i - N_i \bar{\eta}_i (\bar{\chi}_i + p_i^* \chi_i) \end{pmatrix}, \tag{3.12}$$

where $\nabla \hat{N}_i^{JS}$ is the limit of the vector of partial derivatives of $\hat{N}_i^{JS}$ with respect to n_i, u_i, v_i, w_i,

$$\nabla \hat{N}_i^{JS} = \{1/(p_i^* \bar{\chi}_i \bar{\eta}_i)\} \times (-D_i, \bar{\chi}_i + p_i^* \chi_i, \bar{\eta}_i (1 - p_i^*), p_i^*)^\top,$$

χ_i is the probability of not being captured after SP i (it satisfies $\chi_i = (1-\phi_i)+\phi_i(1-p_{i+1}^*)\chi_{i+1}$; $\bar{\chi}_i = 1 - \chi_i$ is the probability of being captured after SP i), η_i is the proportion of unmarked units prior to SP i ; $\bar{\eta}_i = 1 - \eta_i$ is the proportion of marked units just before SP i ($\eta_i = 1 - M_i/N_i$, where M_i is the expected number of marked units in the population prior to SP i), $D_i = 1 - (1 - p_i^*)\chi_i \eta_i - \bar{\eta}_i \bar{\chi}_i$ is the probability, for an unmarked unit, to be captured at SP i or later and, for a marked unit, to be captured for the last time at SP i. Using Proposition 1, the asymptotic multinomial variance of $\hat{N}_i^{JS}$ is

$$\begin{aligned} \mathrm{var}_M \left(\hat{N}_i^{JS} \right) &= (\nabla \hat{N}_i^{JS})^\top \mathrm{cov}(n_i, u_i, v_i, w_i)^\top \nabla \hat{N}_i^{JS} - N_i \\ &= \frac{N_i D_i (1 - p_i^*)}{p_i^* \bar{\chi}_i \bar{\eta}_i}, \end{aligned} \tag{3.13}$$

where

$$\mathrm{cov}(n_i, u_i, v_i, w_i)^\top = N_i \begin{pmatrix} p_i^* & \bar{\eta}_i p_i^* & p_i^* \chi_i & \eta_i p_i^* \\ & \eta_i p_i^* & \eta_i p_i^* \chi_i & 0 \\ & & p_i^* \chi_i & \bar{\eta}_i p_i^* \chi_i \\ & & & \bar{\eta}_i (\bar{\chi}_i + p_i^* \chi_i) \end{pmatrix}.$$

Variance formula (3.13) is presented in Equation (27) of Jolly (1965) and discussed in a more general framework in Yauck *et al.* (2018).

3.4 The estimation of population sizes in a robust design

As was mentioned in the introduction, estimating population sizes in a robust design is a complicated problem. Kendall *et al.* (1995) suggest using a moment estimator that, as will be shown in the next sections, can be inefficient while the maximum likelihood approach of Rivest and Daigle (2004) works only if the total number of capture occasions, $\sum \ell_i$, is small. This section suggests a simple estimator obtained by combining the estimators $\hat{N}_i^{CP}$ and $\hat{N}_i^{CP}$ introduced in the last section. The construction of the combined estimator relies on the following Theorem.

Theorem 1 *If a model satisfying (3.7) is used for the within SP data, then, as N_i goes to ∞,*

$$cov_M(\hat{N}_i^{CP}, \hat{N}_i^{JS}) = 0, \quad and \quad cov(\hat{p}_i^{*,CP}, \hat{p}_i^{*,JS}) = 0, \; i = 2, \ldots, I-1;$$

in other words the within and the between SP estimators of both N_i and p_i^* are asymptotically independent.

Proof 1 *We first consider the two estimators for N_i and evaluate their asymptotic Poisson covariance. Using (3.10) and (3.13), this is given by*

$$(\nabla \hat{N}_i^{JS})^\top cov\{(n_i, u_i, v_i, w_i)^\top, (n_i, \boldsymbol{y}_i^\top \boldsymbol{X}_i)\} \nabla \hat{N}_i^{CP},$$

where $\boldsymbol{X}_i$ is the matrix of explanatory variables for the closed population model of the ith SP. The covariances involve Poisson random variables; a covariance is evaluated as the expected number of units shared by two random variables. This leads to

$$cov\{(n_i, u_i, v_i, w_i)^\top, (n_i, \boldsymbol{y}_i^\top \boldsymbol{X}_i)\} = N_i p_i^* (1, \eta_i, \chi_i, \bar{\eta}_i)^\top (1, \mu_{X_i}^\top / p_i^*).$$

Since

$$(\nabla \hat{N}_i^{JS})^\top (1, \eta_i, \chi_i, \bar{\eta}_i)^\top = 1 \quad and \quad (1, \mu_{X_i}^\top / p_i^*) \nabla \hat{N}_i^{CP} = 1/p_i^* \tag{3.14}$$

the Poisson covariance is equal to N_i and, considering Proposition 1, the multinomial asymptotic covariance is equal to 0.

To calculate the covariance between $\hat{p}_i^{*,CP}$ and $\hat{p}_i^{*,JS}$, we first obtain asymptotic expansions for these two statistics. As $\hat{p}_i^* = n_i/\hat{N}_i$ for both closed and open populations, it is a function of the same sufficient statistics as $\hat{N}_i$. Its asymptotic expansion involves the limit of a vector of partial derivatives that is given by

$$\nabla \hat{p}_i = \frac{\boldsymbol{e}_1}{N_i} - \frac{p_i^*}{N_i} \nabla \hat{N}_i, \tag{3.15}$$

where $\boldsymbol{e}_1 = (1, 0, \ldots, 0)^\top$ and $\nabla \hat{N}_i$ is defined in either (3.9) or (3.12) respectively. Using (3.14), the asymptotic covariance between $\hat{p}_i^{*,CP}$ and $\hat{p}_i^{*,JS}$ is easily evaluated

$$\frac{1}{N_i^2} (\boldsymbol{e}_1 - p_i^* \nabla \hat{N}_i^{JS})^\top cov\{(n_i, u_i, v_i, w_i)^\top, (n_i, \boldsymbol{y}_i^\top \boldsymbol{X}_i)\} (\boldsymbol{e}_1 - p_i^* \nabla \hat{N}_i^{CP}) = 0,$$

as the first column and the first row of $cov\{(n_i, u_i, v_i, w_i)^\top, (n_i, \boldsymbol{y}_i^\top \boldsymbol{X}_i)\}$ are respectively given by $N_i p_i^* (1, \eta_i, \chi_i, \bar{\eta}_i)^\top$ and $N_i (p_i^*, \mu_{X_i})$.

Considering Theorem 1, the optimal estimator of N_i, combining information from two sources, is

$$\hat{N}_i = \frac{\hat{N}_i^{CP}/\mathrm{var}_M(\hat{N}_i^{CP}) + \hat{N}_i^{JS}/\mathrm{var}_M(\hat{N}_i^{JS})}{1/\mathrm{var}_M(\hat{N}_i^{CP}) + 1/\mathrm{var}_M(\hat{N}_i^{JS})}. \tag{3.16}$$

Equation (8) of Rivest and Daigle (2004) gives a log-linear model for a robust design with a within SP closed population model satisfying (3.7). The sufficient statistics for this nested model are given by $\{u_i, v_i, w_i, \boldsymbol{X}_i^\top \boldsymbol{y}_i : \ i = 1, \ldots, I\}$; note that n_i is not among the sufficient statistics. Now using (3.12) and (3.9) one obtains an asymptotic expansion for $\hat{N}_i$ in terms of the robust design sufficient statistics and n_i. In this expansion the coefficient of n_i is equal to 0. Thus the combined estimator $\hat{N}_i$ is asymptotically equal to a function of the sufficient statistics for the robust design model. It is therefore asymptotically equivalent to the maximum likelihood estimator of N_i for the robust design model of Rivest and Daigle (2004).

A similar derivation leads to a combined estimator for p_i^*. First, linearization variances are evaluated using (3.14) and (3.15); this gives

$$\mathrm{var}(\hat{p}_i^{*,CP}) = \frac{-N_i p_i^*(1-p_i^*) + (p_i^*)^2 \mathrm{var}_M(\hat{N}_i^{CP})}{N_i^2} = \frac{p_i^*(1-p_i^*)\mu_{X_i}^\top \Sigma_i^{-1} \mu_{X_i}}{N_i A_i^*},$$

where $A_i^* = (1-p_i^*)^{-1} - 1 - \mu_{X_i}^\top \Sigma_i^{-1} \mu_{X_i}$ depends on the closed population model within SPs, and

$$\mathrm{var}(\hat{p}_i^{*,JS}) = \frac{N_i p_i^*(1-p_i^*) + (p_i^*)^2 \mathrm{var}_M(\hat{N}_i^{JS})}{N_i^2} = \frac{p_i^*(1-p_i^*)\left(D_i + \bar{\chi}_i \bar{\eta}_i\right)}{N_i \bar{\chi}_i \bar{\eta}_i}.$$

As in (3.16), a combined estimator $\hat{p}_i^*$ is easily constructed. In the asymptotic expansion for $\hat{p}_i^*$, n_i gets a coefficient equal to 0. Thus $\hat{p}_i^*$ is asymptotically equal to a linear function of the sufficient statistics for a robust design model; it is therefore asymptotically equivalent to the maximum likelihood estimators for the robust design model. To summarize the findings of this section, closed form formulae for the asymptotic variances of the two combined estimators derived in this section are provided:

$$\mathrm{var}_M(\hat{N}_i) \ = \ \frac{N_i D_i \left(1-p_i^*\right)}{p_i^* \bar{\chi}_i \bar{\eta}_i + D_i \left(1-p_i^*\right) A_i^*}, \tag{3.17}$$

$$\mathrm{var}\left(\hat{p}_i^*\right) \ = \ \frac{p_i^*(1-p_i^*)\left(D_i + \bar{\chi}_i \bar{\eta}_i\right) \mu_{X_i}^\top \Sigma_i^{-1} \mu_{X_i}}{N_i \left[\bar{\chi}_i \bar{\eta}_i \mu_{X_i}^\top \Sigma_i^{-1} \mu_{X_i} + \left(D_i + \bar{\chi}_i \bar{\eta}_i\right) A_i^*\right]}.$$

If either $\bar{\chi}_i$ or $\bar{\eta}_i$ is equal to 0, only the animals caught at the ith SP contribute to the estimation of p_i^* and N_i and one gets the closed population model variances. If there is a single capture occasion for each SP, $A_i^* = 0$ and one gets open population variances.

The independence between closed population and Jolly-Seber estimators is the key for Theorem 1. Closed population models satisfying (3.7) yield such an independence. In a behavioural capture model, labelled M_b, the probability of capture changes after the first capture. This model cannot be expressed as (3.7), see for instance Rivest and Lévesque (2001). The estimation of population sizes in this context is investigated in the Supplementary Material, where Proposition S. 1. shows that Theorem 1 still holds in this special case.

3.5 Comparisons between the maximum likelihood estimator and the moment estimator of Kendall et al. (1995) for N_i

Kendall *et al.* (1995) estimate the parameters of a robust design by combining the likelihood for the underlying Jolly-Seber model with the conditional likelihoods for the closed population models within SPs. This construction leads to efficient estimators of all the parameters except the population sizes $\{N_i\}$ as these parameters do not appear in the combined likelihood. They proposed a moment-type estimator for the population sizes defined as

$$\hat{N}_i^{KEN} = \frac{n_i}{\hat{p}_i^*},\ i = 1, 2, \ldots, I, \tag{3.18}$$

where $\hat{p}_i^*$ is a maximum likelihood estimator. This section investigates the efficiency of $\hat{N}_i^{KEN}$ with respect to the combined estimator (3.16) when the model for the within SP capture-recapture experiments satisfies (3.7).

Proposition 4 *The asymptotic multinomial variance of* $\hat{N}_i^{KEN}$ *is*

$$var_M\left(\hat{N}_i^{KEN}\right) = \frac{N_i D_i \left(1 - p_i^*\right)}{p_i^* \bar{\chi}_i \bar{\eta}_i + D_i \left(1 - p_i^*\right) A_i^*} + \frac{N_i \left(1 - p_i^*\right)^2 A_i^* \bar{\chi}_i \bar{\eta}_i \left(p_i^*\right)^{-1}}{p_i^* \bar{\chi}_i \bar{\eta}_i + D_i \left(1 - p_i^*\right) A_i^*}. \tag{3.19}$$

Proposition 2 is proved in the Appendix. The first part on the right hand side of (3.19) gives the variance of the combined estimator provided in Equation (3.17) while the second part gives the augmentation. To compare the Kendall estimator for N_i to the weighted estimator, we calculate an efficiency index $\text{Eff}_i\left(KEN, COMB\right) = \text{var}_M(\hat{N}_i^{KEN})/\text{var}_M(\hat{N}_i^{COMB})$, where $\text{var}_M(\hat{N}_i^{COMB})$ is the weighted variance for N_i given in (3.17). Straightforward developments lead to

$$\text{Eff}_i\left(KEN, COMB\right) \quad = \quad 1 + \frac{A_i^* \left(1 - p_i^*\right)}{p_i^*} \times \frac{\bar{\chi}_i \bar{\eta}_i}{D_i}. \tag{3.20}$$

The second quantity on the right hand side of Equation (3.20) represents the gain in efficiency of the weighted estimator for N_i over that of Kendall *et al.* (1995). The quantity $\bar{\chi}_i \bar{\eta}_i / D_i$ is an increasing function of $\bar{\chi}_i$. Since $\bar{\chi}_i$, the probability for a captured unit to be seen subsequently, is an increasing function of ϕ_i, the efficiency gain is important when the survival probability ϕ_i is large.

3.6 Numerical investigations

This section presents a Monte Carlo study to investigate the gains in precision of the combined estimator $\hat{N}_i^{COMB}$ when compared with the moment estimator $\hat{N}_i^{KEN}$. It also investigate the performance of a simple variance estimator for $\hat{N}_i^{COMB}$ that is introduced in the next section. Then the methodology proposed in this paper is illustrated through the analysis of a dolphin capture-recapture data set.

3.6.1 Variance estimation

The combined estimator $\hat{N}_i^{COMB}$ is defined in (3.16); it is calculated using the multinomial variance estimators $v(\hat{N}_i^{CP})$ and $v(\hat{N}_i^{JS})$ available when fitting closed and open population models to the data; the Kendall estimator $\hat{N}_i^{KEN}$ is evaluated in a similar way using a combined estimator for p_i^*, see (3.18). In Section 3.6.2, we use a plug-in variance estimator given by $v(\hat{N}_i^{COMB}) = 1/\{1/v(\hat{N}_i^{JS}) + 1/v(\hat{N}_i^{CP})\}$. In a similar way, a plug in variance estimator $v(\hat{N}_i^{KEN})$ is obtained by plugging $\hat{N}_i^{KEN}$, the maximum likelihood estimator $\hat{p}_i^*$, $v(\hat{N}_i^{JS})$ and $v(\hat{N}_i^{CP})$ into (3.19).

3.6.2 Monte Carlo simulations

We conducted a simulation study to assess the importance of the precision gain with the maximum likelihood estimator for N_i and to investigate the sampling properties of the plug-in combined estimator $\hat{N}_i^{COMB}$, when compared to $\hat{N}_i^{RD}$ the maximum likelihood estimator for N_i discussed in Rivest and Daigle (2004). It is evaluated using the functions `robustd.0` of `Rcapture`, see Baillargeon and Rivest (2007), that also provides a multinomial linearization variance estimator, $v(\hat{N}_i^{RD})$. A major limitation of this estimator is that it cannot be calculated in large experiments with more than 20 capture occasions. The data were generated using stationary robust design model M_t^t with $I = 4$ SPs and $\ell = 5$ capture occasions in each one. The capture probabilities at each of the $\ell = 5$ occasions within a SP were set to $p = 0.1,\ 0.2,\ 0.4$ and the between SP survival probabilities to $\phi = 0.5,\ 0.9$. We set the expected number of arrivals to $B = 100,\ 200,\ 400$ when the survival probability is $\phi = 0.5$ and to $B = 20,\ 40,\ 80$ when $\phi = 0.9$ so that the stationary population size for the 2^{nd} SP , $N_2 = B/(1-\phi)$, the parameter to be estimated, is equal to 200, 400 and 800 respectively. The $\tilde{B}$s were generated using a Poisson distribution and the number of weeks of stay for a unit was simulated using a geometric distribution with parameter $1-\phi$. Under this set up, the stochastic process $\{\tilde{N}_i : i = 1, \ldots, 5\}$ is an homogeneous AR(1) Poisson process. We ran 500 repetitions for each set of parameter values. There was a burn-in period of 20 SPs for each replication. We calculated the relative bias and the multinomial root mean squared error of several estimators of N_2,

$$RMSE\left(\hat{N}_2\right) = \left\{\sum_r^{500}\left(\hat{N}_{2r} - \tilde{N}_{2r}\right)^2/500\right\}^{1/2},$$

where $\tilde{N}_{2r}$ is the population size for the second SP at the rth repetition. The simulation setup is similar to that of Kendall et al. (1995) so that the two sets of results can be compared. Because of the limited number of capture occasions, the robust design parameters were estimated using the algorithm of Rivest and Daigle (2004).

Table 3.1 reports the relative bias in percentage and the multinomial root mean square error for four estimators of N_2, the robust design maximum likelihood estimator, the combined

estimator defined in (3.16), the Kendall estimator and the Jolly-Seber estimator. Overall the four methods are similar in terms of bias, which is less than 1% for most estimators when the population size is larger than 400. At low catchability the differences between $\hat{N}_i^{RD}$, $\hat{N}_i^{COMB}$ and $\hat{N}_i^{KEN}$ are small. But when the capture and survival probabilities are high both the robust design and the combined estimators are more precise than the Kendall estimator; the loss of efficiency of the Kendall estimator (defined as 1 minus the ratio of the weighted estimator MSE over the Kendall estimator MSE) is sometimes larger than 50%. When the survival probability is equal to 0.9, the Kendall estimator is less precise than the Jolly-Seber estimator. This result was also reported in Table 2 of Kendall *et al.* (1995). Note also the slight loss of precision of the combined estimator with respect to the maximum likelihood estimator.

Table 3.2 compares the relative bias of the multinomial variance estimator $v(\hat{N}_2)$ and the coverage of the 95% confidence interval of $\hat{N}_2$ for the robust design estimator and the combined estimator. The relative bias for $E\{v(\hat{N}_2)\}$ is calculated as $RB[E\{v(\hat{N}_2)\}] = [E\{v(\hat{N}_2)\} - MSE(\hat{N}_2)]/MSE(\hat{N}_2)$; the 95% confidence interval for $\hat{N}_2$ is $\exp[\log(\hat{N}_2) \pm 1.96 v\{\log(\hat{N}_2)\}^{1/2}]$, where $v\{\log(\hat{N}_2)\} = v(\hat{N}_2)/\hat{N}_2^2$. The bias of the robust design estimator $v(\hat{N}_2^{RD})$ is less than 4% and the coverage of the 95% linearization confidence interval is within 1% of its nominal value. The plug-in variance estimator $v(\hat{N}_2^{COMB})$ underestimates $\text{var}_M(\hat{N}_i^{COMB})$ by up to 39% and the coverage of the 95% confidence interval is below the nominal value of 95% when the population size is 200. In these situations, the methodology presented in Section 5.2 of Yauck *et al.* (2018) could be used to calculate $\hat{N}_i^{RD}$ and estimate its variance where the log-linear model of Rivest and Daigle (2004) cannot be used. As the population size grows, the plug in variance estimator becomes more reliable. Its relative bias is less than 4% in most cases and the coverage of the 95% confidence interval is within less that 1% of its nominal value.

Table 3.1 – Simulation results for the estimation of N_2 under a stationary robust design model for M_t^t.

			$N = 200$		$N = 400$		$N = 800$	
p	ϕ	Method	RB	$RMSE$	RB	$RMSE$	RB	$RMSE$
		COMB	-3	36.64	-2.4	48.44	-1.2	66.95
	0.5	RD	2.5	34.64	<1	46.31	<1	65.31
		KEN	3.6	36.78	1.6	51.87	1.1	72.48
		JS	5.5	68.14	2	91.31	<1	116.12
0.1								
		COMB	-2.8	20.88	-1.6	31.23	<1	43.19
	0.9	RD	< 1	18.24	<1	28.71	<1	41.10
		KEN	1	21.67	<1	31.82	1	46.33
		JS	-1	21.73	<1	35.49	<1	52.24
		COMB	-1	14.78	<1	19.38	<1	26.49
	0.5	RD	<1	14.28	<1	19.04	<1	26.24
		KEN	<1	16.09	<1	21.64	<1	30.87
		JS	<1	21.44	<1	34.26	<1	47.04
0.2								
		COMB	<1	7.88	<1	12.09	<1	16.13
	0.9	RD	<1	7.63	<1	11.98	<1	16.09
		KEN	<1	8.92	<1	14.27	<1	20.13
		JS	<1	9.60	<1	13.80	<1	19.60
		COMB	<1	5.03	<1	5.83	<1	8.23
	0.5	RD	<1	4.21	<1	5.60	<1	8
		KEN	<1	5.37	<1	6.74	<1	9.38
		JS	<1	7.51	<1	10.55	<1	15.45
0.4								
		COMB	<1	2.18	<1	2.98	<1	4.30
	0.9	RD	<1	2.13	<1	2.96	<1	4.21
		KEN	<1	3.79	<1	5.77	<1	8.36
		JS	<1	2.47	< 1	3.40	<1	4.74

Table 3.2 – Simulation results for the estimation of $var(\hat{N}_2)$ under a stationary robust design model for M_t^t.

			$N = 200$		$N = 400$		$N = 800$	
p	ϕ	Method	RB	95%cov.	RB	95%cov.	RB	95%cov.
	0.5	COMB	-19.75	87.6	-14.26	90	-9.7	93.2
		RD	<1	93	<1	93.4	-2.4	95.4
0.1								
		COMB	-17.55	87.4	-18.12	89	-2.7	92.2
	0.9	RD	12.76	92.4	1.2	92.2	1.03	94.8
		COMB	-22.78	90	-7.6	93.2	-2	94
	0.5	RD	-12.64	92.8	-1.8	94.8	1.26	95
0.2								
		COMB	-2	93.2	-13.4	93	-1.79	94.8
	0.9	RD	8.8	95.2	-10.56	94.4	<1	95.2
		COMB	-39.6	86.4	-4.5	92.6	-3.5	95
	0.5	RD	-4	93.8	8	95.2	4.7	94.8
0.4								
		COMB	-11.5	92.4	-2.2	93.6	-7.4	94.2
	0.9	RD	-3.67	95.6	<1	95.2	-2.89	95.6

3.6.3 The analysis of a dolphin data set

This section analyzes the data from Santostasi *et al.* (2016) about a population of three odontocete species (striped dolphins, short-beaked common dolphins and common bottlenose dolphins) living in the Gulf of Corinth, a Mediterranean bay located in Greece. The data collection, which consisted of individual photo-identification of animals from boats, was carried out for 5 years (or SPs) from May 2011 to October 2015. Within a year, a secondary sampling period consisted of three or four consecutive months (May-September) of mark-recapture. A capture history is a list of months on which dolphins are photographed and both their dorsal fin sides properly identified. There were two data sets: one for striped and common dolphins and the other for the bottlenose dolphins; we analyzed the former. A total of 393 striped and common dolphins were identified over the 5 years study period. The goal of the analysis is to estimate yearly abundance for the striped and common dolphins.

Model M_t is selected within SPs following Santostasi *et al.* (2016). Yearly abundance estimates and standard errors are presented in Table 3.3. The results show that yearly abundance estimates for our weighted estimator and the Kendall estimator are similar. The plug in standard errors for $\hat{N}_i^{COMB}$ are smaller than those for $\hat{N}_i^{KEN}$. Considering the simulations of Section 7.2, these estimators are reliable as both the capture probabilities and the survival rates are high. Thus the gain in precision for the combined estimator is important for this data set. This agrees with the conclusion of the simulation study presented in Section 7.1. Santostasi *et al.* (2016) used MARK (White and Burnham, 1999) for their analysis. They

selected a model with equal yearly survival, estimated p_i^* with that model, and used (3.18) to estimate the population sizes. Their standard estimates are comparable to those for $\hat{N}_i^{KEN}$ reported in table 4.5; even if the combined estimator is calculated without accounting for the equal yearly survival it is is more precise than the estimator used by Santostasi *et al.* (2016).

Estimating population sizes is an important objective of capture-recapture experiments for odontocete species. These species have high catchability and high survival, see Silva and Silva JR (2009), Berrow *et al.* (2012), Tyne *et al.* (2014), and Santostasi *et al.* (2016). Thus they should use estimators of population sizes proposed in this paper as its precision can be much higher than that of the estimator of Kendall *et al.* (1995).

Table 3.3 – Abundance estimates for striped and common dolphins, under robust design model M_t^t.

SPs	$\hat{N}_i^{COMB}$	$SE\left(\hat{N}_i^{COMB}\right)$	$\hat{N}_i^{KEN}$	$SE\left(\hat{N}_i^{KEN}\right)$
2011	381	36.02	381	36.02
2012	323	7.70	326	9.70
2013	314	5.19	316	8.19
2014	342	7.73	359	9.52
2015	358	26.55	358	26.55

3.7 Discussion

In some applications of the robust design, such as in the dolphin analysis of Santostasi *et al.* (2016) and the analysis of mobile phone data presented in Chapter 2, the estimation of population sizes is the main objective and determining the best estimator is important. This work has shown a simple method to calculate the maximum likelihood estimator, by combining open and closed population estimators. It has also highlighted, in some instances, the poor efficiency of the simple moment estimator of Kendall *et al.* (1995).

This work could be extended in several directions. The construction of a robust design maximum likelihood estimator for N_i when the survival and the capture probabilities depend on explanatory variables is of interest given the poor performance of the moment estimator in some situations. Investigating whether the asymptotic independence property, between the Jolly-Seber and the closed population estimators, generalizes to situations where the capture probability depends on individual covariates would also be of interest.

Transition

Au chapitre précédent, on a construit un estimateur de la taille de la population qui est asymptotiquement équivalent à l'estimateur du maximum de vraisemblance, en combinant l'estimateur pour le modèle de population ouverte de Jolly-Seber présenté à la Section 1.2.2 et l'estimateur pour le modèle de population fermée présenté à la Section 1.1.5. Dans le chapitre qui suit, on considère une plus grande classe de modèles de population fermée présentée aux Sections 1.2.1 et 1.2.2. Dans le contexte du design robuste, on dérive l'estimateur du maximum de vraisemblance pour la taille de la population; on propose également trois méthodes d'estimation de la variance de l'erreur associée à l'estimateur. On démontre ensuite que l'estimateur du maximum de vraisemblance pour la taille de la population est plus efficace que l'estimateur des moments proposé par Kendall *et al.* (1995) et implémenté dans le logiciel MARK (White and Burnham, 1999).

Chapter 4

Capture-recapture Estimation of Population Sizes Under the Robust Design

Résumé

Ce travail concerne l'estimation de la taille de la population dans le contexte du design robuste, une expérience de capture-recapture imbriquée comportant deux niveaux d'échantillonnage. Le niveau primaire consiste en une étude de long terme dans laquelle la population est ouverte. Le niveau secondaire consiste en une étude de court terme, conduite à l'intérieur de chaque session d'échantillonnage du niveau primaire : la population est supposée fermée à des ajouts et diminutions. Dans ce contexte, on construit un estimateur du maximum de vraisemblance pour la taille de la population ; on propose également trois méthodes d'estimation de la variance de l'estimateur. Ce travail montre que l'estimateur ainsi construit est plus efficace que l'estimateur des moments proposé dans Kendall *et al.* (1995) et mis en oeuvre dans le logiciel MARK (White and Burnham, 1999). La perte de précision associée à l'estimateur des moments de Kendall ainsi que la performance des trois méthodes d'estimation de la variance de l'estimateur du maximum de vraisemblance sont évaluées via une étude de simulation. Ces résultats sont appliqués à travers un exemple sur l'estimation de la taille d'une population de dauphins vivant dans le Golfe de Corinthe (Grèce) et présenté dans Santostasi *et al.* (2016).

*M*ots-clé : Hétérogenéité, Réponse comportementale, Modèle de Jolly-Seber, Capture-recapture, Régression de Poisson, Design robuste.

Abstract

This work is concerned with the estimation of population sizes under Pollock's robust design, a nested mark-recapture experiment featuring two levels of sampling. The primary level consists of a long-term study in which the population is assumed open to addition and deletion. The secondary level consists of a short-term study within each session of the primary level; the population is assumed closed. In this context, we derive the maximum likelihood estimators for the population sizes and propose three methods of estimation for their variances. This work proves that the proposed maximum likelihood estimator is more efficient than the moment estimator provided in Kendall *et al.* (1995). The loss of precision associated with the Kendall estimator and the perfomance of the three methods of estimation for the variance of the maximum likelihood estimator are evaluated in a Monte Carlo study. This is used to analyze a data set about dolphin populations living in the Gulf of Corinth (Greece) and provided in Santostasi *et al.* (2016).

*K*eywords: Heterogeneity, Trap response, Jolly-Seber model, Mark-recapture study, Poisson regression, Robust design

4.1 Introduction

The estimation of population sizes using capture-recapture data is widely studied for open and closed population models. Jolly (1965) and Seber (1965) built a likelihood for the Jolly-Seber model under the assumption that all units are equally catchable and derived simple maximum likelihood estimators. Schwarz and Arnason (1996) proposed a generalization to the Jolly-Seber model by assuming that births come from a super-population that enters the experiment under a multinomial sampling. Cormack (1989) modeled births in a log-linear framework and built a Poisson likelihood. Inference on population sizes in closed population models is discussed in Fewster and Jupp (2009). This work deals with a hierarchical sampling scheme that combines both closed population and open population models. The robust design was introduced by Pollock (1982) and consists of a long-term study featuring a number of sampling periods (SP). Within each SP, a short-term study in which units are sampled over secondary capture occasions is conducted; the population is assumed to be closed. From one SP to the next, the population is assumed to be open to addition and deletion; a Jolly-Seber open population model applies.

In the context of the robust design, recent capture-recapture studies such as the analysis of american black bear populations (Pederson *et al.*, 2012), the analysis of dolphin populations (Santostasi *et al.*, 2016), the analysis of longsnout seahorse populations (Siquiera *et al.*, 2017) and the analysis of data on the activation of applications on mobile phones considered the estimation of population sizes as of primary interest. Kendall *et al.* (1995) built a framework to analyze capture-recapture data using the robust design. They constructed likelihood functions for a large class of models by combining the likelihood for the Jolly-Seber open population model to the likelihoods for the underlying closed population models within secondary capture

occasions; the class of closed population models considered in their work is presented in the seminal paper of Otis *et al.* (1978). The resulting likelihood yields efficient estimators for all demographic parameters except the population sizes as these parameters are not directly included in the likelihood. The robust design models developed in Kendall *et al.* (1995) are implemented in the statistical software MARK (White and Burnham, 1999); they are widely documented and used in the analysis of capture-recapture data.

This paper discusses the estimation of population sizes in the context of the robust design. We consider a broad class of robust design models that can be used to build likelihoods for the robust design data (Amstrup *et al.*, 2005): reduced-parameter models, multiple-group models, time-specific and individual covariate models, multiple-age models (Kendall *et al.* 1997; Kendall and Bjorkland 2001). Considering the likelihood functions for the complete data structure developed by Kendall *et al.* (1995), and applying a conditional argument, we derive the maximum likelihood estimators for the population sizes under the robust design. We show that the macimum likelihood population size estimators are more efficient than the moment-type estimators proposed in Equation (3) of Kendall *et al.* (1995). In Section 3, the precision gained by the maximum likelihood estimator for the population size over that of Kendall *et al.* (1995) is evaluated in a Monte Carlo study. We propose three methods of estimation for the uncertainty associated with the population size estimator: (1) a classical method that takes the Poisson variance of the estimator minus the population size estimator, (2) an alternative method inspired by Kackar and Harville (1984) way of splitting the estimation error and (3) a parametric bootstrap method similar to that of Rivest and Daigle (2004). In Section 4, we compare the relative bias of the three variance estimators using simulation. Section 5 presents a numerical example about the estimation of a population of dolphins living in the Gulf of Corinth (Greece) and discussed in Santostasi *et al.* (2016). The data collection was carried out from 2011 to 2015. Within each of the 4 years of survey, dolphins were sampled for three to four consecutive months. We analyzed the data using the robust design and compared our estimator of abundance to that obtained using the Kendall method and presented in table 4 of Santostasi *et al.* (2016).

4.2 Notation and assumptions

Let i denote SP i, $i = 1, 2, \ldots, I$ (I is the number of SPs), and let j denote a secondary capture occasion within a SP, $j = 1, 2, \ldots, \ell_i$; ℓ_i represents the number of secondary capture occasions within SP i. A unit in the population is given a unique tag when it is captured for the first time; it is re-identified in subsequent captures. For that unit, a capture history is defined as a sequence of 1's (capture) and 0's (miss) summarized in the $(\sum_i \ell_i) \times 1$ vector $\boldsymbol{\omega}$ of secondary capture information with entry $\omega_{ij} = 1$ if the unit has been captured on occasion j of SP i and 0 otherwise. The data set represents the frequencies $\{n_\omega\}$ for the $s = 2^{\sum \ell_i} - 1$ observable capture histories. We define the following statistics:

— u_i is the number of unmarked units captured during SP i;

— m_i is the number of marked units captured during SP i;

— $n_i = u_i + m_i$ is the number of units captured during SP i;

— v_i the number of units captured for the last time at SP i.

We also define the following parameters:

— The survival probability between SP i and $i+1$ is denoted $\phi_i \in (0,1)$ for all units in the population;

— The probability of being captured at least once during SP i is p_i^*;

— $N_i = U_i + M_i$ is the expected population at the start of the i^{th} SP, where U_i denotes the expected number of unmarked units in the population just before SP i, M_i denotes the expected number of marked units in the population just before SP i; it satisfies $M_i = (M_{i-1} + U_{i-1}p_{i-1}^*)\phi_{i-1}$ and $M_1 = 0$;

— $\eta_i = 1 - M_i/N_i$ represents the proportion of unmarked units just before SP i; $\bar{\eta}_i = 1 - \eta_i$.

Following Yauck and Rivest (2018), we assume that the stochastic process of the observed population sizes $\{\tilde{N}_i, i = 1, 2, ...\}$ is a non homogeneous autoregressive Poisson process of order 1 (AR(1)) (McKenzie, 1985). Under this model, the counts $\{n_\omega\}$ have independent Poisson distributions with mean values $\{\mu_\omega\}$. In many applications, we want to quantify the uncertainty associated with $\{\hat{N}_i\}$ as predictors for $\{\tilde{N}_i\}$. The notation $Var(\hat{N}_i)$ will be used to denote the variance of the prediction error $\hat{N}_i - \tilde{N}_i$ while $v(\hat{N}_i)$ denotes its estimate. This is further discussed in the next section.

4.3 Population size estimators for the robust design

In this section, we derive population size estimators for a broad class of robust design models, including reduced-parameter models and time-specific and individual covariate models. Such models are denoted $\{\phi()\, p(), \gamma()\}$ in MARK notation, where $\gamma()$ represents the model for temporary migration (Kendall and Nichols, 1997; Schwarz and Stobo, 1997). Without loss of generality, we assume that there is no temporary migration; the model notation reduces to $\{\phi()\, p()\}$. For models under the robust design, one must specify the variation in capture probability within and between SPs. This will be denoted $p(i.j)$, where i represents the variation within SPs while j represents the variation between SPs; if there is temporal (t) variation in catchability within and between SPs, the capture probability is denoted $p(t.t)$. To fit these models, one need to select the appropriate model for both survival and capture probabilities using model selection criterias such as the Akaike's Information Criterion (AIC). This issue, along with the estimation procedure, is discussed in the manual of the program MARK (Cooch and White, 2018). We also present three methods of estimation for the variances of the errors associated with the new population size estimators.

4.3.1 Derivation of the Kendall and the maximum likelihood population size estimators

The likelihood function for the robust design is the product of three components L_1, L_2 and L_3. As defined in Kendall *et al.* (1995), the likelihood functions L_1 and L_2 are that of the Jolly-Seber model; L_1 represents the model for the capture of unmarked units while L_2 represents the Cormack-Jolly-Seber model (Cormack 1964; Jolly 1965; Seber 1965). The component L_3 is related to the secondary information within each primary sampling period; it is given in Equation (1) of Kendall *et al.* (1995) for the time-dependent within and between SPs, and temporary trap response model. We first consider the saturated model $\{\phi(i)\, p(i.j)\}$ with capture and survival parameters $\{p_{ij}\}$ and $\{\phi_i\}$ respectively. The robust design parameters are estimated using the conditional approach presented in Kendall *et al.* (1995, page 5) and described as follows:

1. First, estimate $\{p_{ij}\}$ and $\{\phi_i\}$ using the conditional likelihood $L_2 \times L_3$, then compute $\{\hat{p}_i^*\}$ as a function of $\{\hat{p}_{ij}\}$. For example, a model with time-dependent capture probabilities within SPs gives $\hat{p}_i^* = \prod_{j=1}^{\ell_i}(1-\hat{p}_{ij})$. This step provides $\{\hat{p}_i^*\}$ and $\{\hat{\phi}_i\}$.
2. Given $\{\hat{p}_i^*\}$ and $\{\hat{\phi}_i\}$, compute the estimators for the number of unmarked units in the population just before SP i:

$$\hat{U}_i \;=\; \frac{u_i}{\hat{p}_i^*}, \quad i = 1, \ldots, I. \tag{4.1}$$

Estimator (4.1) is given in Equation (2) of Kendall *et al.* (1995). They also computed a moment-type estimator for M_i, namely $\hat{M}_i = m_i/\hat{p}_i^*\,(i = 1, 2, \ldots, I)$. Finally, they summed $\{\hat{U}_i\}$ and $\{\hat{M}_i\}$ to obtain the Kendall (KEN) estimators for $\{N_i\}$:

$$\hat{N}_i^{KEN} = \frac{n_i}{\hat{p}_i^*}, \quad i = 1, 2, \ldots, I. \tag{4.2}$$

We propose an alternative estimator for M_i as a function of the sufficient statistics $\{u_i\}$ using a recursive approach. For $i = 2$, $\hat{M}_2 = u_1\hat{\phi}_1$. For $i = 3$, $\hat{M}_3 = u_1\hat{\phi}_1\hat{\phi}_2 + u_2\hat{\phi}_2$. A general formula for the estimator of the number marked units in the population just before SP i is easily derived:

$$\hat{M}_i \;=\; \sum_{k=1}^{i-1} u_k \prod_{s=k}^{i-1} \hat{\phi}_s, \quad i = 2, \ldots, I. \tag{4.3}$$

Finally, summing Equations (4.1) and (4.3) gives the population size estimator:

$$\hat{N}_i \;=\; \frac{u_i}{\hat{p}_i^*} + \sum_{k=1}^{i-1} u_k \prod_{s=k}^{i-1} \hat{\phi}_s, \quad i = 2, \ldots, I; \quad \hat{N}_1 = u_1/\hat{p}_1^*. \tag{4.4}$$

This represents the maximum likelihood estimator (MLE) for N_i. An alternative construction for this estimator, using a log-linear parametrization, is provided in Rivest and Daigle (2004). We rewrite Equation (4.4) as

$$\hat{N}_i \;=\; \boldsymbol{u^i}\hat{\boldsymbol{\beta}}_{\boldsymbol{i}}, \tag{4.5}$$

where $\boldsymbol{u^i} = (u_1, u_2, \ldots, u_i)$ and

$$\hat{\boldsymbol{\beta}}_{\boldsymbol{i}} = \left(\prod_{s=1}^{i-1} \hat{\phi}_s, \ldots, \prod_{s=i-1}^{i-1} \hat{\phi}_s, 1/\hat{p}_i^* \right)^\top \tag{4.6}$$

is a $i \times 1$ column vector. Equation (4.5) represents the MLE of N_i for the general model $\{\phi(i)\, p(i.j)\}$. For special cases, one need to specify the form of (4.6). For the reduced-parameter model $\{\phi(.)\, p(i.j)\}$, $\phi_1 = \phi_2 = \cdots = \phi_{I-1} = \phi$, and (4.6) simplifies to

$$\hat{\boldsymbol{\beta}}_{\boldsymbol{i}} = \left(\hat{\phi}^{i-1}, \hat{\phi}^{i-2}, \ldots, \hat{\phi}, 1/\hat{p}_i^* \right)^\top.$$

In practice, the MLE (4.5) is computed as follows:

1. fit a robust design model using MARK and obtain estimates for the demographic parameters;
2. calculate $\hat{\boldsymbol{\beta}}_{\boldsymbol{i}}$ using the estimates obtained in step 1.
3. Plug $\hat{\boldsymbol{\beta}}_{\boldsymbol{i}}$, obtained in step 2, into (4.5).

The proposed MLE for N_i can be computed for virtually any robust design model available in MARK. Yauck and Rivest (2018) showed that the KEN estimator (4.2) is not as efficient as the MLE estimator under a saturated robust design model. In Section 4, we investigate the efficiency of the KEN estimator, defined as 1 minus the ratio of the MLE estimator mean squared error (MSE) over the KEN estimator MSE, for time-independent survival models.

The special case of the Jolly-Seber model

When there is a single capture occasion within each SP ($\ell_1 = \ell_2 = \cdots = \ell_I = 1$), one obtains the Jolly-Seber mark-recapture experiment. The Jolly-Seber model, with time dependent survival and capture probabilities, is denoted $\{\phi(t)\, p(t)\}$. The estimator for M_i under this model is

$$\hat{M}_i = m_i + \frac{n_i \{\sum_{s=1}^{i-1} (u_s - v_s) - m_i\}}{n_i - v_i},$$

while $\hat{p}_i^* = m_i/\hat{M}_i$ and $\hat{\phi}_i = \hat{M}_{i+1}/(\hat{M}_i + u_i)$. Plugging $\{\hat{p}_i^*\}$ and $\{\hat{\phi}_i\}$ into (4.2) and (4.4) gives

$$\hat{N}_i = \frac{n_i \left[(n_i - u_i)(n_i - v_i) + n_i \{\sum_{s=1}^{i-1} (u_s - v_s) - n_i + u_i\} \right]}{(n_i - u_i)(n_i - v_i)},\ i = 2, \ldots, I-1,$$

the Jolly-Seber estimator for N_i given in Equation (23) of Jolly (1965).

4.3.2 The estimation of the variance of $\hat{N}_i - \tilde{N}_i$

The variance of (4.5) under a Poisson sampling, $\mathrm{Var}_P(\hat{N}_i)$, can be obtained using a linearization argument. It is given by

$$\mathrm{Var}_P(\hat{N}_i) = N_i\eta_i/p_i^* + \sum_{k=1}^{i-1} N_k\eta_k p_k^* \left(\prod_{s=k}^{i-1}\phi_s\right)^2 + E\left\{\boldsymbol{u^i}\left(\boldsymbol{\nabla\hat{\beta}_i}\right)^\top \boldsymbol{A}\left(\boldsymbol{\{\hat{\phi}_i\}, \hat{p}_i^*}\right)\boldsymbol{\nabla\hat{\beta}_i}\left(\boldsymbol{u^i}\right)^\top\right\}, \tag{4.7}$$

where $\boldsymbol{A}\left(\boldsymbol{\{\hat{\phi}_i\}, \hat{p}_i^*}\right)$ is the variance-covariance matrix of $(\hat{\phi}_1, \ldots, \hat{\phi}_{i-1}, \hat{p}_i^*)$ and $\boldsymbol{\nabla\hat{\beta}_i}$ is the limit of the $i \times i$ matrix of partial derivatives of $\boldsymbol{\beta_i}$ with respect to $(\hat{\phi}_1, \ldots, \hat{\phi}_{i-1}, \hat{p}_i^*)$. The developments that led to (4.7) are presented in the Appendix. The estimated Poisson variance of $\hat{N}_i$, obtained by replacing $\{p_i^*\}$ and $\{\phi_i\}$ with their estimates $\{\hat{p}_i^*\}$ and $\{\hat{\phi}_i\}$ and the variance-covariance matrix $\boldsymbol{A}\left(\boldsymbol{\{\hat{\phi}_i\}, \hat{p}_i^*}\right)$ by its estimate (provided in standard computation programs that perform likelihood maximization), will be denoted $v_P(\hat{N}_i)$.

The first method of estimation for the variance of $\hat{N}_i - \tilde{N}_i$, labeled classical (or CLA), gives

$$v_{CLA}(\hat{N}_i) = v_P(\hat{N}_i) - \hat{N}_i. \tag{4.8}$$

Yauck and Rivest (2018) proved that (4.8) gives an asymptotic unbiased estimator when the model is saturated.

Now, let $N_i^* = \boldsymbol{u^i\beta_i}$; it represents the best linear unbiased predictor of $\tilde{N}_i$ if we assume that $\{p_i^*\}$ and $\{\phi_i\}$ are known. The prediction error $\hat{N}_i - \tilde{N}_i$ has two components: one due to the estimation of the parameters $(\hat{N}_i - N_i^*)$ and one due to the stochasticity of $\tilde{N}_i$ $(N_i^* - \tilde{N}_i)$. This way of splitting the error is presented in Kackar and Harville (1984). Under this approach, one have

$$\mathrm{Var}\left(\hat{N}_i - \tilde{N}_i\right) = \mathrm{Var}\left(\hat{N}_i - N_i^*\right) + \mathrm{Var}\left(N_i^* - \tilde{N}_i\right) + 2\mathrm{cov}\left(\hat{N}_i - N_i^*, N_i^* - \tilde{N}_i\right). \tag{4.9}$$

The covariance on the right hand side of (4.9) is generally intractable since $\hat{N}_i$, $\tilde{N}_i$ and N_i^* are correlated; this issue is briefly discussed in Kendall *et al.* (1995) and in Schwarz and Arnason (1996). Kackar and Harville (1984) showed that in the context of mixed linear models, and under certain regularity conditions, $\hat{N}_i - N_i^*$ and $N_i^* - \tilde{N}_i$ are independently distributed and that their covariance is equal to 0. Summing $v(\hat{N}_i - N_i^*)$ and $v(N_i^* - \tilde{N}_i)$, the respective estimations for $\mathrm{Var}(\hat{N}_i - N_i^*)$ and $\mathrm{Var}(N_i^* - \tilde{N}_i)$ in (4.9), gives the alternative (ALT) method of estimation for the variance of the prediction error, namely

$$v_{ALT}(\hat{N}_i) = \underbrace{v_P(\hat{N}_i) - \hat{N}_i}_{v_{CLA}(\hat{N}_i)} + 2\sum_{k=1}^{i-1} \hat{N}_k\hat{\eta}_k\hat{p}_k^* \prod_{s=k}^{i-1}\hat{\phi}_s\left(1 - \prod_{s=k}^{i-1}\hat{\phi}_s\right). \tag{4.10}$$

Estimator (4.10) offers estimates of the variance of $\hat{N}_i - \tilde{N}_i$ that are bigger than those obtained with the CLA method. Its performance will be investigated in Section 4 using simulation.

4.3.3 Parametric bootstrap for the estimation of the variance of $\hat{N}_i - \tilde{N}_i$

The third method of estimation for the variance of the prediction error is to run a parametric bootstrap. We consider the saturated model $\{\phi(i)\, p(i.j)\}$ The procedure is described as follows:

1. Calculate the parameter estimates $\{\hat{p}_{ij}\}, \{\hat{\phi}_i\}, \{\hat{N}_i\}$ and $\{\hat{B}_i\}$ using the two-stage estimation procedure described in Section 4.3;
2. Set the parameters $\{p_{ij}\}, \{\phi_i\}, \{N_i\}$, $\{B_i\}$ equal to the estimates obtained in step 1 and simulate the individual capture histories as follows:
   - 2.1 Generate the number of units in the population at the start of the 1^{st} SP, $\tilde{N}_1$, using a Poisson($\hat{N}_1$);
   - 2.2 Generate the number units from $\tilde{N}_1$ that survive between the 1^{st} and the 2^{nd} SP. The probability that a unit survives between the 1^{st} and the 2^{nd} SP is $\hat{\phi}_1$, and the number of units that survive is $\tilde{N}_{12}^{\hat{\phi}_1}$;
   - 2.3 Generate the individual capture histories within the first SP for each of the $\tilde{N}_1$ units using the model selected in step 1 of the estimation procedure;
   - 2.4 Generate the number of units that enter the population between the 1^{st} and the 2^{nd} SPs, $\tilde{B}_1$, using a Poisson($\hat{B}_1$), and calculate $\tilde{N}_2$ by summing $\tilde{B}_1$ and $\tilde{N}_{12}^{\hat{\phi}_1}$ (calculated in step 2.2);
   - 2.5 Repeat steps 2.2 and 2.3, then repeat steps 2.4 and 2.5 for each SP $i\,(i = 1, 2, \ldots, I)$.
3. For each simulated sample $r\,(r = 1, 2, \ldots, R)$, calculate the parameter estimates $\{\hat{N}_{ir}\}$ as in step 1.

The bootstrap variance of $\hat{N}_i - \tilde{N}_i$ is $v_{BOOT}(\hat{N}_i) = \{\sum_r (\hat{N}_{ir} - \tilde{N}_{ir})^2 / R\}^{1/2}$, where $\tilde{N}_{ir}$ is the realisation of the random population size for the i^{th} SP of the r^{th} repetition. This procedure can be adapted for any robust design model. The performance of the bootstrap method will be investigated in Section 4 with the CLA and the ALT methods.

4.4 Numerical investigations

We conducted a simulation study to (1) compare the bias and the precision of the maximum likelihood estimator (MLE) and the Kendall (KEN) estimators for N_i and to (1) assess the accuracy of the classical (CLA), the alternative (ALT) and the bootstrap (BOOT) methods of estimation for the variance of $\hat{N}_i - \tilde{N}_i$ under the robust design. Without loss of generality, we assume that there is no temporary emigration between SPs (Kendall *et al.*, 1995). The data were generated using two robust design models with $I = 5$ SPs and $\ell = 7$ occasions within each SP:

(1) $\{\phi(.)\, p(t.t)\}$ with time variation in catchability within and between SPs and constant survival;

(2) $\{\phi(.)\, p(.b)\}$ with trap response within SPs and constant survival.

For simplicity of notations, models (1) and (2) will be denoted M_t^t and M_b^0 respectively (the subscript denotes variation in catchability within SPs while the superscript denotes variation between SPs; 0 represents no variation in catchability). For model M_t^t, the weekly capture probabilities for each SP were set to $p^* = 0.3, 0.5$, corresponding to average daily capture probabilities $p = 0.05, 0.1$. For model M_b^0, the vector of the daily capture and recapture probabilities was set to $(p, c) = (0.05, 0.3)$ and $(p, c) = (0.1, 0.3)$, corresponding to average weekly capture probabilities $p^* = 0.3, 0.5$. The between SPs survival probabilities were set to $\phi = 0.5, 0.9$ for both models. The initial population size was set to $N_1 = 200, 400, 800$ and the expected number of arrivals at each SP is set to compensate for mortality, so that the expected population sizes remain constant. The $\tilde{B}s$ were generated using a Poisson distribution and the number of SPs for a unit's stay was generated using a geometric distribution with parameter $1 - \phi$. The stochastic process $\{\tilde{N}_i : i = 1, 2, \ldots, 5\}$ obtained in this set up is an autoregressive of order 1 (AR(1)) Poisson process. We ran 500 repetitions for each set of parameter values.

We fitted models $\{\phi(.)\, p(t.t)\}$ and $\{\phi(.)\, p(.b)\}$ using the programming language R, package RMark (Laake, 2013) to build robust design models from program MARK. For details on how to fit capture-recapture models using package RMark, see Laake (2013).

4.4.1 Comparison of the maximum likelihood and the Kendall estimators of N_i

In this section, we compare the bias and the precision of the MLE and the KEN estimators for N_3. We calculated the bias of $\hat{N}_3$ as $RB(\hat{N}_3) = \{\sum_r (\hat{N}_{3r} - \tilde{N}_{3r})/\tilde{N}_{3r}\}/500$ and its root mean squared error as $RMSE(\hat{N}_3) = \{\sum_r (\hat{N}_{3r} - \tilde{N}_{3r})^2/500\}^{1/2}$, where $\tilde{N}_{3r}$ is the realisation of the random population size for the 3^{rd} SP of the r^{th} repetition. The results for models M_t^t and M_b^0 are reported in Tables 4.1 and 4.2 respectively. The two estimators are similar in terms of relative bias, which is less than 1% when $N = 400, 800$; for model M_b^0 the bias for the KEN estimator is 37% larger than the MLE estimator when $(p, c) = (0.1, 0.3)$, $\phi = 0.9$ and $N = 200$. When the survival probability is large, our estimator is more precise than the Kendall estimator; the loss of efficiency of the Kendall estimator in this scenario is, in some cases, larger than 25% for model M_b^0.

Table 4.1 – Simulation results for the estimation of N_3 under the robust design model M_t^t.

			$N = 200$		$N = 400$		$N = 800$	
p	ϕ	Method	RB(%)	RMSE	RB(%)	RMSE	RB(%)	RMSE
	0.5	MLE	1.72	46.73	<1	54.73	<1	76.72
		KEN	1.85	46.84	<1	55.05	<1	77.86
0.05								
	0.9	MLE	<1	27.36	<1	38.89	<1	53.46
		KEN	<1	28.08	<1	40.03	<1	54.79
	0.5	MLE	<1	18.41	<1	27.57	<1	37.01
		KEN	<1	18.79	<1	28.17	<1	38.25
0.1								
	0.9	MLE	<1	10.74	<1	14.26	<1	18.07
		KEN	<1	12.78	< 1	17.36	<1	20.46

Table 4.2 – Simulation results for the estimation of N_3 under the robust design model M_b^0.

			$N = 200$		$N = 400$		$N = 800$	
(p, c)	ϕ	Method	RB(%)	RMSE	RB(%)	RMSE	RB(%)	RMSE
	0.5	MLE	-2.96	43.49	<1	63.95	<1	92.78
		KEN	-3.41	44.59	<1	64.88	-1.12	94.09
(0.05, 0.3)								
	0.9	MLE	-1.20	25.36	<1	38.93	<1	50.06
		KEN	-2.18	29.18	<1	43.07	<1	56.85
	0.5	MLE	-1.43	19.25	<1	22.94	<1	38.16
		KEN	-2.03	20.19	<1	26.03	<1	43.18
(0.1, 0.3)								
	0.9	MLE	-8.31	12.07	<1	16.33	<1	19.26
		KEN	-11.07	15.18	<1	22.16	<1	24.03

4.4.2 Comparison of the classical, alternative and the bootstrap methods of estimation for the variance of $\hat{N}_3 - \tilde{N}_3$

In this section, we compare the CLA, the ALT and the BOOT methods of estimation for the variance of $\hat{N}_3 - \tilde{N}_3$. The bias for $E\{v(\hat{N}_3)\}$, where $v(\hat{N}_3)$ is the estimated MSE for $\hat{N}_3$, is computed as $RB[E\{v(\hat{N}_3)\}] = [E\{v(\hat{N}_3)\} - MSE(\hat{N}_3)]/MSE(\hat{N}_3)$ and the 95% confidence interval for $\hat{N}_3$ is $\exp[\log(\hat{N}_3) \pm 1.96 v\{\log(\hat{N}_3)\}^{1/2}]$, where $v\{\log(\hat{N}_3)\} = v(\hat{N}_3)/\hat{N}_3^2$. Tables 4.3 and 4.4 report the relative bias of $v(\hat{N}_3)$ and the coverage of the 95% confidence interval of $\hat{N}_3$ for models M_t^t and M_b^0 respectively.

For the time variation model M_t^t, the CLA method underestimates the MSE by up to 11% while the ALT method overestimates the MSE by up to 13% when $p = 0.05$; the coverage of the 95% confidence interval is within 2% of its nominal value. The BOOT estimator overestimates the MSE by up to 6% and the coverage of the 95% confidence interval is within

1% of its nominal value When $p = 0.1$ and $\phi = 0.5$, the CLA method underestimates the MSE by 14% on average with a 95% confidence interval within 2% of its nominal value; the ALT method overestimates the MSE by 9% on average with a coverage of the 95% confidence interval within 1% of its nominal value; the BOOT method overestimates the MSE by 3% on average with a coverage of the 95% confidence interval within 1% of its nominal value. When $p = 0.1$ and $\phi = 0.9$, both the CLA and the ALT methods give accurate approximations (bias of 4% in absolute value and on average with a 95% confidence interval within 1% of its nominal value) while the BOOT method gives a better estimation (bias of 2% on average with a 95% confidence interval within 1% of its nominal value). For model M_b^0, when $\phi = 0.5$ and $N = 800$, the CLA approximation underestimates the MSE by up to 3% on average and coverage of the 95% confidence interval is within 2% of its nominal value on average; the ALT approximation does much better by overestimating the MSE by up to 2% on average and the coverage of the 95% confidence interval is within 1% of its nominal value on average.

Overall, the BOOT method performs better than the ALT and the CLA methods; it becomes more reliable (bias less than 1% and a nominal value of 95% for the coverage of the 95% confidence interval) as the population size increases. The ALT method performs better than the CLA method for almost all sets of parameter values, particularly for small population sizes; it becomes more reliable as the population size increases. The ALT method can be used alongside the BOOT method for medium and large population sizes as will be seen in Section 4.5. In practice, the BOOT method takes some time to run ($\sim$ 15 minutes for 500 bootstrap repetitions) and generates several MARK files ($\sim$ 4 files for each bootstrap repetition). In this case, the ALT method offers the best compromise between running time and precision.

Table 4.3 – Simulation results for the estimation of the MSE of $\hat{N}_3$ under the robust design model M_t^t.

			$N = 200$		$N = 400$		$N = 800$	
p	ϕ	Method	RB(%)	95%cov.	RB(%)	95%cov.	RB(%)	95%cov.
0.05	0.5	CLA	-7.09	93.60	-2.98	93.40	-3.08	94.00
		ALT	6.14	94.00	<1	93.80	<1	95.00
		BOOT	2.16	94.00	<1	94.00	<1	95.00
	0.9	CLA	-11.07	93.60	-9.38	93.60	-5.07	94.00
		ALT	10.24	95.40	12.41	94.40	10.48	94.40
		BOOT	6.37	94.20	1.08	94.00	<1	94.80
0.1	0.5	CLA	-12.04	93.60	-11.42	93.60	-15.08	92.20
		ALT	13.17	95.60	4.08	95.40	6.14	94.00
		BOOT	6.19	94.80	<1	94.00	<1	95.00
	0.9	CLA	-4.61	94.40	-5.01	95.00	-2.68	94.80
		ALT	5.14	94.20	6.11	94.80	2.05	94.80
		BOOT	3.64	94.20	<1	95.00	<1	95.00

Table 4.4 – Simulation results for the estimation of the MSE of $\hat{N}_3$ under the robust design model M_b^0.

			$N = 200$		$N = 400$		$N = 800$	
(p, c)	ϕ	Method	RB(%)	95%cov.	RB(%)	95%cov.	RB(%)	95%cov.
(0.05, 0.3)	0.5	CLA	-3.25	92.80	-5.01	93.20	-2.07	94.20
		ALT	7.02	93.20	4.19	94.00	<1	94.80
		BOOT	4.22	93.00	2.05	95.00	<1	95.00
	0.9	CLA	-2.39	93.40	<1	95.00	<1	95.00
		ALT	2.38	94.80	<1	94.80	<1	95.00
		BOOT	1.78	94.40	<1	94.80	<1	95.00
(0.1, 0.3)	0.5	CLA	-6.02	93.20	<1	94.40	<1	94.00
		ALT	5.32	95.00	<1	94.80	<1	94.80
		BOOT	2.59	94.80	<1	95.00	<1	95.00
	0.9	CLA	-6.21	92.80	-11.07	92.20	<1	94.00
		ALT	8.31	94.80	5.12	94.40	<1	95.00
		BOOT	3.16	94.40	<1	95.00	<1	95.00

4.5 Case Study

To illustrate these findings, we analyzed a data set from a 5-year study of three odontocete species (*S*tenella coeruleoalba or "striped", *D*elphinus delphis or "common" and *T*ursiops truncatus or "bottlenose"). Within each year, dolphins were individually photo-identified from boats for three or four consecutive months. This study analyzes the striped and common dolphins data using the robust design. Following Santostasi *et al.* (2016), we fitted a time variation model within SPs and equal yearly survival: $\{\phi(.)\, p(t.t)\}$. We compared our estimates to those obtained with the Kendall method. Abundance estimates are computed following the steps described in Section 2. Standard errors are calculated using the CLA, the ALT and the BOOT methods. The results are presented in Table 4.5. For the 1^{st} SP (year 2011), the MLE and the KEN abundance estimates are equal since $\hat{N}_1^{MLE} = \hat{N}_1^{KEN} = u_1/\hat{p}_1^*$. The MLE and the KEN abundance estimates are similar. The bootstrap standard errors for the MLE are smaller than those obtained using the KEN estimator as seen in Section 4.4.1 of the Monte Carlo study; the loss of efficiency is larger than 30% for SPs 2012 to 2014 on average. For the MLE estimator, the bootstrap standard errors and those obtained using both the ALT and the CLA methods of approximation for years 2012-2014 are similar; the ALT method gives bigger estimates while the CLA underestimates the true variance, as seen in Section 4.4.2 of the Monte Carlo study.

Table 4.5 – Abundance estimates for striped and common dolphins, under robust design model $\{\phi(.)\, p(t.t)\}$.

SPs	$\hat{N}_i^{MLE}$	$SE(\hat{N}_i^{MLE})^{a}$	$SE(\hat{N}_i^{MLE})^{b}$	$SE(\hat{N}_i^{MLE})^{c}$	$\hat{N}_i^{KEN}$	$SE(\hat{N}_i^{KEN})^{a}$	$SE(\hat{N}_i^{KEN})^{d}$
2011	374	19.80	34.59	34.59	374	19.83	34.38
2012	329	7.90	8.21	6.82	326	10.66	13.58
2013	325	5.67	7.36	3.85	318	8.79	13.80
2014	341	7.90	8.48	5.87	356	10.51	14.00
2015	344	8.66	9.83	8.60	343	11.71	18.55

[a] Standard error obtained by bootstrap.
[b] Standard error computed using the ALT method of approximation.
[c] Standard error computed using the CLA method of approximation.
[d] Standard error computed using MARK.

4.6 Discussion

In this work, we have developed a maximum likelihood population size estimator for the robust design. We have compared the accuracy and the precision of the maximum likelihood estimator with that of Kendall *et al.* (1995). The results have shown that our method yields more precise estimates than those obtained with the Kendall method. We have also proposed three methods of estimation for the uncertainty associated with the maximum likelihood estimator: a classical method discussed in Chapter 3, an alternative method inspired by Kackar and Harville (1984) way of splitting the estimation error and a parametric bootstrap. The results show that the bootstrap method gives estimates that are more precise than those obtained using both the classical and the alternative methods, particularly for small population sizes. For medium and large population sizes, the alternative method performs as well as the bootstrap method.

The maximum likelihood population size estimates and the associated variances can be computed for virtually any robust design model available in the software MARK White and Burnham (1999). This new method of estimation for population sizes under the robust design is implemented in the programming language R, package RMark Laake (2013), as a complement to MARK. We have provided, as a Supplementary Material, the programs implementing most of the computations presented in this work.

Conclusion

Dans de nombreuses applications des méthodes de capture-recapture, notamment en biologie, en démographie et sciences sociales, en sciences médicales et en applications mobiles, l'estimation de la taille de la population est définie comme étant l'objectif principal. La collecte des données dans le cadre de ces applications s'opère de plus en plus selon une structure d'échantillonnage imbriquée, qui combine les méthodes de population fermée sur des occasions de capture secondaires (qui peuvent être des jours ou mois consécutifs) à l'intérieur des périodes primaires (qui peuvent être des mois ou années consécutifs), et de population ouverte d'une période primaire à une autre. Dans cette thèse, on a proposé une méthodologie d'estimation de la taille de la population et de l'incertitude associée aux estimateurs obtenus dans le contexte du design robuste.

La méthodologie d'estimation des paramètres du design robuste proposée dans cette thèse offre plusieurs avantages. D'abord, la procédure d'estimation séquentielle présentée au chapitre 2 ne nécessite pas la maximisation d'une fonction de vraisemblance est facile à mettre en œuvre; elle constitue une alternative aux méthodes traditionnelles d'estimation des paramètres du design robuste qui se heurtent aux gros problèmes mettant en jeu plus de 20 occasions de capture. Ensuite, l'estimateur du maximum de vraisemblance pour la taille de la population proposée au chapitre 4, au-delà d'être facile à implémenter dans les logiciels d'estimation standard, constitue une alternative à l'estimateur proposé par Kendall *et al.* (1995) et implémenté dans le logiciel MARK (White and Burnham, 1999), très prisé par les praticiens du domaine. Enfin, les méthodes d'estimation de la variance des estimateurs du design robuste, y compris les deux versions du bootstrap paramétrique présentées aux chapitres 2 et 4 respectivement, constituent une approche simple et flexible qui vient compléter la littérature sur l'estimation de l'incertitude associée à l'estimation de la taille de la population.

La procédure d'estimation séquentielle développée au chapitre 2 trouve son avantage lorsqu'on travaille avec des données de capture-recapture en grande dimension; il s'agit par exemple de données d'activation d'applications dans les téléphones intelligents, constituées d'enregistrements des activations d'à peu près 10.000 individus sur 76 semaines consécutives. Les difficultés liées à la manipulation d'une telle quantité d'information par les logiciels standard d'ajustement de modèles linéaires généralisés prouvent l'importance du nouvel algorithme d'estimation

des paramètres du design robuste. Outre la simplicité computationnelle de l'algorithme d'estimation, la nouvelle technique de calcul de la variance des estimateurs par bootstrap paramétrique rend la méthologie facile à implémenter dans la plupart des logiciels standard de calcul.

Dans le contexte du design robuste, le théorème d'indépendance asymptotique des estimateurs de la taille de la population obtenus avec le modèle de Jolly-Seber (Jolly 1965; Seber 1965) et le modèle de population fermée, démontré au chapitre 3, offre principalement deux avantages. Premièrement, ce résultat montre qu'on peut calculer, de façon très simple, un estimateur de la taille de la population asymptotiquement équivalent à l'estimateur du maximum de vraisemblance du design robuste par une simple pondération des estimateurs de population ouverte et de population fermée. Deuxièment, les simulations ont montré que l'estimateur de la variance pour la taille de la population, obtenue par simple combinaison des variances de Jolly-Seber et de population fermée, a des propriétés asymptotiques très intéressantes; il peut donc être substitué à l'estimateur obtenu par bootstrap paramétrique lorsqu'on travaille par exemple avec des données de capture-recapture en grande dimension.

Le chapitre 4 est une extension du chapitre 3 à des scénarios de modélisation de la survie et/ou de la capture des unités. L'estimateur de la taille de la population proposée dans ce contexte représente l'estimateur du maximum de vraisemblance pour les modèles de design robuste développés par Kendall *et al.* (1995). L'estimateur de la variance de l'erreur associée à l'estimateur du maximum de vraisemblance pour la taille de la population, inspirée de l'approche de décomposition de l'erreur de prédiction discutée par Kackar and Harville (1984), a des propriétés asymptotiques souvent meilleures que celui obtenu avec l'approche classique inspirée de Jolly (1965) et Schwarz and Arnason (1996). Les programmes informatiques développés dans le cadre de ce travail sont complémentaires au logiciel MARK, rendant la méthodologie facile d'implémentation pour les praticiens.

Les travaux effectués dans le cadre de cette thèse peuvent être étendus dans plusieurs directions. La méthodologie d'estimation proposée au chapitre 2 concerne la classe des modèles de design robuste considérés dans Rivest and Daigle (2004). Il serait intéressant de généraliser la méthodologie à la classe des modèles considérés dans Kendall *et al.* (1995). La généralisation de la propriété d'indépendance asymptotique entre les estimateurs de Jolly-Seber et de population fermée à des situations où la probabilité de capture dépend de covariables individuelles serait d'un grand intérêt.

Annexe A

Arguments techniques et matériel supplémentaire du chapitre 2

A.1 Arguments techniques

The following notation is used throughout the calculations. We distinguish (*i*) the notation used to describe the data themselves, (*ii*) important summary statistics and (*iii*) the robust design parameters for its analysis.

Capture-recapture data

- Subscript i denotes primary sampling period i, $i = 1, \ldots, I$;
- Subscript j denotes a secondary capture occasion within a PSP, $j = 1, ..., \ell_i$; ℓ_i is the number of secondary capture occasions within PSP i ;
- $\boldsymbol{\omega}$ is then a $\sum \ell_i \times 1$ vector containing for secondary capture occasion j in PSP i the outcome $\omega_{ij} = 1$ if the unit has been captured on that occasion and 0 otherwise ;
- The between PSP capture information can be summed up in a $I \times 1$ vector $\boldsymbol{\delta}$ which has entry $\delta_i = 1$ if the unit has been captured at least once during PSP i, that is if $\sum_j \omega_{ij} > 0$, and $\delta_i = 0$ otherwise ; $\boldsymbol{\delta}(\boldsymbol{\omega})$ denotes the $I \times 1$ between PSP capture history derived from $\boldsymbol{\omega}$.

Statistics

- The frequency of the number of units that have capture history $\boldsymbol{\omega}$ is denoted n_ω ;
- u_i is the number of unmarked units captured during PSP i ;
- m_i represents the number of marked units captured during PSP i ;
- $n_i = u_i + m_i$ is the number of units captured during PSP i ;
- v_i is the number of units captured for the last time at PSP i ;

— $w_i = \sum_{j=1}^{i-1}(u_j - v_j)$ is the number of units captured at least once during the first $i-1$ PSPs that will be seen at least once more, either in PSP i or later ;

— z_i represents the number of units captured before PSP i, not seen at PSP i, but captured subsequently ;

— n is the total number of units captured at least once during the whole sampling process.

Parameters

— The survival probability, for all units in the population, between primary periods i and $i+1$ is denoted $\phi_i \in (0,1)$;

— The probability of being captured at least once during PSP i is denoted $p_i^* = Pr(\delta_i = 1)$, this depends on the closed population model describing the captures within PSP i ;

— the probability of not being seen after PSP i, χ_i, satisfies the following recursive relationship, $\chi_i = (1-\phi_i) + \phi_i(1-p_{i+1}^*)\chi_{i+1}$; $\bar{\chi}_i = 1 - \chi_i$ is the probability of being seen after PSP i ;

— $N_i, i = 1, \ldots, I$ is the expected population at the start of the i^{th} PSP ;

— B_i is the expected number of new units joining the population before the start of PSP $i+1$ such that $N_{i+1} = N_i\phi_i + B_i$.

— The expected number of unmarked units in the population just before PSP i, U_i, satisfies $U_i = U_{i-1}\,(1-p_i^*)\,\phi_{i-1} + B_{i-1}$ for $i = 1, ..., I-1$; $U_1 = N_1$;

— M_i is the expected number of marked units in the population just before PSP i such that $M_i = (M_{i-1} + U_{i-1}p_{i-1}^*)\phi_{i-1}$; $M_1 = 0$;

— $\eta_i = 1 - M_i/N_i$ is the proportion of unmarked units just before PSP i ; $\bar{\eta}_i = 1 - \eta_i$ is the proportion of marked units just before PSP i ;

— $D_i = 1 - (1-p_i^*)\chi_i\eta_i - \bar{\eta}_i\bar{\chi}_i$ is the probability, for an unmarked unit, to be captured at least once at PSP i or later and, for a marked unit, to be captured for the last time at PSP i.

For the asymptotic variance derivations, N_i is assumed large for every PSP i.

A.1.1 Evaluation of var $\left(\hat{N}_i\right)$ for model M_0

This section presents the main steps to evaluate the asymptotic variance of $\hat{N}_i$. The following notation is used throughout this section.

— $P_{2i} = 1-(1-p_i)^{\ell_i} - \ell_i p_i (1-p_i)^{\ell_i - 1}$ is the probability of obtaining two captures or more at PSP i ,

— $p_{1i} = \ell_i p_i (1-p_i)^{\ell_i - 1}$ is the probability of being captured once at PSP i,

and note that $p_i^* = P_{2i} + p_{1i}$. In this section, we give further details to the proof of the asymptotic variance of $\hat{N}_i$,

$$\text{var}\left(\hat{N}_i\right) = N_i D_i (1 - p_i^*)/(p_i^* \bar{\chi}_i \bar{\eta}_i + D_i P_{2i}). \tag{A.1}$$

In the derivations, we approximate the sufficient statistics u_i, v_i, and w_i by their expectations,

$$u_i \approx N_i \eta_i p_i^*, \quad v_i \approx N_i p_i^* \chi_i, \text{ and } w_i \approx N_i \bar{\eta}_i \bar{\chi}_i + N_i p_i^* \bar{\eta}_i \chi_i, \tag{A.2}$$

as, when N_i goes to ∞, the ratio of a statistic over its expectation converges to 1 in probability. Throughout the derivations, we will refer to the estimating equation that leads to the maximum likelihood estimator for N_i,

$$f_{i,N}(N_i) = N_i \left[n_i^*(w_i - n_i^* + u_i) / \left\{ n_i^*(w_i - v_i) + u_i v_i \right\} - \left\{ 1 - C_i/(\ell_i N_i) \right\}^{\ell_i} \right], \tag{A.3}$$

where

$$n_i^* = N_i [1 - \left\{ 1 - C_i/(N_i \ell_i) \right\}^{\ell_i}]. \tag{A.4}$$

The limit in probability of the derivative of $f_{i,N}(N_i)$ with respect to N_i is proven in the next section to be

$$A_i = -(p_i^* \bar{\chi}_i \bar{\eta}_i + D_i \times P_{2i})/(\bar{\eta}_i \bar{\chi}_i). \tag{A.5}$$

Evaluation of A_i

Using (A.4), the derivative of n_i^* with respect to N_i, evaluated at the expectation, $N_i p_i \ell_i$, of C_i, is equal to P_{2i}. Now, differentiating (A.3) using the chain rule and applying (A.2) gives

$$\begin{aligned}
\frac{\partial f_{i,N}}{\partial N_i} &= N_i \frac{\partial n_i^*}{\partial N_i} \frac{\partial}{\partial n_i^*} \left[n_i^*(w_i - n_i^* + u_i) / \left\{ n_i^*(w_i - v_i) + u_i v_i \right\} \right] - N_i \frac{\partial}{\partial N_i} \{1 - C_i/(N_i \ell_i)\}^{\ell_i} \\
&= -N_i \frac{\partial n_i^*}{\partial N_i} \left[\left\{ (n_i^*)^2 (w_i - v_i) + (2n_i^* - w_i - u_i) u_i v_i \right\} / \left\{ n_i^*(w_i - v_i) + u_i v_i \right\}^2 \right] \\
&- (C_i/N_i) \left\{ 1 - C_i/(N_i \ell_i) \right\}^{\ell_i - 1} \\
&\approx -P_{2i} \left\{ (1 - \eta_i \chi_i + p_i^* \eta_i \chi_i)/(\bar{\eta}_i \bar{\chi}_i) \right\} - p_{1i} = A_i,
\end{aligned}$$

where the last equality is obtained by setting $p_{1i} = p_i^* - P_{2i}$ and A_i is defined by (A.5).

Evaluation of Σ_i

The first three random variables (u_i, v_i, w_i) are, under Poisson sampling, Poisson random variables. Their variances are equal to their expectations given in (A.2) while the covariances are the expected number of units common to two random variables. Now C_i does not have a Poisson distribution. To find the elements of Σ_i for C_i, let $\tilde{N}_i$ be a Poisson random variable equal to the actual number of units in the population during PSP i. Given $\tilde{N}_i$, C_i has a binomial distribution with $\tilde{N}_i \times \ell_i$ trials and probability p_i. Conditioning on $\tilde{N}_i$ yields to the following expression for the variance of C_i,

$$\text{var}(C_i) = E(\tilde{N}_i) \ell_i p_i (1 - p_i) + \text{var}(\tilde{N}_i) \ell_i^2 p_i^2 = N_i \ell_i p_i (\ell_i p_i + 1 - p_i).$$

The covariances with C_i are calculated in a similar way. For instance that with u_i involves $\tilde{U}_i$, the units that are not marked before the ith PSP. Given $(\tilde{N}_i, \tilde{U}_i)$ the conditional covariance between C_i and u_i is $\tilde{U}_i \ell_i p_i (1 - p_i^*)$ while their respective expectations are $\tilde{U}_i \ell_i p_i$ and $\tilde{U}_i p_i^*$. Thus

$$\mathrm{cov}(C_i, u_i) = E\{\tilde{U}_i \ell_i p_i (1 - p_i^*)\} + \mathrm{cov}(\tilde{U}_i \ell_i p_i, \tilde{U}_i p_i^*) = U_i \ell_i p_i = N_i \eta_i \ell_i p_i.$$

Similar derivations lead to the covariance matrix of the sufficient statistics u_i, v_i, w_i, C_i,

$$\Sigma_i = N_i \begin{pmatrix} \eta_i p_i^* & \eta_i p_i^* \chi_i & 0 & \eta_i \ell_i p_i \\ & p_i^* \chi_i & \bar{\eta}_i p_i^* \chi_i & \ell_i p_i \chi_i \\ & & \bar{\eta}_i (\bar{\chi}_i + p_i^* \chi_i) & \bar{\eta}_i \ell_i p_i \\ & & & \ell_i p_i (1 - p_i + \ell_i p_i) \end{pmatrix}. \tag{A.6}$$

Derivation of $\nabla f_{i,N}$

We have,

$$\nabla f_{i,N} = \left(\frac{\partial f_{i,N}}{\partial u_i}, \frac{\partial f_{i,N}}{\partial v_i}, \frac{\partial f_{i,N}}{\partial w_i}, \frac{\partial f_{i,N}}{\partial C_i} \right).$$

Now, differentiating $f_{i,N}$ with respect to u_i and using (A.2) gives

$$\begin{aligned} \frac{\partial f_{i,N}}{\partial u_i} &= N_i n_i^* w_i (n_i - v_i) / \{n_i^* (w_i - v_i) + u_i v_i\}^2 \\ &\approx (\bar{\chi}_i + p_i^* \chi_i) / (\bar{\eta}_i \bar{\chi}_i). \end{aligned}$$

All of the other derivatives are calculated in a similar way.

Evaluation of $\nabla f_{i,N}^\top \Sigma_i \nabla f_{i,N}$

The denominator $(\bar{\eta}_i \bar{\chi}_i)$ of $\nabla f_{i,N}$ simplifies with that of A_i ; only the numerator is considered here. To evaluate it we calculate separately the quadratic form Q_1 involving the covariance matrix of (u_i, v_i, w_i) and Q_2, involving the variance and the covariances with C_i. As $D_i = -2\chi_i \eta_i + \chi_i + \eta_i + p_i^* \chi_i \eta_i$, patient calculations lead to

$$\begin{aligned} Q_1 &= p_i^* \{\eta_i + \chi_i - 2\eta_i \chi_i - \eta_i^2 \chi_i - \eta_i \chi_i^2 + 2\eta_i^2 \chi_i^2 \\ &\quad + p_i^* (1 - \eta_i - \chi_i + \eta_i \chi_i + 2\eta_i^2 \chi_i + 2\eta_i \chi_i^2 - 4\eta_i^2 \chi_i^2) \\ &\quad + (p_i^*)^2 (\eta_i \chi_i - \eta_i^2 \chi_i - \eta_i \chi_i^2 + 2\eta_i^2 \chi_i^2)\} \\ &= p_i^* \{D_i^2 + \eta_i + \chi_i - 4\eta_i \chi_i - \eta_i^2 - \chi_i^2 + 3\eta_i \chi_i^2 + 3\eta_i^2 \chi_i - 2\eta_i^2 \chi_i^2 \\ &\quad + p_i^* (1 - \eta_i)(1 - \chi_i) + (p_i^*)^2 \eta_i \chi_i (1 - \eta_i)(1 - \chi_i)\}. \end{aligned}$$

Since $\eta_i + \chi_i - 4\eta_i \chi_i - \eta_i^2 - \chi_i^2 + 3\eta_i^2 \chi_i + 3\eta_i \chi_i^2 - 2\eta_i^2 \chi_i^2 = (1 - \eta_i)(1 - \chi_i)(\eta_i + \chi_i - 2\eta_i \chi_i)$, we have

$$Q_1 = p_i^* D_i^2 + p_i^* \bar{\eta}_i \bar{\chi}_i \{\chi_i + \eta_i - 2\chi_i \eta_i + p_i^* + (p_i^*)^2 \eta_i \chi_i\}.$$

To evaluate Q_2 note that $p_{1i}(1-p_i+\ell_i p_i)=\ell_i p_i(1-P_{2i})$. Thus

$$\begin{aligned} Q_2 &= -2p_{1i}D_i\{\eta_i(1-\chi_i+\chi_i p_i^*)+\chi_i(1-\eta_i)(1-p_i^*)+p_i^*(1-\eta_i)\} \\ &+ D_i^2 p_{1i}^2(1-p_i+\ell_i p_i)/(\ell_i p_i) \\ &= -2p_{1i}D_i\{p_i^*(1-\eta_i)(1-\chi_i)+D_i\}+D_i^2 p_{1i}(1-P_{2i}) \\ &= -2p_{1i}D_i p_i^*\bar{\eta}_i\bar{\chi}_i - D_i^2 p_{1i}(1+P_{2i}). \end{aligned}$$

Since $p_i^*-p_{1i}-p_{1i}P_{2i}=P_{2i}(1-p_{1i})$, the quadratic form is

$$Q_1+Q_2 = p_i^*\bar{\eta}_i\bar{\chi}_i\{\chi_i+\eta_i-2\chi_i\eta_i+p_i^*+(p_i^*)^2\eta_i\chi_i-2p_{1i}D_i\}+D_i^2P_{2i}(1-p_{1i}).$$

As $D_i^2P_{2i}(1-p_{1i}-P_{2i})=D_i^2P_{2i}(1-p_i^*)$, subtracting (A.5) squared, that is $\{p_i^*(1-\eta_i)(1-\chi_i)\}^2+2D_iP_{2i}p_i^*(1-\eta_i)(1-\chi_i)+D_i^2P_{2i}^2$, leads to

$$Q_3 = -\{p_i^*\bar{\eta}_i\bar{\chi}_i\}^2+D_i^2P_{2i}(1-p_i^*)+p_i^*\bar{\eta}_i\bar{\chi}_i\times Q_4,$$

where

$$\begin{aligned} Q_4 &= \eta_i+\chi_i-2\eta_i\chi_i+p_i^*+(p_i^*)^2\eta_i\chi_i-2(p_{1i}+P_{2i})D_i \\ &= \eta_i+\chi_i-2\eta_i\chi_i+p_i^*+(p_i^*)^2\eta_i\chi_i+2p_i*(2\eta_i\chi_i-\eta_i-\chi_i-p_i^*\eta_i\chi_i) \\ &= p_i^*\bar{\eta}_i\bar{\chi}_i+(1-p_i^*)D_i. \end{aligned}$$

Thus $Q_3=(1-p_i^*)D_i\{D_iP_{2i}+p_i^*\bar{\eta}_i\bar{\chi}_i\}$. It gives (A.1) when divided by the numerator of A_i^2 and multiplied by N_i ; A_i is defined by (A.5).

A.1.2 Derivation of var $\left(\hat{\phi}_i\right)$ for model M_0

In this section, we show the calculations that lead to the variance of $\hat{\phi}_i$,

$$\begin{aligned} \mathrm{var}(\hat{\phi}_i) &= \phi_i^2\left\{\frac{(1-p_{i+1}^*)\chi_{i+1}\left\{\bar{\eta}_{i+1}+p_{i+1}^*\eta_{i+1}\right\}}{N_{i+1}p_{i+1}^*\bar{\eta}_{i+1}\bar{\chi}_{i+1}}+\frac{(1-p_i^*)\bar{\eta}_i\chi_i}{N_ip_i^*\bar{\chi}_i(\bar{\eta}_i+\eta_ip_i^*)}+\frac{1-\phi_i}{N_{i+1}\bar{\eta}_{i+1}}\right\} \quad \text{(A.7)} \\ &- \phi_i^2\left\{\frac{(1-p_{i+1}^*)(\bar{\eta}_{i+1}+p_{i+1}^*\eta_{i+1})^2\chi_{i+1}^2P_{2,i+1}}{N_{i+1}\bar{\eta}_{i+1}\bar{\chi}_{i+1}p_{i+1}^*(\bar{\eta}_{i+1}\bar{\chi}_{i+1}p_{i+1}^*+D_{i+1}P_{2,i+1})}+\frac{(1-p_i^*)\bar{\eta}_i\chi_i^2P_{2,i}}{N_i\bar{\chi}_ip_i^*(\bar{\eta}_i\bar{\chi}_ip_i^*+D_iP_{2,i})}\right\}. \end{aligned}$$

It is useful to define the conditional expectation of the units that are marked before session i and available for capture at session i or later,

$$\hat{M}_i = n_i^*-u_i+n_i^*(w_i-n_i^*+u_i)/(n_i^*-v_i). \quad \text{(A.8)}$$

There are two intermediate steps, the evaluation of the variance of $\hat{n}_i$ and of $\hat{M}_i$,

$$\mathrm{var}\left(\hat{M}_i\right) = N_i(1-p_i^*)\bar{\eta}_i\chi_i(\bar{\eta}_i+p_i^*\eta_i)(p_i^*\bar{\eta}_i+P_{2i}\eta_i)/\left\{p_i^*(p_i^*\bar{\chi}_i\bar{\eta}_i+D_i\times P_{2i})\right\}. \quad \text{(A.9)}$$

Derivation of var $(\hat{n}_i)$

Since $\hat{n}_i = \hat{N}_i \left\{1 - \left(1 - C_i/(\ell_i \hat{N}_i)\right)^{\ell_i}\right\}$, a standard linearization argument leads to the following approximation, $\hat{n}_i \approx N_i p_i^* + P_{2i}(\hat{N}_i - N_i) + (1-p_i)^{\ell_i - 1}(C_i - N_i \ell_i p_i)$. The covariance between C_i and $\hat{N}_i$ is evaluated as N_i times the fourth entry of the vector $-\Sigma_i \nabla f_{i,N}/A_i$. It is found to be equal to $N_i \ell_i p_i$. Thus the linearization approximation to the variance of $\hat{n}_i$ is

$$\begin{aligned}
\text{var}\,(\hat{n}_i) &= P_{2i}^2\{\text{var}(\hat{N}_i) + N_i\} + 2P_{2i}(1-p_i)^{\ell_i - 1} N_i \ell_i p_i \\
&\quad + (1-p_i)^{2\ell_i - 2} N_i \ell_i p_i (1 - p_i + \ell_i p_i) \\
&= P_{2i}^2 \text{var}(\hat{N}_i) + N_i\{(p_i^*)^2 + (p_i^* - P_{2i})(1 - p_i^*)\} \\
&= N_i \left[P_{2i}(1-p_i^*)\left\{P_{2i} D_i/(p_i^* \bar{\eta}_i \bar{\chi}_i + P_{2i} D_i) - 1\right\} + p_i^*\right] \\
&= N_i p_i^* \left\{1 - (1-p_i^*)\bar{\eta}_i \bar{\chi}_i P_{2i}/(p_i^* \bar{\eta}_i \bar{\chi}_i + D_i \times P_{2i})\right\}.
\end{aligned}$$

Thus the variance of $\hat{n}_i$ is equal to $N_i p_i^*$, the variance under an open population model, minus a term representing the variance reduction under a robust design.

Derivation of var $\left(\hat{M}_i\right)$

From equation (A.8),

$$\hat{M}_i = \hat{n}_i - u_i + \hat{n}_i(w_i - \hat{n}_i + u_i)/(\hat{n}_i - v_i).$$

The linearization approximation to $\hat{M}_i$ is

$$\hat{M}_i = M_i + \begin{pmatrix} u_i - N_i \eta_i p_i^* \\ v_i - N_i p_i^* \chi_i \\ w_i - N_i \bar{\eta}_i \{1 - (1-p_i^*)\chi_i\} \\ \hat{n}_i - N_i p_i^* \end{pmatrix}^\top \nabla M_i$$

and ∇M_i is limit of the vector of partial derivatives of $\hat{M}_i$ with respect to $(u_i, v_i, w_i, \hat{n}_i)$. Using (A.2), it is given by

$$\nabla M_i = (1/\bar{\chi}_i) \begin{pmatrix} \chi_i \\ \bar{\eta}_i (1 - p_i^*)/p_i^* \\ 1 \\ -\chi_i(\bar{\eta}_i + \eta_i p_i^*)/p_i^* \end{pmatrix}.$$

Following Jolly (1965), as $M_i = N_i \bar{\eta}_i$, the quadratic form for the variance is $\text{var}\left(\hat{M}_i\right) = \nabla M_i^\top \Sigma_i^* \nabla M_i - N_i \bar{\eta}_i$, where Σ_i^* is the covariance matrix of $u_i, v_i, w_i, \hat{n}_i$. The covariance matrix for u_i, v_i, w_i is already discussed in sections (A.1.1) and (A.1.1). Using the linearization for $\hat{n}_i$, one has

$$\text{cov}(\hat{n}_i, u_i) = P_{2i}\text{cov}(\hat{N}_i, u_i) + (1-p_i)^{\ell_i - 1}\text{cov}(C_i, u_i).$$

The covariance between u_i and $\hat{N}_i$ is evaluated as N_i times the first entry of the vector $-\Sigma_i \nabla f_{i,N}/A_i$. It is equal to $N_i \eta_i p_i^*$. The covariance between C_i and u_i is $N_i \eta_i \ell_i p_i$, see (A.6). This gives $\text{cov}(\hat{n}_i, u_i) = N_i \eta_i p_i^*$. In the same fashion, we derive $\text{cov}(\hat{n}_i, v_i) = N_i p_i^* \chi_i$ and $\text{cov}(\hat{n}_i, w_i) = N_i \bar{\eta}_i p_i^*$. Thus, the covariance matrix between $u_i, v_i, w_i, \hat{n}_i$ is,

$$\Sigma_i^* = N_i \begin{pmatrix} \eta_i p_i^* & \eta_i p_i^* \chi_i & 0 & \eta_i p_i^* \\ & p_i^* \chi_i & \bar{\eta}_i p_i^* \chi_i & p_i^* \chi_i \\ & & \bar{\eta}_i \{1 - (1 - p_i^*)\chi_i\} & \bar{\eta}_i p_i^* \\ & & & \text{var}\,(\hat{n}_i) \end{pmatrix}.$$

To evaluate the quadratic form $\nabla M_i^\top \Sigma_i^* \nabla M_i$, first observe that the contribution of the covariance matrix of (u_i, v_i, w_i) is

$$Q_1 = (N_i/\bar{\chi}_i^2) \left[\eta_i p_i^* \chi_i^2 + \bar{\eta}_i + (1 - p_i^*) \bar{\eta}_i \chi_i \left\{ \bar{\eta}_i / p_i^* - \eta_i + 2\eta_i \chi_i \right\} \right].$$

The sum of the the three terms involving a covariance with $\hat{n}_i$ is

$$Q_2 = \left\{ -2N_i p_i^* \chi_i (\bar{\eta}_i + \eta_i p_i^*)/(p_i^* \bar{\chi}_i) \right\} \left[\bar{\eta}_i + \chi_i (\bar{\eta}_i + \eta_i p_i^*)/(p_i^* \bar{\chi}_i) \right].$$

The last term, involving the variance of $\hat{n}_i$, is given by

$$\begin{aligned} Q_3 &= N_i p_i^* \left\{ (\bar{\eta}_i + p_i^* \eta_i) \chi_i / p_i^* \bar{\chi}_i \right\}^2 \left\{ 1 - (1 - p_i^*) \bar{\eta}_i \bar{\chi}_i P_{2i} / (p_i^* \bar{\chi}_i \bar{\eta}_i + D_i \times P_{2i}) \right\} \\ &= \underbrace{N_i p_i^* \left\{ (\bar{\eta}_i + p_i^* \eta_i) \chi_i / (p_i^* \bar{\chi}_i) \right\}^2}_{Q_{3,1}} \\ &- \underbrace{N_i p_i^* \left\{ (\bar{\eta}_i + p_i^* \eta_i) \chi_i / (p_i^* \bar{\chi}_i) \right\}^2 \left\{ (1 - p_i^*) \bar{\eta}_i \bar{\chi}_i P_{2i} / (p_i^* \bar{\chi}_i \bar{\eta}_i + D_i \times P_{2i}) \right\}}_{Q_{3,2}}. \end{aligned}$$

First we calculate the variance under an open population model, corresponding to the case $Q_{3,2} = P_{2i} = 0$. This gives

$$Q_1 + Q_2 + Q_{3,1} = N_i \chi_i \bar{\eta}_i (1 - p_i^*) \left\{ \bar{\eta}_i + p_i^* \eta_i \right\} / (p_i^* \bar{\chi}_i) + N_i \bar{\eta}_i.$$

Now adding $Q_{3,2}$ yields the expression for $\nabla M_i^\top \Sigma_i^* \nabla M_i - N_i \bar{\eta}_i$ given in (A.9) .

In the next sections, we give some details on the last calculations that lead to the variance of $\hat{\phi}_i$.

Derivation of $\nabla \hat{\phi}_i$

We have,

$$\nabla \hat{\phi}_i = \left(\frac{\partial \hat{\phi}_i}{\partial \hat{M}_{i+1}}, \frac{\partial \hat{\phi}_i}{\partial \hat{M}_i}, \frac{\partial \hat{\phi}_i}{\partial u_i} \right).$$

Differentiating $\hat{\phi}_i$ with respect to $\hat{M}_{i+1}$ and using (A.2) leads to

$$\begin{aligned} \frac{\partial \hat{\phi}_i}{\partial \hat{M}_{i+1}} &= 1/(\hat{M}_i + u_i) \\ &\approx 1/\left\{ N_i (\bar{\eta}_i + \eta_i p_i^*) \right\}. \end{aligned}$$

The derivations with respect to $\hat{M}_i$ and u_i are calculated in a similar way.

Evaluation of Γ_i

The variance of $\hat{M}_{i+1}$ and $\hat{M}_i$ have been derived in section (A.1.2) ; the expectation of u_i is calculated in equation (A.2). The covariance between $\hat{M}_{i+1}$ and $\hat{M}_i$ is evaluated as

$$\text{cov}(\hat{M}_{i+1}, \hat{M}_i) = \nabla M_{i+1}^{\top} \Gamma_i^* \nabla M_i,$$

where Γ_i^* is the 4×4 matrix of the covariances between the vectors $(u_i, v_i, w_i, \hat{n}_i)$ and $(u_{i+1}, v_{i+1}, w_{i+1}, \hat{n}_{i+1})$. The covariance between u_i and u_{i+1} is 0. The covariance between u_i and v_{i+1} is $N_i \eta_i p_i^* \phi_i p_{i+1}^* \chi_{i+1}$. The other covariances involving the random variables u_i, v_i, w_i and $u_{i+1}, v_{i+1} w_{i+1}$ are derived in a similar way. The covariances involving $\hat{n}_i$ and $\hat{n}_{i+1}$ are derived using the approach discussed in section (A.1.2). For example, the covariance between $\hat{n}_i$ and $\hat{n}_{i+1}$ gives $N_i p_i^* \phi_i p_{i+1}^*$. Finally, the covariance matrix Γ_i^* is,

$$\Gamma_i^* = N_i \begin{pmatrix} 0 & \eta_i p_i^* \phi_i p_{i+1}^* \chi_{i+1} & \eta_i p_i^* \phi_i (\bar{\chi}_{i+1} + p_{i+1}^* \chi_{i+1}) & \eta_i p_i^* \phi_i p_{i+1}^* \\ 0 & 0 & 0 & 0 \\ 0 & \bar{\eta}_i \phi_i p_{i+1}^* \chi_{i+1} & \bar{\eta}_i (1 - p_i^*) \phi_i (1 - p_{i+1}^*) \bar{\chi}_{i+1} & \bar{\eta}_i (1 - p_i^*) \phi_i p_{i+1}^* \\ 0 & p_i^* \phi_i p_{i+1}^* \chi_{i+1} & p_i^* \phi_i (\bar{\chi}_{i+1} + p_{i+1}^* \chi_{i+1}) & p_i^* \phi_i p_{i+1}^* \end{pmatrix}^{\top}.$$

To evaluate the quadratic form $\nabla M_{i+1}^{\top} \Gamma_i^* \nabla M_i$, we first calculate the component featuring the covariances between (u_i, v_i, w_i) and $(u_{i+1}, v_{i+1}, w_{i+1})$,

$$\begin{aligned} Q_1 &= \{N_i \phi_i / (\bar{\chi}_i \bar{\chi}_{i+1})\} \left[\eta_i \chi_i \left\{\bar{\eta}_{i+1}(1 - p_{i+1}^*)\chi_{i+1} + p_i^*(\bar{\chi}_{i+1} + p_{i+1}^* \chi_{i+1})\right\}\right] \\ &+ \{N_i \phi_i / (\bar{\chi}_i \bar{\chi}_{i+1})\} \left[\bar{\eta}_i (1 - p_{i+1}^*) \left\{\bar{\eta}_{i+1} \chi_{i+1} + (1 - p_i^*) \bar{\chi}_{i+1}\right\}\right]. \end{aligned}$$

The sum of the covariances between (u_i, v_i, w_i) and $\hat{n}_{i+1}$ is

$$Q_2 = -N_i \chi_i \chi_{i+1} \phi_i (\bar{\chi}_{i+1} + \eta_{i+1} p_{i+1}^*)(\bar{\chi}_i + \eta_i p_i^*) / \bar{\chi}_i.$$

That involving $(u_{i+1}, v_{i+1}, w_{i+1})$ and $\hat{n}_i$ is,

$$Q_3 = -N_i p_i^* \phi_i \chi_i (\bar{\eta}_i + \eta_i p_i^*) \left\{\bar{\eta}_{i+1}(1 - p_{i+1}^*) + \bar{\chi}_{i+1} + p_{i+1}^* \chi_{i+1}\right\} / (p_i^* \bar{\chi}_i \bar{\chi}_{i+1}).$$

The component involving the covariance between $\hat{n}_{i+1}$ and $\hat{n}_i$ is

$$Q_4 = N_i \phi_i \chi_i \chi_{i+1} (\bar{\eta}_i + \eta_i p_i^*)(\bar{\eta}_{i+1} + \eta_{i+1} p_{i+1}^*) / (\bar{\chi}_i \bar{\chi}_{i+1}).$$

Now adding up Q_1, Q_2, Q_2 and Q_4 gives the quadratic form $\nabla M_{i+1}^{\top} \Gamma_i^* \nabla M_i$. It leads to $\text{cov}(\hat{M}_{i+1}, \hat{M}_i) = N_i \bar{\eta}_i \phi_i$. In the same fashion, we derive $\text{cov}(\hat{M}_{i+1}, u_i) = N_i \eta_i p_i^* \phi_i$ and $\text{cov}(\hat{M}_i, u_i) = 0$. The covariance matrix Γ_i is now completely derived.

Evaluation of the variance of $\hat{\phi}_i$

To evaluate the quadratic form $\nabla\hat{\phi}_i^\top \Gamma_i \nabla\hat{\phi}_i$, first observe that the contribution of the three covariances between $\hat{M}_{i+1}$,$\hat{M}_i$, and u_i is

$$Q_1 = -2\phi_i^2 / \left\{N_i(\bar{\eta}_i + \eta_i p_i^*)\right\},$$

while the contribution of the the variances of $(\hat{M}_{i+1}, \hat{M}_i, u_i)$ to the quadratic form is

$$\begin{aligned} Q_2 &= \phi_i^2 \frac{(1-p_{i+1}^*)\chi_{i+1}\left\{\bar{\eta}_{i+1} + \eta_{i+1}p_{i+1}^*\right\}\left\{p_{i+1}^*\bar{\eta}_{i+1} + P_{2,i+1}\eta_{i+1}\right\}}{N_{i+1}p_{i+1}^*\bar{\eta}_{i+1}\left\{p_{i+1}^*\bar{\chi}_{i+1}\bar{\eta}_{i+1} + D_{i+1}\times P_{2,i+1}\right\}} \\ &+ \phi_i^2\left[\frac{(1-p_i^*)\bar{\eta}_i\chi_i\left\{p_i^*\bar{\eta}_i + P_{2i}\eta_i\right\}}{N_i p_i^*\left\{\bar{\eta}_i + \eta_i p_i^*\right\}\left\{p_i^*\bar{\chi}_i\bar{\eta}_i + D_i\times P_{2i}\right\}}\right] \\ &+ \phi_i^2\left[\frac{1}{N_{i+1}\bar{\eta}_{i+1}} + \frac{1}{N_i(\bar{\eta}_i + \eta_i p_i^*)}\right]. \end{aligned}$$

Finally, adding up Q_1 and Q_2 gives (A.7) .

A.1.3 Derivation of var $\left(\hat{N}_i\right)$ for model M_t

In this section, we show the calculations that lead to the variance of $\hat{N}_i$. The following notation is used throughout this section.

— $p_{1i}^* = \sum_{j=1}^{\ell_i}\left\{p_{ij}\prod_{\substack{s=1\\s\neq j}}^{\ell_i}(1-p_{is})\right\}$ the probability of being captured once at PSP i,

— $P_{2i}^* = 1 - \prod_{j=1}^{\ell_i}(1-p_{ij}) - \sum_{j=1}^{\ell_i}\left\{p_{ij}\prod_{\substack{s=1\\s\neq j}}^{\ell_i}(1-p_{is})\right\}$ the probability of obtaining two captures or more at PSP i.

Furthermore, for a PSP i, let p_{ij} and n_{ij} be (resp.) the probability of being captured and the number of captures in a secondary sampling level j , $j = 1, 2, ..., \ell_i$; thus, $p_i^* = 1 - \prod_{j=1}^{\ell_i}(1-p_{ij})$.

The estimating equation that leads to the maximum likelihood estimator for N_i is

$$f_{i,N}^*(N_i) = N_i\left[n_i^*(w_i - n_i^* + u_i)/\left\{n_i^*(w_i - v_i) + u_i v_i\right\} - \prod_{j=1}^{\ell_i}(1 - n_{ij}/N_i)\right], \tag{A.10}$$

where

$$n_i^* = N_i\left\{1 - \prod_{j=1}^{\ell_i}(1 - n_{ij}/N_i)\right\}. \tag{A.11}$$

This section give further details to the proof of the asymptotic variance of $\hat{N}_i$,

$$\text{var}\left(\hat{N}_i\right) = N_i D_i(1-p_i^*)/(p_i^*\bar{\chi}_i\bar{\eta}_i + D_i P_{2i}^*). \tag{A.12}$$

Evaluation of A_i^*

We have,

$$\begin{aligned} n_i^* &= N_i \left\{ 1 - \prod_{j=1}^{\ell_i} (1 - n_{ij}/N_i) \right\} \\ &= N_i \left[1 - \exp \left\{ \sum_{j=1}^{\ell_i} \log (1 - n_{ij}/N_i) \right\} \right]. \end{aligned}$$

Then, its derivative with respect to N_i, evaluated at the expectation of n_{ij}, $j = 1, 2, ..., \ell_i$, is equal to P_{2i}^*, the probability of being captured more than once during PSP i. Now, differentiating (A.10) using the chain rule and applying (A.2) gives

$$\begin{aligned} \frac{\partial f_{i,N}^*}{\partial N_i} &\approx -N_i P_{2i}^* \left[(1 - p_i^*)\chi_i / (N_i \bar{\chi}_i) + \{1 - (1 - p_i^*)\chi_i\} / (M_i \bar{\chi}_i) \right] - p_{1i}^* \\ &= -\left\{ 1/(\bar{\chi}_i \bar{\eta}_i) \right\} (p_i^* \bar{\chi}_i \bar{\eta}_i + D_i \times P_{2i}^*) = A_i^*. \end{aligned}$$

Evaluation of Σ_i^*

The variance and the covariance of the three Poisson random variables u_i, v_i, w_i are already discussed in section (A.1.1). Now, for $j = 1, 2, ..., \ell_i$, the variance of n_{ij}, a Poisson random variable, is

$$\text{var}(n_{ij}) = N_i p_{ij}.$$

The covariances with n_{ij} are calculated in a similar way. For instance that with u_i involves $\tilde{U}_i$, the units that are not marked before the ith PSP. Given $(\tilde{N}_i, \tilde{U}_i)$ the conditional covariance between n_{ij} and u_i is $\tilde{U}_i \ell_i p_{ij}(1 - p_i^*)$ while their respective expectations are $\tilde{N}_i p_{ij}$ and $\tilde{U}_i p_i^*$. Thus,

$$\text{cov}(n_{ij}, u_i) = E\{\tilde{U}_i p_{ij}(1 - p_i^*)\} + \text{cov}(\tilde{N}_i p_{ij}, \tilde{U}_i p_i^*) = U_i p_{ij} = N_i \eta_i p_{ij}.$$

Finally, the covariance matrix of $u_i, v_i, w_i, n_{i1}, n_{i2}, ..., n_{i\ell_i}$, $\mathbf{\Sigma_i^*}$, is

$$\mathbf{\Sigma_i^*} = N_i \begin{pmatrix} \eta_i p_i^* & \eta_i p_i^* \chi_i & 0 & \eta_i p_{i1} & \dots & \eta_i p_{i\ell_i} \\ & p_i^* \chi_i & \bar{\eta}_i p_i^* \chi_i & p_{i1}\chi_i & \dots & p_{i\ell_i}\chi_i \\ & & \bar{\eta}_i \{1 - (1 - p_i^*)\chi_i\} & \bar{\eta}_i p_{i1} & \dots & \bar{\eta}_i p_{i\ell_i} \\ & & & p_{i1} & \dots & p_{i1}p_{i\ell_i} \\ & & & & \ddots & p_{i\ell_i} \end{pmatrix}. \tag{A.13}$$

Derivation of $\nabla f_{i,N}^*$

We have,

$$\nabla f_{i,N}^* = \left(\frac{\partial f_{i,N}}{\partial u_i}, \frac{\partial f_{i,N}}{\partial v_i}, \frac{\partial f_{i,N}}{\partial w_i}, \left(\frac{\partial f_{i,N}}{\partial n_{ij}} \right)_{j=1,\dots,\ell_i} \right).$$

The first three partial derivates are the same as calculated in section (A.1.1). Now, for a fixed j $(j = 1, 2, ..., \ell_i)$, deriving $f^*_{i,N}$ with respect to n_{ij} using (A.2) gives

$$\begin{aligned}
\frac{\partial f_{i,N}}{\partial n_{ij}} &\approx \prod_{\substack{s=1\\s\neq j}}^{\ell_i}(1-p_{is}) - \prod_{\substack{s=1\\s\neq j}}^{\ell_i}(1-p_{is})\left\{(\bar{\chi}_i + p^*_i\chi_i)/(\bar{\eta}_i\bar{\chi}_i) + (1-p^*_i)\chi_i/(\bar{\chi}_i)\right\} \\
&= \prod_{\substack{s=1\\s\neq j}}^{\ell_i}(1-p_{is})\left\{1-(\bar{\chi}_i + p^*_i\chi_i)/(\bar{\eta}_i\bar{\chi}_i) - (1-p^*_i)\chi_i/(\bar{\chi}_i)\right\} \\
&= -D_i\prod_{\substack{s=1\\s\neq j}}^{\ell_i}(1-p_{is})/(\bar{\eta}_i\bar{\chi}_i).
\end{aligned}$$

The limit of the vector of partial derivatives of $f^*_{i,N}(N_i)$ with respect to u_i, v_i, w_i, n_{i1},n_{i2},...,$n_{i\ell_i}$, is

$$\nabla f^*_{i,N} = \{1/(\bar{\chi}_i\bar{\eta}_i)\} \times (1-(1-p^*_i)\chi_i, \bar{\eta}_i(1-p^*_i), p^*_i, -D_i\mathbf{\Psi_i})^\top,$$

where

$$\mathbf{\Psi_i} = \left(\prod_{\substack{s=1\\s\neq 1}}^{\ell_i}(1-p_{is}), \prod_{\substack{s=1\\s\neq 2}}^{\ell_i}(1-p_{is}), ..., \prod_{\substack{s=1\\s\neq \ell_i}}^{\ell_i}(1-p_{is})\right).$$

Evaluation of $\nabla f^{*\top}_{i,N}\Sigma^*_i\nabla f^*_{i,N}$

From the previous evaluation of $\nabla f^\top_{i,N}\Sigma_i\nabla f_{i,N}$, the only quantity that changes is the one involving the variances and the covariances with n_{ij}. Let that be A_{ij}. We have,

$$\begin{aligned}
A_{ij} &= \sum_{j=1}^{\ell_i}\left\{D_i\prod_{\substack{s=1\\s\neq j}}^{\ell_i}(1-p_{is})\right\}^2 p_{ij} + \sum_{j=1}^{\ell_i}\sum_{\substack{j'=1\\j'\neq j}}^{\ell_i}\left\{D_i^2\prod_{\substack{s=1\\s\neq j}}^{\ell_i}(1-p_{is})\prod_{\substack{t=1\\t\neq j'}}^{\ell_i}(1-p_{it})\,p_{ij}p_{ij'}\right\} \\
&- 2D_i\sum_{j=1}^{\ell_i}\prod_{\substack{s=1\\s\neq j}}^{\ell_i}(1-p_{is})\left\{\frac{\partial f^*_{i,N}}{\partial u_i}\eta_i p_{ij} + \frac{\partial f^*_{i,N}}{\partial v_i}p_{ij}\chi_i + \frac{\partial f^*_{i,N}}{\partial w_i}\bar{\eta}_i p_{ij}\right\} \\
&= D_i^2\sum_{j=1}^{\ell_i}p_{ij}\prod_{\substack{s=1\\s\neq j}}^{\ell_i}(1-p_{is})\left[\prod_{\substack{s=1\\s\neq j}}^{\ell_i}(1-p_{is}) - p_{ij}\prod_{\substack{s=1\\s\neq j}}^{\ell_i}(1-p_{is}) + \sum_{j=1}^{\ell_i}p_{ij}\prod_{\substack{s=1\\s\neq j}}^{\ell_i}(1-p_{is})\right] \\
&- 2D_ip^*_{1i}\left\{\frac{\partial f^*_{i,N}}{\partial u_i}\eta_i + \frac{\partial f^*_{i,N}}{\partial v_i}\chi_i + \frac{\partial f^*_{i,N}}{\partial w_i}\bar{\eta}_i\right\} \\
&\approx D_i^2p^*_{1i}(1-p^*_i+p^*_{1i}) - 2D_ip^*_{1i}\left[\eta_i\left\{1-(1-p^*_i)\chi_i\right\} + \bar{\eta}_i(1-p^*_i)\chi_i + \bar{\eta}_ip^*_i\right] \\
&= -2p^*_{i1}D_ip^*_i\bar{\eta}_i\bar{\chi}_i - D_i^2p^*_{i1}(1+P^*_{2i}),
\end{aligned}$$

which only depends on PSP i. Finally, A_{ij} has the same form as Q_2 defined in the evaluation of $\nabla f_{i,N}^\top \Sigma_i \nabla f_{i,N}$. Therefore, $\nabla f_{i,N}^{*\top} \Sigma_i^* \nabla f_{i,N}^*$ and $\nabla f_{i,N}^\top \Sigma_i \nabla f_{i,N}$ have the same quadratic form and the vaiance of $\hat{N}_i$ is given by (A.12).

A.1.4 Derivation of var $\left(\hat{\phi}_i\right)$ for model M_t

In this section, we show the calculations that lead to the variance of $\hat{\phi}_i$,

$$\begin{aligned}\mathrm{var}(\hat{\phi}_i) &= \phi_i^2 \left\{ \frac{(1-p_{i+1}^*)\chi_{i+1}\left\{\bar{\eta}_{i+1}+p_{i+1}^*\eta_{i+1}\right\}}{N_{i+1}p_{i+1}^*\bar{\eta}_{i+1}\bar{\chi}_{i+1}} + \frac{(1-p_i^*)\bar{\eta}_i\chi_i}{N_i p_i^*\bar{\chi}_i(\bar{\eta}_i+\eta_i p_i^*)} + \frac{1-\phi_i}{N_{i+1}\bar{\eta}_{i+1}} \right\} \quad \text{(A.14)} \\ &- \phi_i^2 \left\{ \frac{(1-p_{i+1}^*)(\bar{\eta}_{i+1}+p_{i+1}^*\eta_{i+1})^2\chi_{i+1}^2 P_{2,i+1}^*}{N_{i+1}\bar{\eta}_{i+1}\bar{\chi}_{i+1}p_{i+1}^*(\bar{\eta}_{i+1}\bar{\chi}_{i+1}p_{i+1}^* + D_{i+1}P_{2,i+1}^*)} + \frac{(1-p_i^*)\bar{\eta}_i\chi_i^2 P_{2,i}^*}{N_i\bar{\chi}_i p_i^*(\bar{\eta}_i\bar{\chi}_i p_i^* + D_i P_{2,i}^*)} \right\}.\end{aligned}$$

Throughout the section, we give details on the evaluation of the variance of $\hat{n}_i$ and of $\hat{M}_i$,

$$\mathrm{var}\left(\hat{M}_i\right) = N_i(1-p_i^*)\bar{\eta}_i\chi_i(\bar{\eta}_i+p_i^*\eta_i)(p_i^*\bar{\eta}_i+P_{2i}^*\eta_i)/\left\{p_i^*(p_i^*\bar{\chi}_i\bar{\eta}_i+D_i\times P_{2i}^*)\right\}. \quad \text{(A.15)}$$

Derivation of var $(\hat{n}_i)$

The covariance between C_i and $\hat{N}_i$ is evaluated as N_i times the fourth entry of the vector $-\Sigma_i\nabla f_{i,N}^*/A_i^*$. It is found to be equal to $N_i\sum_{j=1}^{\ell_i} p_{ij}$. Thus the linearization approximation to the variance of $\hat{n}_i$ is

$$\begin{aligned}\mathrm{var}\left(\hat{n}_i\right) &= (P_{2i}^*)^2\mathrm{var}(\hat{N}_i) + N_i\{(p_i^*)^2 + (p_i^*-P_{2i}^*)(1-p_i^*)\} \\ &= N_i\left[P_{2i}^*(1-p_i^*)\left\{P_{2i}^*D_i/(p_i^*\bar{\eta}_i\bar{\chi}_i+P_{2i}^*D_i)-1\right\}+p_i^*\right] \\ &= N_i p_i^*\left\{1-(1-p_i^*)\bar{\eta}_i\bar{\chi}_i P_{2i}^*/(p_i^*\bar{\eta}_i\bar{\chi}_i+D_i\times P_{2i}^*)\right\}.\end{aligned}$$

Derivation of var $\left(\hat{M}_i\right)$

The only term that changes in the evaluation of var $\left(\hat{M}_i\right)$ from the previous derivations of (A.1.2) is the last term, involving the variance of $\hat{n}_i$,

$$\begin{aligned}Q_3 &= N_i p_i^*\left\{(\bar{\eta}_i+p_i^*\eta_i)\chi_i/p_i^*\bar{\chi}_i\right\}^2\left\{1-(1-p_i^*)\bar{\eta}_i\bar{\chi}_i P_{2i}^*/(p_i^*\bar{\chi}_i\bar{\eta}_i+D_i\times P_{2i}^*)\right\} \\ &= \underbrace{N_i p_i^*\left\{(\bar{\eta}_i+p_i^*\eta_i)\chi_i/(p_i^*\bar{\chi}_i)\right\}^2}_{Q_{3,1}} \\ &- \underbrace{N_i p_i^*\left\{(\bar{\eta}_i+p_i^*\eta_i)\chi_i/(p_i^*\bar{\chi}_i)\right\}^2\left\{(1-p_i^*)\bar{\eta}_i\bar{\chi}_i P_{2i}^*/(p_i^*\bar{\chi}_i\bar{\eta}_i+D_i\times P_{2i}^*)\right\}}_{Q_{3,2}}.\end{aligned}$$

First we calculate the variance under an open population model, corresponding to the case $Q_{3,2} = P_{2i}^* = 0$. This gives

$$Q_1+Q_2+Q_{3,1} = N_i\chi_i\bar{\eta}_i(1-p_i^*)\left\{\bar{\eta}_i+p_i^*\eta_i\right\}/(p_i^*\bar{\chi}_i)+N_i\bar{\eta}_i.$$

Now adding $Q_{3,2}$ gives (A.15). The derivations of $\nabla\hat{\phi}_i$ and Γ_i are the exact same obtained in (A.1.2) and (A.1.2) respectively.

Evaluation of the variance of $\hat{\phi}_i$

The only term that changes from prior calculations in (A.1.2) is the contribution of the variances of $(\hat{M}_{i+1}, \hat{M}_i, u_i)$ to the quadratic form,

$$\begin{aligned} Q_2 &= \phi_i^2 \frac{(1-p_{i+1}^*)\chi_{i+1}\left\{\bar{\eta}_{i+1}+\eta_{i+1}p_{i+1}^*\right\}\left\{p_{i+1}^*\bar{\eta}_{i+1}+P_{2,i+1}^*\eta_{i+1}\right\}}{N_{i+1}p_{i+1}^*\bar{\eta}_{i+1}\left\{p_{i+1}^*\bar{\chi}_{i+1}\bar{\eta}_{i+1}+D_{i+1}\times P_{2,i+1}^*\right\}} \\ &+ \phi_i^2\left[\frac{(1-p_i^*)\bar{\eta}_i\chi_i\left\{p_i^*\bar{\eta}_i+P_{2i}^*\eta_i\right\}}{N_i p_i^*\left\{\bar{\eta}_i+\eta_i p_i^*\right\}\left\{p_i^*\bar{\chi}_i\bar{\eta}_i+D_i\times P_{2i}^*\right\}}\right] \\ &+ \phi_i^2\left[\frac{1}{N_{i+1}\bar{\eta}_{i+1}}+\frac{1}{N_i(\bar{\eta}_i+\eta_i p_i^*)}\right]. \end{aligned}$$

Finally, adding up Q_1 obtained in (A.1.2) and Q_2 gives (A.14) .

A.2 Matériel supplémentaire

A.2.1 Simulation study

This section is a complement to Section 5.2 presenting the Monte Carlo validation of the bootstrap variance estimation method. The within PSP capture probabilites were generated using Darroch model,

$$p_{ik} = \binom{7}{k}\exp\left\{\beta k+\tau k^2/2\right\}/\sum_{j=0}^{7}\binom{7}{j}\exp\left\{\beta+\tau j^2/2\right\}, \quad k=0,\ldots,7 \qquad \text{(A.16)}$$

with parameters $(\beta,\ \tau) = (-3.3,\ 0.6)\,,\ (-2.85,\ 0.6)$ corresponding to $p^* = 0.3,\ 0.5$. The daily capture probability for a unit was simulated using the method proposed in Section 2.2 of Rivest and Baillargeon (2007). We set the survival probabilities at $\phi = 0.6,\ 0.8$.

We ran 1000 replications for the Monte Carlo study ; for each replication there was a burn-in period of 20 PSPs. We calculated the relative bias of $\hat{N}_5$ as $RB\left(\hat{N}_5\right) = \left(\sum_i \hat{N}_{5i}/1000 - N_5\right)/N_5$. The mean squared error, $MSE\left(\hat{N}_5\right) = \left\{\sum_i\left(\hat{N}_{5i}-N_5\right)^2/1000\right\}$, the relative root mean squared error, $RRMSE\left(\hat{N}_5\right) = MSE\left(\hat{N}_5\right)^{1/2}/N_5$. For each replication i of the Monte Carlo simulation, we ran $L = 100$ replications of the bootstrap described in Section 5.2 to calculate the bootstrap variance for $\hat{N}_5$, $v\left(\hat{N}_5\right)$. The relative bias for $E\left\{v\left(\hat{N}_5\right)\right\}$ is calculated as $RB\left[E\left\{v\left(\hat{N}_5\right)\right\}\right] = \left[E\left\{v\left(\hat{N}_5\right)\right\} - MSE\left(\hat{N}_5\right)\right]/MSE\left(\hat{N}_5\right)$; the 95% confidence interval for $\hat{N}_5$ is $\exp\left[\log\left(\hat{N}_5\right) \pm 1.96 s.e\left\{\log\left(\hat{N}_5\right)\right\}\right]$; the expected relative length of the confidence interval is calculated as $RLCI\left(\hat{N}_5\right) = (UB-LB)/N_5$, where UB and LB are respectively the expected upper and lower bounds of the confidence interval for N_5. In Table A.1, the Monte Carlo standard errors, $RRMSE\left(\hat{N}_5\right)/1000^{1/2}$, of $RB\left(\hat{N}_5\right)$ are also provided in parenthesis. They show that $\hat{N}_5$ has a negative bias which is important in scenario ED2.

TABLE A.1 – Simulation results for the estimation of N_5 under the robust design model for M^t_{Dh} with $p^* = 0.3,\ 0.5$, $\phi = 0.6,\ 0.8$ and three scenarios for the entry process. All the results are presented in percentages

ϕ	p^*	Scenario	$RB(\hat{N}_5)$	$RB(E(v(\hat{N}_5)))$	$RRMSE$	95% Cov.	$RLCI(\hat{N}_5)$
0.8	0.5	RD	-0.02 (0.14)	12.79	4.54	93.3	18.34
		ED1	-0.94 (0.15)	14.92	4.73	95.9	19.58
		ED2	-6.68 (0.26)	-63.23	8.12	74.2	18.99
	0.3	RD	-2.2 (0.25)	5.8	7.8	95.5	30.7
		ED1	-3.28 (0.24)	11.41	7.54	95	30.54
		ED2	-9.17 (0.35)	-53.08	11.19	78.9	29.44
0.6	0.5	RD	0.72 (0.29)	-0.76	9.31	93.9	35.3
		ED1	-0.2 (0.30)	-2.45	9.6	93.4	36.12
		ED2	-11.36 (0.48)	-63.83	15.10	72.4	34.56
	0.3	RD	-4.34 (0.46)	8.35	14.61	93.3	58.12
		ED1	-5.3 (0.50)	-5.16	15.7	93.7	58.14
		ED2	-17.98 (0.73)	-60.07	23.17	73.8	55.59

Additional simulations, using 1000 replications, were carried out for scenario ED2 to investigate whether the relative change in population size, defined as $inc = (N_{35} - N_5)/N_5$, could be estimated accurately even if absolute population estimators were biased. The set-up of this second simulation is presented in Figure A.1 : after a burn-in period of 21 PSPs data are collected in weeks 1 to 9 ; this is used to estimate N_5. Starting on week 10 daily births are multiplied by $1 + inc$ and the process runs for 29 more weeks. Data are collected in weeks 31 to 39 to estimate N_{35}.

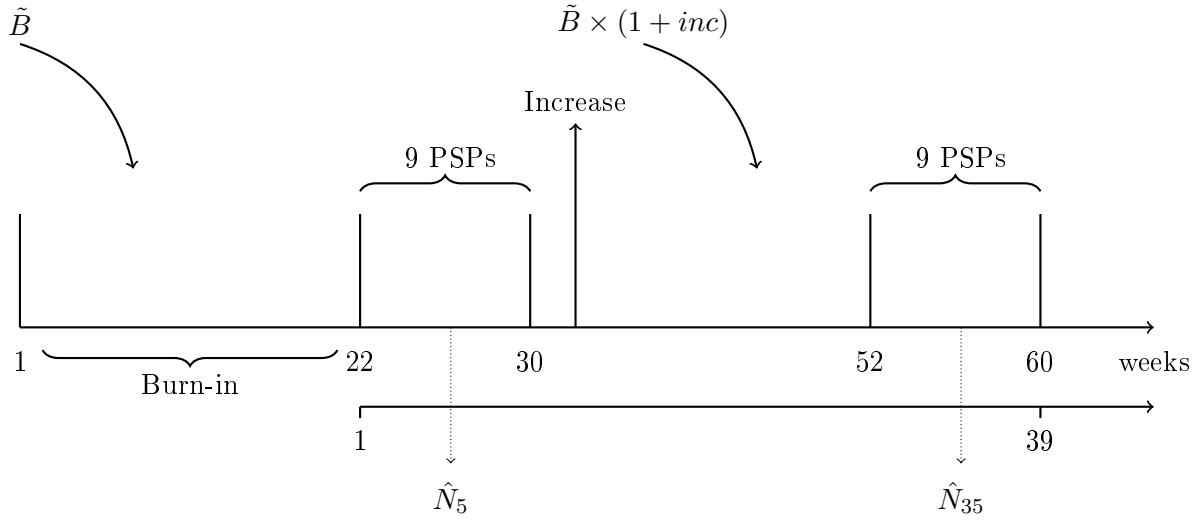

FIGURE A.1 – Data simulation process

For each replication i, the relative increase $\hat{inc}_i = \left(\hat{N}_{35,i} - \hat{N}_{5,i}\right)/\hat{N}_{5,i}$ is calculated. A bootstrap variance for $\hat{inc}$, calculated using 100 bootstrap samples, was computed using the formula $v\left(\hat{inc}\right) = \left(\hat{N}_{35}/\hat{N}_5\right)^2 \times \left[v\left\{\log\left(\hat{N}_5\right)\right\} + v\left\{\log\left(\hat{N}_{35}\right)\right\}\right]$, where $v\left\{\log\left(\hat{N}_5\right)\right\}$ and $v\left\{\log\left(\hat{N}_{35}\right)\right\}$ are the bootsrap variances for $\log\left(\hat{N}_5\right)$ and $\log\left(\hat{N}_{35}\right)$ respectively ; the 95% confidence interval for $\hat{inc}$ is $\left[\hat{inc} \pm 1.96\sqrt{v\left(\hat{inc}\right)}\right]$. The bias of $\hat{inc}$, with its Monte Carlo standard error, its root mean squared error and the coverage of its 95% confidence interval are reported in Table A.2. In general, the population increase is well estimated and the bootstrap confidence level is equal to the target value. In most cases, there is a small positive bias, which is always less than 10% of the true value of inc. Relatively large RMSEs are found when both p^* and ϕ are small.

TABLE A.2 – The bias, with its Monte Carlo standard error in parenthesis, the root mean squared error and the 95% coverage of the estimator of inc. The results are expressed in percentages.

ϕ	p^*	inc	$B(\hat{inc})$	$RMSE(\hat{inc})$	95%Cov.
0.8	0.5	20	0.45 (0.2)	8	98
		50	0.08 (0.3)	9	96
		80	0.26 (0.3)	10	98
	0.3	20	0.71 (0.4)	12	97
		50	0.8 (0.5)	14	97
		80	0.59 (0.7)	24	97
0.6	0.5	20	0.74 (0.5)	16	96
		50	1.75 (0.7)	21	94
		80	-1.40 (0.7)	22	94
	0.3	20	2.32 (1.0)	32	95
		50	1.82 (1.1)	35	94
		80	5.30 (1.3)	42	95

A.2.2 Robustness investigations

The robustness of the estimates obtained was investigated by fitting the same model to capture-recapture data from a different metropolitan area. The two sets of estimates for $\{p_i^*\}$ are provided in Figure A.2. They are very similar supporting the statement that the capture mechanism is the same in the two metropolitan areas.

Alternative data sets were created by removing the first few days of data. In Figure A.3, the Sunday data set is the original data set. The Monday one is obtained by dropping the first day : a PSP starts one day later than in the original data. In the Tuesday data set, it starts two days later. The population size estimates appear to be invariant with respect to a redefinition of the starting day of a PSP.

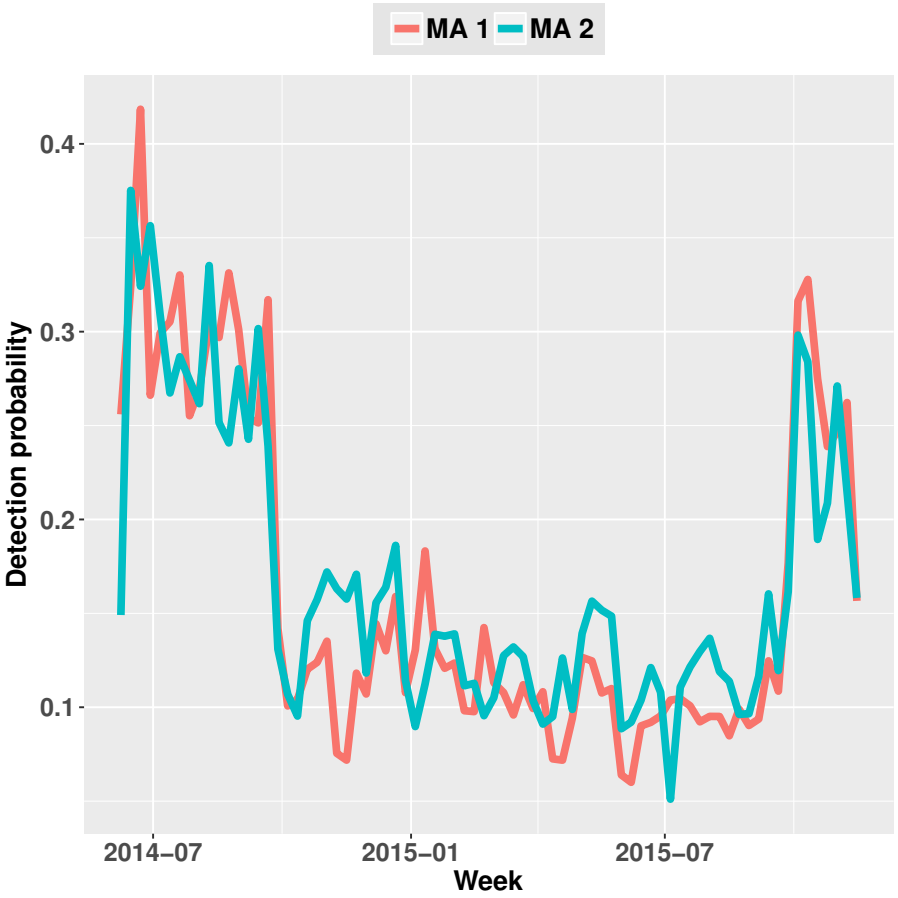

FIGURE A.2 – Evolution of the detection probability for 76 weeks and for Metropolitan Area (MA) 1 and 2.

A.2.3 Demographic parameter and capture probability estimates for the app data

In this section, estimates of the demographic parameters and the capture probabilities are presented along with their coefficients of variation. The results are presented for the first 20 weeks of the experiment ; results for the 76 PSPs can be obtained in .xlsx file named *E*stimations provided as a supplementary material.

Figure A.4 present boxplots of the relative efficiencies computed with respect to the robust design estimators for the 76 PSPs ; for the Jolly Seber estimators this efficiency is defined by $\{CV(\hat{N}_i^{JS})/CV(\hat{N}_i^{RD})\}^2$. The results show that the robust design provides estimates of N_i that are much more efficient that those of the closed population and the Jolly-Seber models. Furthermore, the gain in precision is more important for the closed population estimates.

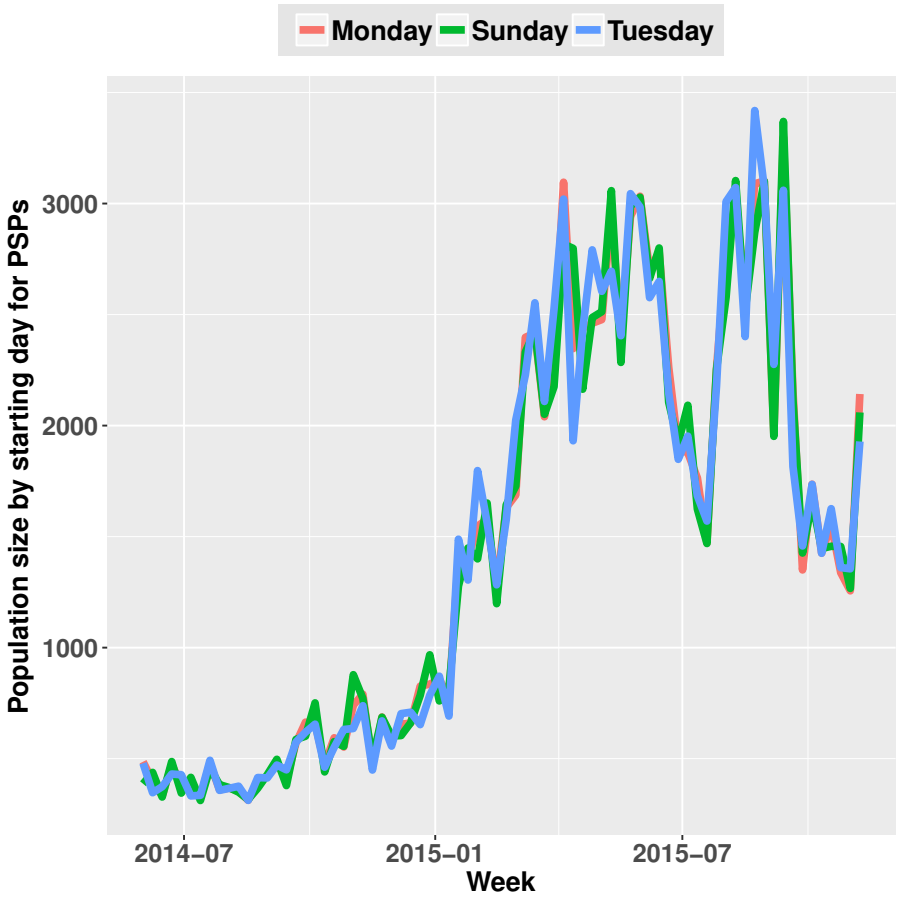

FIGURE A.3 – Evolution of the population size for 76 weeks and for three starting days (Sunday, Monday, Tuesday) for the PSPs.

TABLE A.3 – Clientele size estimates and their coefficients of variation for the first 20 weeks of the experiment, under models M_h for a closed population, Jolly-Seber and M_{Dh}^t.

Parameter	Closed Pop.	Robust Design M_h	Open Pop.
$\hat{N}_1$	387 (27)	387 (27)	-
$\hat{N}_2$	672 (22)	438 (17)	315 (26)
$\hat{N}_3$	461 (21)	327 (14)	275 (18)
$\hat{N}_4$	708 (26)	487 (16)	411 (21)
$\hat{N}_5$	291 (26)	345 (14)	365 (16)
$\hat{N}_6$	446 (24)	416 (13)	405 (17)
$\hat{N}_7$	290 (27)	312 (13)	317 (15)
$\hat{N}_8$	375 (27)	457 (15)	492 (18)
$\hat{N}_9$	300 (27)	383 (14)	416 (17)
$\hat{N}_{10}$	311 (24)	370 (13)	392 (15)
$\hat{N}_{11}$	332 (25)	349 (12)	354 (15)
$\hat{N}_{12}$	280 (23)	316 (12)	327 (13)
$\hat{N}_{13}$	355 (24)	367 (12)	370 (14)
$\hat{N}_{14}$	368 (28)	428 (13)	446 (16)
$\hat{N}_{15}$	465 (25)	497 (13)	509 (16)
$\hat{N}_{16}$	248 (22)	379 (12)	464 (15)
$\hat{N}_{17}$	686 (48)	587 (21)	565 (28)
$\hat{N}_{18}$	292 (37)	601 (21)	793 (28)
$\hat{N}_{19}$	560 (49)	751 (22)	818 (29)
$\hat{N}_{20}$	212 (41)	440 (20)	533 (26)

TABLE A.4 – Survival probability estimates and their coefficients of variation for the first 20 weeks of the experiment, under Jolly-Seber and M_{Dh}^t models.

Parameter	Closed Pop.	Robust Design M_h	Open Pop.
$\hat{\phi}_1$	-	0.964 (6)	0.810 (3)
$\hat{\phi}_2$	-	0.554 (12)	0.543 (15)
$\hat{\phi}_3$	-	0.768 (11)	0.742 (12)
$\hat{\phi}_4$	-	0.632 (12)	0.689 (14)
$\hat{\phi}_5$	-	0.835 (9)	0.804 (9)
$\hat{\phi}_6$	-	0.642 (12)	0.656 (14)
$\hat{\phi}_7$	-	0.869 (8)	0.896 (8)
$\hat{\phi}_8$	-	0.766 (11)	0.784 (12)
$\hat{\phi}_9$	-	0.704 (12)	0.699 (14)
$\hat{\phi}_{10}$	-	0.705 (11)	0.694 (13)
$\hat{\phi}_{11}$	-	0.699 (10)	0.707 (11)
$\hat{\phi}_{12}$	-	0.842 (9)	0.835 (9)
$\hat{\phi}_{13}$	-	0.894 (7)	0.912 (7)
$\hat{\phi}_{14}$	-	0.829 (9)	0.826 (9)
$\hat{\phi}_{15}$	-	0.755 (10)	0.873 (10)
$\hat{\phi}_{16}$	-	0.872 (7)	0.749 (9)
$\hat{\phi}_{17}$	-	0.648 (20)	0.749 (23)
$\hat{\phi}_{18}$	-	0.988 (8)	0.932 (10)
$\hat{\phi}_{19}$	-	0.715 (17)	0.776 (20)
$\hat{\phi}_{20}$	-	0.758 (14)	0.691 (17)

TABLE A.5 – Capture probability estimates and their coefficients of variation for the first 20 weeks of the experiment, under models M_h for a closed population, Jolly-Seber and M_{Dh}^t.

Parameter	Closed Pop.	Robust Design M_h	Open Pop.
$\hat{p}_1^*$	0.256 (25)	0.256 (22)	-
$\hat{p}_2^*$	0.222 (20)	0.324 (16)	0.474 (27)
$\hat{p}_3^*$	0.310 (18)	0.418 (13)	0.520 (19)
$\hat{p}_4^*$	0.191 (23)	0.266 (16)	0.328 (22)
$\hat{p}_5^*$	0.347 (23)	0.299 (15)	0.276 (18)
$\hat{p}_6^*$	0.287 (22)	0.305 (14)	0.316 (19)
$\hat{p}_7^*$	0.352 (24)	0.330 (14)	0.321 (16)
$\hat{p}_8^*$	0.304 (25)	0.255 (16)	0.232 (20)
$\hat{p}_9^*$	0.334 (24)	0.269 (15)	0.240 (19)
$\hat{p}_{10}^*$	0.354 (22)	0.303 (13)	0.281 (17)
$\hat{p}_{11}^*$	0.310 (23)	0.297 (13)	0.291 (16)
$\hat{p}_{12}^*$	0.368 (20)	0.331 (12)	0.315 (14)
$\hat{p}_{13}^*$	0.310 (21)	0.301 (13)	0.297 (16)
$\hat{p}_{14}^*$	0.290 (26)	0.254 (15)	0.240 (18)
$\hat{p}_{15}^*$	0.267 (23)	0.251 (14)	0.244 (17)
$\hat{p}_{16}^*$	0.455 (20)	0.317 (13)	0.243 (17)
$\hat{p}_{17}^*$	0.123 (48)	0.142 (22)	0.149 (29)
$\hat{p}_{18}^*$	0.195 (34)	0.101 (22)	0.072 (30)
$\hat{p}_{19}^*$	0.138 (51)	0.104 (22)	0.094 (29)
$\hat{p}_{20}^*$	0.063 (42)	0.120 (21)	0.094 (27)

TABLE A.6 – Estimates of new arrivals and their coefficients of variation for the first 20 weeks of the experiment, under Jolly-Seber and M_{Dh}^t models.

Parameter	Closed Pop.	Robust Design M_h	Open Pop.
$\hat{B}_1$	-	65 (62)	-
$\hat{B}_2$	-	84 (90)	20 (534)
$\hat{B}_3$	-	236 (29)	262 (38)
$\hat{B}_4$	-	37 (189)	60 (346)
$\hat{B}_5$	-	128 (43)	153 (45)
$\hat{B}_6$	-	45 (134)	0 (382)
$\hat{B}_7$	-	186 (34)	284 (38)
$\hat{B}_8$	-	33 (146)	0 (198)
$\hat{B}_9$	-	100 (59)	66 (120)
$\hat{B}_{10}$	-	88 (63)	81 (106)
$\hat{B}_{11}$	-	72 (76)	81 (118)
$\hat{B}_{12}$	-	101 (43)	139 (50)
$\hat{B}_{13}$	-	100 (49)	137 (66)
$\hat{B}_{14}$	-	142 (48)	102 (72)
$\hat{B}_{15}$	-	3 (263)	44 (190)
$\hat{B}_{16}$	-	256 (40)	160 (56)
$\hat{B}_{17}$	-	221 (86)	370 (177)
$\hat{B}_{18}$	-	157 (73)	224 (74)
$\hat{B}_{19}$	-	0 (237)	0 (315)
$\hat{B}_{20}$	-	243 (70)	178 (98)

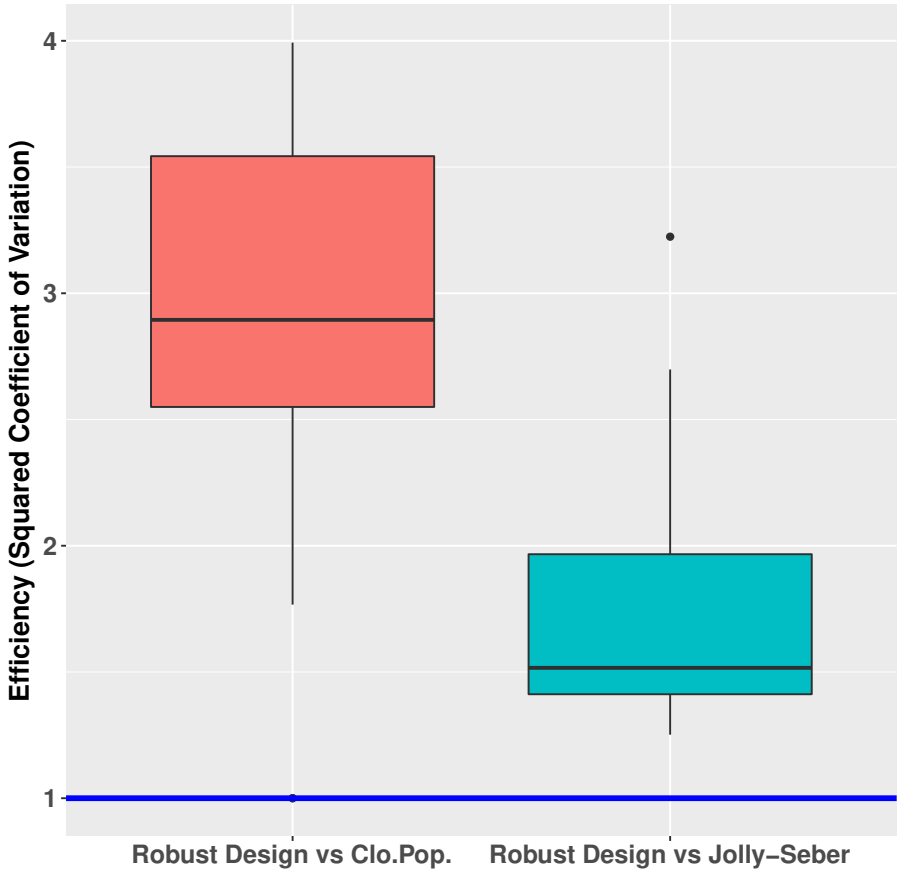

FIGURE A.4 – Efficiency comparison (Squared Coefficient of variation) between the robust design estimate of N_i and those obtained under models M_h closed population and Jolly-Seber. The relative efficiencies are calculated with 1000 bootstrap samples and their values from the 76 PSP plotted.

Annexe B

Arguments techniques et matériel supplémentaire du chapitre 3

B.1 Arguments techniques

B.1.1 Proof of Proposition 1

Clearly

$$E\left\{\sum_{\Omega_i^*} n_\omega\right\} \leq N_i$$

One can write $\hat{N}_i = g(n_\omega; \boldsymbol{\omega} \in \Omega_i^*)$ and $N_i = g(\mu_\omega; \boldsymbol{\omega} \in \Omega_i^*)$ where g satisfies (3.3). Taking the partial derivative with respect to c on the two sides of the equation $cN_i = g(c\mu_\omega; \boldsymbol{\omega} \in \Omega_i^*)$ and setting $c = 1$ give

$$N_i = \sum_{\boldsymbol{\omega} \in \Omega_i^*} \frac{\partial}{\partial \mu_\omega} g(\mu_\omega; \boldsymbol{\omega} \in \Omega_i^*)\mu_\omega = \nabla^\top \mu,$$

where ∇ is the vector of partial derivatives of g with respect to $\{\mu_\omega\}$ and μ is the vector of the μ_ω for $\boldsymbol{\omega} \in \Omega_i^*$. Now

$$\begin{aligned} \mathrm{var}(\hat{N}_i - \tilde{N}_i) &\approx \mathrm{var}_P(\hat{N}_i) - 2\mathrm{cov}(\hat{N}_i, \tilde{N}_i) + N_i \\ &= \mathrm{var}_P(\hat{N}_i) - N_i, \end{aligned}$$

since

$$\mathrm{cov}(\hat{N}_i, \tilde{N}_i) \approx \sum_{\boldsymbol{\omega} \in \Omega_i^*} \frac{\partial}{\partial \mu_\omega} g(\mu_\omega; \boldsymbol{\omega} \in \Omega_i^*)\mathrm{cov}(n_\omega, \tilde{N}_i),$$

and $\mathrm{cov}(n_\omega, \tilde{N}_i) = \mu_\omega$ for $\boldsymbol{\omega} \in \Omega_i$. Since $\hat{N}_i - N_i$ and $\tilde{N}_i - N_i$ can be approximated by a linear function of asymptotically normal random variables, the asymptotic distribution of $\hat{N}_i - \tilde{N}_i$ is normal with asymptotic variance $\mathrm{var}_P(\hat{N}_i) - N_i$. The derivation of the asymptotic covariance between $\hat{N}_i^a - \tilde{N}_i$ and $\hat{N}_i^b - \tilde{N}_i$ relies on a similar argument.

B.1.2 Derivation of Equation (3.9)

For models satisfying (3.7) the Poisson Fisher information matrix for $\boldsymbol{\theta} = (\gamma, \beta^\top)^\top$ is

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

Its inverse is

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

The asymptotic expansion for the Poisson maximum likelihood estimator of $\boldsymbol{\theta}$ is

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

where $S_p(\boldsymbol{\theta})$ is the Poisson score function :

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

where $p^* = 1 - e^\gamma/N$ is the probability of being captured at least once. A simple linearization gives

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

From (B.1), one gets

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

Plugging this in the linearization for $\hat{N}^{CP}$ yields Equation (3.9).

B.1.3 Proof of Proposition 2

Considering Section 3.4, the estimator $\hat{p}_i^*$ is asymptotically equivalent to a linear combination of $\hat{p}_i^{*,CP}$ and $\hat{p}_i^{*,JS}$. Plugging this combination in (3.18) gives

$$\hat{N}_i^{KEN} \approx \frac{n_i}{\hat{p}_i^{*,CP}/\mathrm{var}(\hat{p}_i^{*,CP}) + \hat{p}_i^{*,JS}/\mathrm{var}(\hat{p}_i^{*,JS})} \times \left\{ 1/\mathrm{var}(\hat{p}_i^{*,CP}) + 1/\mathrm{var}(\hat{p}_i^{*,JS}) \right\}. \tag{B.2}$$

This is a function of $\hat{N}_i^{CP}$ and N_i^{JS} whose large sample linearization is given by

$$\hat{N}_i^{KEN} - N_i \approx \frac{(\hat{N}_i^{CP} - N_i)/\mathrm{var}(\hat{p}_i^{*,CP}) + (\hat{N}_i^{JS} - N_i)/\mathrm{var}(\hat{p}_i^{*,JS})}{1/\mathrm{var}(\hat{p}_i^{*,CP}) + 1/\mathrm{var}(\hat{p}_i^{*,JS})}. \tag{B.3}$$

In this expansion the weight given to $\hat{N}_i^{CP}$ is

$$\frac{\text{var}(\hat{p}_i^{*,JS})}{\text{var}(\hat{p}_i^{*,JS}) + \text{var}(\hat{p}_i^{*,CP})} = \frac{\text{var}_M(\hat{N}_i^{JS})}{\text{var}_M(\hat{N}_i^{JS}) + \text{var}_M(\hat{N}_i^{CP})} + \frac{N_i(1-p_i^*)/p_i^*}{\text{var}_M(\hat{N}_i^{JS}) + \text{var}_M(\hat{N}_i^{CP})}.$$

This uses the formulae, given in Section 3.4, for the variance of $\hat{p}_i^*$ in terms of that for $\hat{N}_i$. As the weight given to $\hat{N}_i^{JS}$ is 1 minus that for N_i^{CP}, this leads to the following expansion

$$\hat{N}_i^{KEN} \approx \frac{\hat{N}_i^{CP}/\text{var}_M(\hat{N}_i^{CP}) + \hat{N}_i^{JS}/\text{var}_M(\hat{N}_i^{JS})}{1/\text{var}_M(\hat{N}_i^{CP}) + 1/\text{var}_M(\hat{N}_i^{JS})} + \frac{N_i(1-p_i^*)\left(\hat{N}_i^{CP} - \hat{N}_i^{JS}\right)}{p_i^*\left\{\text{var}_M(\hat{N}_i^{CP}) + \text{var}_M(\hat{N}_i^{JS})\right\}}. \tag{B.4}$$

The first part of (B.4) is clearly the combined estimator for N_i defined in (3.16). By taking the variance of (B.4) the covariance part is equal to 0 and the conditional variance of $\hat{N}_i^{KEN}$ is

$$\text{var}_M\left(\hat{N}_i^{KEN}\right) = \frac{1}{1/\text{var}_M(\hat{N}_i^{CP}) + 1/\text{var}_M(\hat{N}_i^{JS})} + \frac{N_i^2(1-p_i^*)^2}{(p_i^*)^2\left\{\text{var}_M(\hat{N}_i^{CP}) + \text{var}_M(\hat{N}_i^{JS})\right\}}.$$

Straighforward developments lead to

$$\text{var}_M\left(\hat{N}_i^{KEN}\right) = \frac{N_i D_i\,(1-p_i^*)}{p_i^*\bar{\chi}_i\bar{\eta}_i + D_i\,(1-p_i^*)\,A_i^*} + \frac{N_i\,(1-p_i^*)^2\,A_i^*\bar{\chi}_i\bar{\eta}_i\,(p_i^*)^{-1}}{p_i^*\bar{\chi}_i\bar{\eta}_i + D_i\,(1-p_i^*)\,A_i^*}, \tag{B.5}$$

where $A_i^* = (1-p_i^*)^{-1} - 1 - \mu_{X_i}^\top \Sigma_i^{-1}\mu_{X_i}$.

B.2 Matériel supplémentaire

B.2.1 The special case of model M_b within SP

Under model M_b, the sufficient statistics for estimating N are $\{u_j : j = 1, ..., \ell\}$, where u_j is the number of units captured for the first time on occasion j. One has $E(u_j) = \gamma_j = Np(1-p)^{j-1}$, where p is the capture probability. In other words

$$\log\gamma_j = \beta_0 + (j-1)\beta, \quad j = 1, 2, ..., \ell, \tag{B.6}$$

where $\beta_0 = \log Np$ and $\beta = \log(1-p)$. In terms of the log-linear parameters, $N = e_0^\beta/\left(1-e^\beta\right)$ and the number of missed units is $\mu_0 = e_0^\beta e^{\beta\ell}/\left(1-e^\beta\right)$ while the probability of being captured is $p^* = 1 - e^{\beta\ell}$.

Define the random variable X as the number of misses before the first capture, for a unit that has been captured ; X has a truncated geometric distribution with probability mass function

$$P(X = k) = \frac{e^{\beta k}}{c(\beta)}, \quad k = 0, \ldots, \ell - 1,$$

where $c(\beta) = \left(1-e^{\beta\ell}\right)/\left(1-e^\beta\right)$. The Poisson score vector and Fisher information matrix for M_b can be expressed in terms of moments of this random variable. For instance $E(\sum(j -$

$1)u_j) = Np^*\mu$ where $\mu = (1-p)/p - \ell(1-p^*)/p^*$ is the expectation of X. It can be evaluated as the first derivative of $\log\{c(\beta)\}$ with respect to β. Differentiating twice gives the variance, $\Sigma = (1-p)/p^2 - \ell^2(1-p^*)/(p^*)^2$.

The population size estimator for this model is $\hat{N}^b = e^{\hat{\beta}_0}/(1-e^{\hat{\beta}})$. For model (B.6) the Poisson Fisher information matrix for $\theta = (\beta_0, \beta)^\top$) is

$$I_p(\theta) = \sum_{j=1}^{\ell} \mu_j \begin{pmatrix} 1 \\ j-1 \end{pmatrix} \begin{pmatrix} 1 \\ j-1 \end{pmatrix}^\top = Np^* \begin{pmatrix} 1 & \mu \\ \mu & \mu^2+\Sigma \end{pmatrix}.$$

Its inverse is

$$I_p^{-1}(\theta) = \frac{1}{\Sigma N p^*} \begin{pmatrix} \mu^2+\Sigma & -\mu \\ -\mu & 1 \end{pmatrix}. \tag{B.7}$$

As $\hat{N}^{CP} = e^{\hat{\beta}_0}/(1-e^{\hat{\beta}})$, a simple linearization gives

$$\hat{N}^{CP} - N \approx \left\{e_0^\beta/(1-e^\beta)\right\}(\hat{\beta}_0 - \beta_0) + \left\{e_0^\beta e^\beta/(1-e^\beta)^2\right\}(\hat{\beta}-\beta). \tag{B.8}$$

The standard asymptotic expansion for Poisson maximum likelihood estimator gives,

$$\begin{pmatrix} \hat{\beta}_0 - \beta_0 \\ \hat{\beta} - \beta \end{pmatrix} = \frac{1}{\Sigma N p^*} \begin{pmatrix} \mu^2+\Sigma & -\mu \\ -\mu & 1 \end{pmatrix} \begin{pmatrix} n - Np^* \\ \sum_{j=1}^{\ell}(j-1)u_j - Np^*\mu. \end{pmatrix}$$

From (B.8), $\hat{N}^{CP} - N \approx N\{(\hat{\beta}_0 - \beta_0) + (1-p)(\hat{\beta}-\beta)/p\}$. This leads to the asymptotic expansion

$$\hat{N}^b - N \approx (\nabla \hat{N}^b)^\top \begin{pmatrix} n - Np^* \\ \sum_{j=1}^{\ell}(j-1)u_j - Np^*\mu \end{pmatrix}, \tag{B.9}$$

where

$$\nabla \hat{N}^b = \{1/(\Sigma p^*)\} \begin{pmatrix} \Sigma - \mu f & , & f \end{pmatrix}^\top,$$

and $f = \ell(1-p^*)/p^*$. The asymptotic multinomial variance of $\hat{N}^b$ is

$$\mathrm{var}_M\left(\hat{N}^b\right) = (\nabla \hat{N}^b)^\top \mathrm{cov} \begin{pmatrix} n \\ \sum(j-1)u_j \end{pmatrix} \nabla \hat{N}^b - N = N \frac{1 - p^* + \Sigma^{-1}\{\ell(1-p^*)/p^*\}^2}{p^*}. \tag{B.10}$$

The closed population model estimator $\hat{N}_i^b$ for SP i can be obtained when a trap response is assumed within SPs. The corresponding Jolly-Seber estimator $\hat{N}_i^{JS}$ is obtained by pooling the data related to SP i. To evaluate the covariance between $\hat{N}_i^b$ and $\hat{N}_i^{JS}$, we consider the expansions for the Jolly-Seber estimator given in Equation (12) of the manuscript and (B.9). The covariance is $(\nabla \hat{N}_i^{JS})^\top \mathrm{cov}\{(n_i, u_i, v_i, w_i)^\top, (n_i, \sum_{j=1}^{\ell_i}(j-1)u_j^i\}\nabla \hat{N}_i^b$, where u_j^i is the statistic

u_j for the ith SP . The covariances involving Poisson random variables are easily derived, as discussed in Section 4 of the manuscript. The Poisson covariance is given by

$$N_i p_i^* (\nabla \hat{N}_i^{JS})^\top (1, \eta_i, \chi_i, \bar{\eta}_i)^\top (1, \mu_i) \nabla \hat{N}_i^b.$$

We have have already proved that $(\nabla \hat{N}_i^{JS})^\top (\eta_i, \chi_i, \bar{\eta}_i, 1)^\top = 1$, see equation (14). Straightforward developments show $(1, \mu_i)\nabla \hat{N}_i^b = \{1/(\Sigma_i p_i^*)\}(\Sigma_i - \mu_i f_i + \mu_i f_i) = 1/p_i^*$. The Poisson covariance is then equal to N_i and the multinomial covariance is equal to 0. This result is summarized in the following proposition.

Proposition 5 *The asymptotic covariance between $\hat{N}_i^b$ and $\hat{N}_i^{JS}$ satisfies*

$$cov_M(\hat{N}_i^b\ , \hat{N}_i^{JS}) = 0,\ i = 2, \ldots, I-1. \tag{B.11}$$

and the estimators for N_i obtained from within and between information pertaining to the same SP i are independent.

Under the robust design model M_b^t (superscript t means that the parameter for M_b varies with the SP), the maximum likelihood estimators for N_i are unknown as the log-linear framework proposed in Rivest & Daigle (2004) does not work in this case. Still, considering Proposition 5, an optimal estimator of N_i is obtained by using Equation (16) of the manuscript. A closed form expression for the asymptotic variance of the combined estimator for N_i is

$$\mathrm{var}_M(\hat{N}_i) \quad = \quad \frac{N_i D_i (1 - p_i^*) \left\{1 + \Sigma_i^{-1}(1 - p_i^*)(\ell_i / p_i^*)^2\right\}}{D_i p_i^* + p_i^* \bar{\chi}_i \bar{\eta}_i \left\{1 + \Sigma_i^{-1}(1 - p_i^*)(\ell_i / p_i^*)^2\right\}}.$$

Using the expansion for the Jolly-eber estimator for $\hat{N}_i^{JS}$ and (B.9), one can linearize the combined estimator $\hat{N}_i$ for M_b^t. In the calculations, one finds a non-null coefficient for n_i. Thus $\{(u_i, v_i, n_i\}$ and $\sum (j-1)u_j^i$ for $i = 1, \ldots, I$ are the sufficient statistics for M_b^t. The dimension of the vector of sufficient statistics, $4I-1$, is larger than the dimension of the parameter space, $3I - 1$. This is an important difference with the robust design models constructed with closed population model satisfying Equation (7) of the manuscript.

Annexe C

Arguments techniques et matériel supplémentaire du chapitre 4

In the derivations, we approximate u_i by its expectation,

$$u_i \approx N_i \eta_i p_i^*$$

as, when N_i goes to ∞, the ratio of u_i over its expectation converges to 1 in probability.

C.1 Derivation of Equation (4.7)

The Poisson variance of $\hat{N}_i$ is

$$\mathrm{Var}_P\left(\hat{N}_i\right) \approx E\left\{\mathrm{Var}\left(\hat{N}_i|\boldsymbol{u^i}\right)\right\} + \mathrm{Var}\left(N_i^*\right), \tag{C.1}$$

where N_i^* is defined in Section 4.3.2 .The second quantity on the right hand side of (C.1) is easily obtained as

$$\mathrm{Var}\left(N_i^*\right) = \frac{N_i \eta_i}{p_i^*} + \sum_{k=1}^{i-1} N_k \eta_k p_k^* \left(\prod_{s=k}^{i-1} \phi_s\right)^2, \quad i = 2, \ldots, I. \tag{C.2}$$

To evaluate $E\left\{\mathrm{var}\left(\hat{N}_i|\boldsymbol{u^i}\right)\right\}$, one define an asymptotic expansion for $\hat{N}_i$ conditional on $\boldsymbol{u^i}$. One have

$$\boldsymbol{u^i}\hat{\boldsymbol{\beta}}_{\boldsymbol{i}} - \boldsymbol{u^i}\boldsymbol{\beta_i} \approx \boldsymbol{u^i}\boldsymbol{\nabla}\hat{\boldsymbol{\beta}}_{\boldsymbol{i}} \begin{pmatrix} \hat{\phi}_1 - \phi_1 \\ \hat{\phi}_2 - \phi_2 \\ \ldots \\ \hat{\phi}_{i-1} - \phi_{i-1} \\ \hat{p}_i^* - p_i^* \end{pmatrix}, \tag{C.3}$$

where $\boldsymbol{\nabla\hat{\beta}_i}$ is the limit of the $i \times i$ matrix of partial derivatives of $\boldsymbol{\beta_i}$ with respect to $\hat{\phi}_1, \ldots, \hat{\phi}_{i-1}, \hat{p}_i^*$,

$$\boldsymbol{\nabla\hat{\beta}_i} = \begin{pmatrix} \prod_{\substack{s=1 \\ s\neq 1}}^{i-1} \phi_s & \prod_{\substack{s=1 \\ s\neq 2}}^{i-1} \phi_s & \ldots & \prod_{\substack{s=1 \\ s\neq i-1}}^{i-1} \phi_s & 0 \\ 0 & \prod_{\substack{s=2 \\ s\neq 2}}^{i-1} \phi_s & \ldots & \prod_{\substack{s=2 \\ s\neq i-1}}^{i-1} \phi_s & 0 \\ \ldots & \ldots & \ldots & \ldots & 0 \\ 0 & 0 & \ldots & 1 & 0 \\ 0 & 0 & \ldots & 0 & -1/(p_i^*)^2 \end{pmatrix}. \tag{C.4}$$

Using (C.3), $E\left\{\text{Var}\left(\hat{N}_i|\boldsymbol{u^i}\right)\right\}$ can be expressed as

$$E\left\{\text{Var}\left(\hat{N}_i|\boldsymbol{u^i}\right)\right\} \approx E\left\{\boldsymbol{u^i}\left(\boldsymbol{\nabla\hat{\beta}_i}\right)^\top \boldsymbol{A}\left(\boldsymbol{\{\hat{\phi}_i\}, \hat{p}_i^*}\right)\boldsymbol{\nabla\hat{\beta}_i}\left(\boldsymbol{u^i}\right)^\top\right\}, \tag{C.5}$$

where $\boldsymbol{A}\left(\boldsymbol{\{\hat{\phi}_i\}, \hat{p}_i^*}\right)$ is the variance-covariance matrix of $\left(\hat{\phi}_1, \ldots, \hat{\phi}_{i-1}, \hat{p}_i^*\right)$. Finally, summing (C.2) and (C.5) gives (4.7).

C.2 Derivation of Equation (4.10)

Let $\tilde{N}_i$ be defined as $\tilde{N}_i = \tilde{U}_i + \sum_{k=1}^{i-1} \tilde{U}_{ki}$ where $\tilde{U}_{ki}$ is the number of units captured for the first time at SP k that survives till SP i. Equation (4.10) has two components : $\text{Var}\left(\hat{N}_i - N_i^*\right)$ and $\text{Var}\left(N_i^* - \tilde{N}_i\right)$.

The first component is given by

$$\text{Var}\left(\hat{N}_i - N_i^*\right) = E\left\{\text{Var}\left(\hat{N}_i - N_i^*|\boldsymbol{u^i}\right)\right\} + \text{Var}\left\{E\left(\hat{N}_i - N_i^*|\boldsymbol{u^i}\right)\right\}. \tag{C.6}$$

One can show that $E\left(\hat{N}_i - N_i^*|\boldsymbol{u^i}\right) \approx 0$. Hence, $\text{Var}\left\{E\left(\hat{N}_i - N_i^*|\boldsymbol{u^i}\right)\right\} = 0$. Thus, (C.6) is equal to (C.5).

To calculate the second component of (4.10), one need to rewrite $N_i^* - \tilde{N}_i$. This gives :

$$N_i^* - \tilde{N}_i = u_i/p_i^* - \tilde{U}_i + \sum_{k=1}^{i-1} u_k \prod_{s=k}^{i-1} \phi_s - \sum_{k=1}^{i-1} \tilde{U}_{ki}. \tag{C.7}$$

Since $\tilde{U}_i - u_i/p_i^*$ and $\sum_{k=1}^{i-1} u_k \prod_{s=k}^{i-1} \phi_s - \sum_{k=1}^{i-1} \tilde{U}_{ki}$ are two independent components, and $u_i|\tilde{U}_i \sim \text{Binomial}(\tilde{U}_i, p_i^*)$ and $\tilde{U}_{ki}|u_k \sim \text{Binomial}(u_k, \prod_{s=k}^{i-1} \phi_s)$, taking the Poisson variance of (C.7) gives

$$\text{Var}\left(N_i^* - \tilde{N}_i\right) = N_i\eta_i\left(1 - p_i^*\right)/p_i^* + \sum_{k=1}^{i-1} N_k\eta_k p_k^* \prod_{s=k}^{i-1} \phi_s \left(1 - \prod_{s=k}^{i-1} \phi_s\right). \tag{C.8}$$

Summing (C.5) and (C.8) gives

$$\begin{aligned}\mathrm{Var}_{ALT}\left(\hat{N}_i\right) = &\quad N_i\eta_i\left(1-p_i^*\right)/p_i^* + \textstyle\sum_{k=1}^{i-1} N_k\eta_k p_k^* \prod_{s=k}^{i-1}\phi_s\left(1-\prod_{s=k}^{i-1}\phi_s\right)\\ + &\quad \mathrm{Var}_P\left(\hat{N}_i\right) - \mathrm{Var}_P\left(N_i^*\right) + N_i\end{aligned}$$

The quantity $\mathrm{Var}_P\left(N_i^*\right)$ is given in (C.2). Straightforward developments lead to

$$\mathrm{Var}_{ALT}(\hat{N}_i) = \mathrm{Var}_P\left(\hat{N}_i\right) - N_i + 2\sum_{k=1}^{i-1} N_k\eta_k p_k^* \prod_{s=k}^{i-1}\phi_s\left(1-\prod_{s=k}^{i-1}\phi_s\right). \tag{C.9}$$

Finally, replacing all parameters with their estimates in (C.9) gives (4.10).

C.3 Equation (C.5) in the special case of model M_t^t with constant survival

The matrix of partial derivatives $\boldsymbol{\nabla\hat{\beta}_i}$ given in (C.4) simplifies to

$$\boldsymbol{\nabla\hat{\beta}_i} = \begin{pmatrix} (i-1)\phi^{i-2} & 0\\ (i-2)\phi^{i-3} & 0\\ \vdots & \vdots\\ \phi & 0\\ 1 & 0\\ 0 & -1/(p_i^*)^2\end{pmatrix}. \tag{C.10}$$

For model M_t^t, MARK provides the $(I\times J+1)\times(I\times J+1)$ variance-covariance matrix $\boldsymbol{A}\left(\hat{\boldsymbol{\phi}},\{\hat{\boldsymbol{p}}_{\boldsymbol{ij}}\}\right)$ for $\left(\hat{\phi},\{\hat{p}_{ij}\}\right)$. To obtain the $(I+1)\times(I+1)$ variance-covariance matrix $\boldsymbol{A}\left(\hat{\boldsymbol{\phi}},\{\hat{\boldsymbol{p}}_{\boldsymbol{i}}^*\}\right)$ for $(\hat{\phi},\{\hat{p}_i^*\})$, we use standard linearization methods. One can represent $\boldsymbol{A}\left(\hat{\boldsymbol{\phi}},\{\hat{\boldsymbol{p}}_{\boldsymbol{ij}}\}\right)$ as

$$\boldsymbol{A}\left(\hat{\boldsymbol{\phi}},\{\hat{\boldsymbol{p}}_{\boldsymbol{ij}}\}\right) = \left(\begin{array}{c|c}\mathrm{var}(\hat{\phi}) & \mathrm{cov}(\hat{\phi},\{\hat{p}_{ij}\})\\ \hline & \mathrm{var}(\{\hat{p}_{ij}\})\end{array}\right).$$

The probability of being captured at least once during SP i $(i=1,2,\ldots,I)$, p_i^*, is

$$p_i^* = 1-\prod_{j=1}^{\ell_i}\left(1-p_{ij}\right).$$

By linearization, one have

$$\mathrm{var}(\{\hat{p}_i^*\}) = \boldsymbol{\Psi_i}^\top \mathrm{var}(\{\hat{p}_{ij}\})\boldsymbol{\Psi_i}$$

and

$$\mathrm{cov}(\hat{\phi},\{\hat{p}_i^*\}) = \boldsymbol{\Psi_i}^\top \mathrm{cov}(\hat{\phi},\{\hat{p}_{ij}\}),$$

where $\mathbf{\Psi_i}$ is the limit of the vector of partial derivatives of $\hat{p}_i^*$ with respect to $\{\hat{p}_{ij}\}$,

$$\mathbf{\Psi_i} = \left(\prod_{\substack{s=1 \\ s \neq 1}}^{\ell_i} (1 - p_{is}), \prod_{\substack{s=1 \\ s \neq 2}}^{\ell_i} (1 - p_{is}), ..., \prod_{\substack{s=1 \\ s \neq \ell_i}}^{\ell_i} (1 - p_{is}) \right)^\top .$$

Now, the $(I+1) \times (I+1)$ variance-covariance matrix $\boldsymbol{A}\left(\hat{\boldsymbol{\phi}}, \{\hat{\boldsymbol{p}}_{\boldsymbol{i}}^*\}\right)$ is

$$\boldsymbol{A}\left(\hat{\boldsymbol{\phi}}, \{\hat{\boldsymbol{p}}_{\boldsymbol{i}}^*\}\right) = \left(\begin{array}{c|c} \text{var}(\hat{\phi}) & \text{cov}(\hat{\phi}, \{\hat{p}_i^*\}) \\ \hline & \text{var}(\{\hat{p}_i^*\}) \end{array} \right). \tag{C.11}$$

C.4 Equation (C.5) in the special case of model M_b^0 with constant survival

In the case of model M_b^0, the matrix of partial derivatives $\boldsymbol{\nabla \hat{\beta}_i}$ given in (C.4) simplifies to

$$\boldsymbol{\nabla \hat{\beta}_i} = \begin{pmatrix} (i-1)\phi^{i-2} & 0 \\ (i-2)\phi^{i-3} & 0 \\ \vdots & \vdots \\ \phi & 0 \\ 1 & 0 \\ 0 & -1/(p^*)^2 \end{pmatrix}. \tag{C.12}$$

MARK provides $(\hat{\phi}, \hat{p})$ and its estimated variance-covariance matrix $\boldsymbol{A}\left(\hat{\boldsymbol{\phi}}, \hat{\boldsymbol{p}}\right)$, which can be represented as

$$\boldsymbol{A}\left(\hat{\boldsymbol{\phi}}, \hat{\boldsymbol{p}}\right) = \begin{pmatrix} \text{var}(\hat{\phi}) & \text{cov}(\hat{\phi}, \hat{p}) \\ & \text{var}(\hat{p}) \end{pmatrix}. \tag{C.13}$$

The probability of being captured at lease once is $p^* = 1 - (1-p)^\ell$ ($\ell_1 = \ell_2 = \cdots = \ell_I = \ell$). Using standard linearization methods, one have

$$\text{var}(\hat{p}^*) = \{\ell(1-p)^{\ell-1}\}^2 \text{var}(\hat{p})$$

and

$$\text{cov}(\hat{\phi}, \hat{p}^*\}) = \ell(1-p)^{\ell-1} \text{cov}(\hat{\phi}, \hat{p}\}).$$

This gives

$$\boldsymbol{A}\left(\hat{\boldsymbol{\phi}}, \hat{\boldsymbol{p}}^*\right) = \begin{pmatrix} \text{var}(\hat{\phi}) & \text{cov}(\hat{\phi}, \hat{p}^*) \\ & \text{var}(\hat{p}^*) \end{pmatrix}. \tag{C.14}$$

Annexe D

Aspects computationnels

Cette section détaille les aspects computationnels liés aux trois chapitres de cette thèse. Les programmes ayant servi à construire les figures, à mener les études de simulation et à analyser les données sont fournis. Ils ont été produits à l'aide du logiciel R (version 3.5.1). Les données dupliquées, utilisées dans les études de cas, sont accessibles via des liens fournis dans les prochaines lignes.

D.1 Programmes et données du chapitre 2

D.1.1 Programmes

Les programmes R ainsi que les résultats (simulations, analyses des données) sont disponibles en téléchargement gratuit via le lien suivant : https://figshare.com/articles/Capture-Recapture_Methods_for_Data_on_the_Activation_of_Applications_on_Mobile_Phones/6220322.

D.1.2 Données

Les données utilisées pour l'étude de cas sont disponibles en téléchargement gratuit via le lien suivant : https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/H4AOQP.

D.2 Données des chapitres 3 & 4

Les programmes informatiques liés à ces chapitres ne sont pas encore rendus disponibles ; ils le seront une fois les révisions des articles complètes. Les données utilisées dans les études de cas présentées aux Chapitres 3 et 4 sont tirées de l'article de Santostasi *et al.* (2016) ; elles sont disponibles en téléchargement gratuit via le lien suivant : https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0166650#sec025.

Bibliographie

Agresti, A. (**1994**). "Simple capture-recapture models permitting unequal catchability and variable sampling effort", Biometrics **50**, 494–500.

Amstrup, S. C., McDonald, T. L., and Manly, B. F. J. (**2005**). *Handbook of capture-recapture analysis* (Princeton University Press).

Baillargeon, S. and Rivest, L.-P. (**2007**). "The Rcapture package : Loglinear models for capture-recapture in R", Journal of Statistical Software **19**, http://www.jstatsoft.org/v19/i05, Rcapture CRAN URL : http://CRAN.R-project.org/package=Rcapture.

Berrow, S., O'Brien, J., Groth, L., Foley, A., and Voigt, K. (**2012**). "Abundance estimate of bottlenose dolphins (tursiops truncatus) in the lower river shannon candidate special area of conservation, ireland", Aquatic Mammals **38**, 136–144.

Bishop, Y. M. M., Fienberg, S. E., and Holland, P. W. (**1975**). *Discrete Multivariate Analyses : Theory and Practice* (MIT Press).

Bonner, S. J. and Schwarz, C. J. (**2006**). "An extension of the cormack–jolly–seber model for continuous covariates with application to microtus pennsylvanicus", Biometrics **62**, 142–149.

Borchers, D. L. (**2012**). "A non-technical overview of spatially explicit capture-recapture models", Journal of Ornithology **152**, 435–444.

Borchers, D. L., Buckland, S. T., and Zucchini, W. (**2002**). "Estimating animal abundance : closed populations", The Annals of Applied Statistics **3**.

Borchers, D. L. and Efford, M. G. (**2008**). "Spatially explicit maximum lieklihood methods for capture-recapture studies", Biometrics **64**, 377–385.

Burnham, K. P. and Overton, W. S. (**1978**). "Estimation of the size of a closed population when capture probabilities vary among animals (Corr : V68 p345)", Biometrika **65**, 625–634.

Chao, A. (**1987**). "Estimating the population size for capture-recapture data with unequal catchability", Biometrics **43**, 783–791.

Chao, A. (**1989**). "Estimating population size for sparse data in capture-recapture experiments", Biometrics **45**, 427–438.

Chao, A., Lee, S.-M., and Jeng, S.-L. (**1992**). "Estimating population size for capture-recapture data when capture probabilities vary by time and individual animal", Biometrics **48**, 201–216.

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

Cooch, E. G. and White, G. C. (**2018**). *Program MARK : A Gentle Introduction.*

Cormack, R. M. (**1964**). "Estimates of survival from the sighting of marked animals", Biometrika **51**, 429–438.

Cormack, R. M. (**1985**). "Example of the use of glim to analyze capture-recapture studies", in *Lecture Notes in Statistics 29 : Statistics in Ornithology*, edited by B. Morgan and P. North, 242–274 (Springer-Verlag, New York).

Cormack, R. M. (**1989**). "Loglinear models for capture-recapture", Biometrics **45**, 395–413.

Cormack, R. M. (**1992**). "Interval estimation for mark-recapture studies of closed populations (Ack : V49 p315 ; Ref : 91StatMed V10 p717-721)", Biometrics **48**, 567–576.

Cormack, R. M. and Jupp, P. E. (**1991**). "Inference for Poisson and multinomial models for capture-recapture experiments", Biometrika **78**, 911–916.

Coull, B. A. and Agresti, A. (**1999**). "The use of mixed logit models to reflect heterogeneity in capture-recapture studies", Biometrics **55**, 294–301.

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

Darroch, J. N., Fienberg, S. E., Glonek, G. F. V., and Junker, B. W. (**1993**). "A three-sample multiple-recapture approach to census population estimation with heterogeneous catchability", Journal of the American Statistical Association **88**, 1137–1148.

Dorazio, R. M. and Royle, J. A. (**2003**). "Mixture models for estimating the size of a closed population when capture rates vary among individuals", Biometrics **59**, 351–364.

Farcomeni, A. (**2016**). "A general class of recapture models based on the conditional capture probabilities", Biometrics **72**, 116–124.

Fewster, R. and Jupp, P. (**2009**). "Inference on population size in binomial detectability models", Biometrika **96**, 805–820.

Fienberg, S. E. (**1972**). "The multiple recapture census for closed populations and incomplete 2^k contingency tables", Biometrika **59**, 591–603.

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

Hook, E. B. and Regal, R. R. (**1993**). "Effect of variation in probability of ascertainment by sources ("variable catchability") upon "capture-recapture" estimates of prevalence", American Journal of Epidemiology **137**, 1148–1166.

Huggins, R. M. (**1989**). "On the statistical analysis of capture experiments", Biometrika **76**, 130–140.

Huggins, R. M. (**2001**). "A note on the difficulties associated with the analysis of capture-recapture experiments with heterogeneoous capture probabilities", Statistics and Probability Letters **54**, 147–152.

Hwang, W.-H. and R. Huggins, R. M. (**2005**). "An examination of the effect of heterogeneity on the estimation of population size using capture-recapture data", Biometrika **92**, 229–233.

Illian, J., Penttinen, P., Stoyan, H., and Stoyan, D. (**2008**). *Statistical Analysis and Modelling of Spatial Point Patterns* (Statistics in Practice, Wiley, New York).

Ivie, G. W. (**2017**). "Mrc location-based advertising measurement guidelines", URL http://www.mediaratingcouncil.org/MRCLocation-BasedAdvertisingMeasurementGuidelinesFinalMarch2017.pdf.

IWGDMF (International Working Group for Disease Monitoring and Forecasting) (**1995**a). "Capture-recapture and multiple record systems estimation i : history and theoretical development", Am. J. Epidemiol. **142**, 1047–1058.

IWGDMF (International Working Group for Disease Monitoring and Forecasting) (**1995**b). "Capture-recapture and multiple record systems estimation ii : Applications in human diseases", Am. J. Epidemiol. **142**, 1059–1068.

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

Jolly, G. M. (**1982**). "Mark-recapture models with parameters constant in time", Biometrics **38**, 301–321.

Jolly, G. M. and Dickson, J. M. (**1980**). "Mark-recapture suite of programs", Proceedings in Computational Statistics **4**, 570–576.

Kackar, R. N. and Harville, D. A. (**1984**). "Approximations for standard errors of estimators of fixed and random effect in mixedlinear models", Journal of the American Statistical Association **79**, 853–862.

Kendall, W. L. and Bjorkland, R. (**2001**). "Using open robust design models to estimate temporary emigration from capture-recapture data", Biometrics **57**, 1113–1122.

Kendall, W. L., Nichols, J. D., and Hines, J. E. (**1997**). "Estimating temporary emigration using capture-recapture data with Pollock's robust design (Corr : P2248 ; 97V78 p2248)", Ecology **78**, 563–578.

Kendall, W. L., Pollock, K. H., and Brownie, C. (**1995**). "A likelihood-based approach to capture-recapture estimation of demographic parameters under the robust design", Biometrics **51**, 293–308.

King, R., Brooks, S. P., Mazzetta, C., Freeman, S. N., and Morgan, B. J. T. (**2008**). "Identifying and diagnosing population declines : A bayesian assessment of lapwings in the uk", Journal of the Royal Statistical Society, Series C **57**, 607–632.

King, R., Morgan, B. J. T., Gimenez, O., and Brooks, S. P. (**2010**). "Bayesian analysis for population ecology", Chapman and Hall/CRC, Boca Raton, Florida .

Laake, J. L. (**2013**). "Rmark : an r interface for analysis of capture-recapture data with mark", AFSC Processed Rep. 2013-01 1–25.

Lebreton, J.-D., Burnham, K. P., Clobert, J., and Anderson, D. R. (**1992**). "Modeling survival and testing biological hypotheses using marked animals : a unified approach with case studies", Ecological Monographs **62**, 67–118.

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

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

Lindsay, B. G. (**1986**). "Exponential family mixture models (with least-squares estimators)", The Annals of Statistics **14**, 124–137.

Link, W. A. (**2003**). "Nonidentifiability of population size from capture-recapture data with heterogeneous detection probabilities", Biometrics **59**, 1123–1130.

Link, W. A. and Barker, R. J. (**2009**). *Bayesian Inference : with Ecological Applications* (Academic Press, Amsterdam).

McCrea, R. S. and Morgan, B. J. T. (**2015**). *Analysis of Capture-Recapture Data* (Chapman et Hall/CRC).

McCullagh, P. and Nelder, J. A. (**1989**). *Generalized Linear Models*, 2nd edition (Chapman & Hall, London).

McKenzie, E. (**1985**). "Some simple models for discrete variate time series", Water Resources Bulletin **21**, 645–650.

Millar, T., Domingo-Salvany, A., Eastwood, C., and Hay, G. (**2008**). "Glossary of terms relating to capture-recapture methods", Journal of Epidemiology & Community Health **62**, 677–681.

Moran, P. A. P. (**1951**). "A mathematical theory of animal trapping", Biometrika **38**, 307–311.

Nichols, J., Conroy, M., Anderson, D., and Burnham, K. P. (**1984**). "Compensatory mortality in waterfowl populations : A review of the evidence and implications for research and management", Trans. Nth Amer. Wildl. and Natural Resour. Conf **49**, 535–554.

Norris, James L., I. and Pollock, K. H. (**1996**). "Nonparametric MLE under two closed capture-recapture models with heterogeneity", Biometrics **52**, 639–649.

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

Pederson, J. C., Bunnell, K. D., Conner, M. M., and McLaughlin, C. R. (**2012**). "A robust-design analysis to estimate american black bear population parameters in utah", BioOne **23**, 104–116.

Pledger, S. (**2000**). "Unified maximum likelihood estimates for closed capture-recapture models using mixtures", Biometrics **56**, 434–442.

Pollock, K. H. (**1975**). "A k-sample tag-recapture model allowing for unequal survival and catchability.", Biometrika **62**, 577–583.

Pollock, K. H. (**1982**). "A capture-recapture design robust to unequal probability of capture", J. Wildl. Manage. **46**, 752–757.

Pollock, K. H., Nichols, J. D., Brownie, C., and Hines, J. E. (**1990**). *Statistical Inference for Capture-recapture Experiments*, volume 107 of *Wildlife Monographs* (Wildlife Society).

Pollock, K. H. and Otto, M. C. (**1983**). "Robust estimation of population size in closed animal populations from capture-recapture experiments", Biometrics **39**, 1035–1049.

Quenouille, M. (**1949**). "Approximate tests for correlation in time series", Journnal of the Royal Society **11**, 68–84.

Rivest, L.-P. (**2008**). "Why a time effect often has a limited impact on capture-recapture estimates in closed populations", The Canadian Journal of Statistics **36**, 75–84.

Rivest, L.-P. and Baillargeon, S. (**2007**). "Applications and extensions of Chao's moment estimator for the size of a closed population", Biometrics **63**, 999–1006.

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

Rivest, L.-P. and Daigle, G. (**2004**). "Loglinear models for the robust design in mark-recapture experiments", Biometrics **60**, 100–107.

Rivest, L.-P. and Lévesque, T. (**2001**). "Improved loglinear model estimators of abundance in capture-recapture experiments", The Canadian Journal of Statistics **29**, 555–572.

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

Sandland, R. L. and Cormack, R. M. (**1984**). "Statistical inference for Poisson and multinomial models for capture-recapture experiments", Biometrika **71**, 27–33.

Santostasi, N. L., Bonizzoni, S., Bearzi, G., Eddy, L., and Gimenez, O. (**2016**). "A robust design capture-recapture analysis of abundance, survival and temporary emigration of three odontocete species in the gulf of corinth, greece", PLoS ONE **11**, 1–21.

Schnabel, Z. E. (**1938**). "The estimation of the the total fish population in a lake", American Mathematical Monthly **45**, 348–352.

Schwarz, C. J. and Arnason, A. N. (**1996**). "A general methodology for the analysis of capture-recapture experiments in open populations", Biometrics 860–873.

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 edition (Macmillan, New York).

Sekar, C. C. and Deming, W. E. (**1949**). "On a method of estimating birth and death rates and the extent of registration", Journal of the American Statistical Association **44**, 101–115.

Silva, F. J. and Silva JR, J. M. (**2009**). "Circadian and seasonal rhythms in the behavior of spinnerdolphins (stenella longirostris)", Marine Mammal Science **25**, 176–186.

Siquiera, A. C., Quimbayo, J. P., Cantor, M., Silveira, R. B., and Daura-Jorge, F. G. (**2017**). "Estimating population parameters of longsnout seahorses, hippocampus reidi (teleostei : Syngnathidae) through mark-recapture", Neotropical Ichthyology **15**, 1–7.

Smith, A. (**2017**). "Mobile fact sheet", URL http://www.pewinternet.org/fact-sheet/mobile/.

Smith, B. J. (**2015**). "Marketing roi and location data", URL http://www.iab.com/insights/marketing-roi-and-location-data/.

Tyne, J. A., Pollock, K. H., Johnston, D. W., and Bejder, L. (**2014**). "Abundance and survival rates of the hawai'i island associated spinner dolphin (stenella longirostris) stock", PLoS ONE **9**, 1–10.

White, G. and Burnham, K. (**1999**). "Program mark : Survival estimation from populations of marked animals", Bird Study **46**, 120–139.

Yang, H.-C. and Chao, A. (**2005**). "Modeling animals' behavioral response by markov chain models for capture-recapture experiments", Biometrics **61**, 1010–1017.

Yauck, M. and Rivest, L.-P. (**2018**). "On the estimation of population sizes in capture-recapture experiments", Submitted to the Journal of Multivariate Analysis .

Yauck, M., Rivest, L.-P., and Rothman, G. (**2018**). "Capture-recapture methods for data on the activation of applications on mobile phones", Journal of the American Statistical Association URL https://www.tandfonline.com/doi/pdf/10.1080/01621459.2018.1469991.

Zippin, C. (**1956**). "An evaluation of the removal method of estimating animal populations", Biometrics **12**, 163–189.