Development of statistical capture-recapture models in the presence of individual misidentifications

Rémi Fraysse

THÈSE POUR OBTENIR LE GRADE DE DOCTEUR
DE L'UNIVERSITÉ DE MONTPELLIER

En Biostatistiques

École doctorale Information, Structures et Systèmes
Unité de recherche Centre d'écologie fonctionnelle et évolutive – UMR 5175

Développement de modèles statistiques de capture-recapture en présence d'erreurs d'identification individuelle

Présentée par Rémi FRAYSSE
Le 29 février 2024

Sous la direction de Roger PRADEL
et Rémi CHOQUET

Devant le jury composé de

Brett McCLINTOCK, Professeur, NOAA, Alaska Fisheries Science Center	Rapporteur
Nathalie PEYRARD, Professeure, INRAE, MIAT	Rapportrice
Bénédicte FONTEZ, Professeure, Institut Agro (Montpellier), MISTEA	Présidente
Éric PETIT, Directeur de recherche, INRAE, DECOD	Examinateur
Roger PRADEL, Directeur de recherche, CNRS, CEFE	Directeur de thèse
Rémi CHOQUET, Ingénieur de recherche, CNRS, CEFE	Co-encadrant de thèse

UNIVERSITÉ DE MONTPELLIER

The harmony of the world is made manifest in Form and Number, and the heart and soul and all the poetry of Natural Philosophy are embodied in the concept of mathematical beauty.

– D'Arcy Wentworth Thompson

Résumé

La capture-recapture est une méthode largement utilisée pour estimer la taille des populations animales ou des processus démographiques tels que la survie ou la migration. Au cours des dernières années, l'utilisation de marques artificielles pour marquer les animaux a été remplacée par l'utilisation des marques naturelles telles que des motifs visuels ou des empreintes génétiques. Les marques naturelles présentent l'énorme avantage de permettre la "capture" de l'individu de manière non invasive, à l'aide de pièges photographiques ou par la collecte de feces, de poils ou de plumes. Il n'est donc pas nécessaire de voir les individus pour les capturer.

Bien que l'échantillonnage non invasif soit de plus en plus utilisé dans les expériences de capture-recapture, il comporte un risque d'erreur d'identification individuelle qui ne peut être ignoré. Dans de nombreuses situations, les données susceptibles d'erreur d'identification sont tout simplement écartées. La proportion de données ainsi rejetées peut être non négligeable. Pour pallier à ce problème et mieux exploiter les données, des modèles capables de traiter les erreurs d'identification individuelles ont donc été proposés. Toutefois, ces modèles ne tiennent pas compte de plusieurs caractéristiques communes aux données d'échantillonnage non-invasif. Premièrement, la qualité de l'identification n'est évaluée que globalement par les modèles alors qu'une mesure de la qualité de l'identification est souvent disponible au niveau de l'échantillon. Deuxièmement, les modèles ne tiennent pas compte des observations répétées, c'est à dire des différents échantillons appartenant au même individu et obtenus à la même occasion. Troisièmement, la plupart des modèles n'ont été proposés que pour des populations fermées ce qui ne concerne qu'un nombre restreint d'études.

Mon travail aura permis de développer une large gamme de modèles autour du modèle latent multtinomial (LMM) en présence d'erreurs d'identification individuelle. Dans cette thèse, je propose en effet des extensions du LMM qui couvrent les populations fermées et ouvertes, avec un ou plusieurs états et avec ou sans covariable d'identification. En outre, j'ai validé ces extensions par simulation et j'ai appliqué un de ces modèles à des données de loutres.

J'ai implémenté ces extensions au LMM dans le langage R, avec la bibliothèque NIMBLE, dans une approche bayésienne. À l'aide de simulations, j'ai testé les modèles étendus développés et je les ai comparés aux modèles pré-existants appropriés.

Ce travail donne également un exemple du potentiel de modélisation des erreurs d'identification individuelle à travers une étude de simulation d'une expérience de capture-recapture sur des larves de moustiques, dans laquelle l'élimination des échantillons de faible qualité conduirait probablement à ne conserver presque aucun échantillon.

Enfin, la mise en œuvre de ces modèles dans le langage R et sous environnement NIMBLE, très répandu dans le milieu des utilisateurs potentiels, devrait permettre leur adaptation à des cas particuliers et contribuer à leur diffusion dans un contexte plus large.

Abstract

Capture-recapture is a widely used method for estimating population size and inferring demographic processes such as survival or migration rates. In recent years, the use of man-made tags to mark animals has been replaced by natural tags such as visual patterns or genetic fingerprints. Natural marks have the advantage of enabling the individual to be 'captured' non-invasively, using photographic traps or by collecting feces, hairs, or feathers for instance. It is therefore not necessary to see the individuals in order to capture them.

Although non-invasive sampling is increasingly used in capture-recapture experiments, it carries a risk of individual misidentification that cannot be ignored. In most studies, data susceptible to individual misidentification are simply discarded. As a result, the proportion of discarded data may be significant. To overcome this problem, models have been proposed that can deal with individual identification errors in order to use a larger amount of data. However, these models do not take into account several characteristics common to non-invasive sampling data. First, identification quality is only modelled globally, although a measure of identification quality is often available at the sample level. Second, the models do not take into account repeated observations, i.e. different samples belonging to the same individual and collected on the same occasion. Third, most models have only been proposed for closed populations, which only concerns a limited number of studies.

By going through the different models available, I selected one that had the potential to address all these limitations. I implemented the selected model in the R language, specifically the NIMBLE package, in a Bayesian approach, and extended it to overcome the identified limitations. I used simulations to test the performances of the models I had developed and compared them with appropriate pre-existing models.

My work has allowed the development of a complete framework for all basic cases of capture-recapture in the presence of individual misidentification. It covers closed or open populations, in single or multiple states, and with or without an identification quality covariate. This work also provides an example of the potential of modelling misidentification through a simulation study of capture-recapture on mosquito larvae, where discarding the poor quality samples would likely result in almost no sample being retained. Finally, the implementation of the model will make it usable by modellers and should contribute to the dissemination of these new models in a wider context.

Remerciements

Il y a un peu plus de trois ans, je me suis dit "Et si je faisais de l'écologie ?". Je me renseigne un peu et j'apprends vite que même les ingénieurs ont une thèse dans le domaine. Je ne me décourage pas, et dans le mois qui suit je vois passer une offre de thèse au CEFE sur un sujet qui me plaisait bien (beaucoup même). Je n'avais encore jamais entendu parler du CEFE à ce moment mais ça me permettait de rester à Montpellier, ce qui était déjà très bien. La découverte, dès le début, du CEFE et des gens qui s'y trouvaient aura été la touche finale pour bien démarrer la thèse. Ce sont donc trois années d'épanouissement tant professionnel que personnel qui touchent à sa fin. Cette expérience n'aurait pas été la même sans toutes les personnes qui m'ont accompagné et je me dois donc de commencer par les remercier.

Avant toute chose, je veux bien sûr remercier **Brett Mc Clintock**, **Nathalie Peyrard**, **Bénédicte Fontez** et **Éric Petit** d'avoir accepté de relire et d'évaluer mon travail. Ce sont trois années de recherche dont il est bien difficile de savoir à quel point cela peut passionner un autre que soit. J'espère donc qu'à défaut de vous émouvoir, ce travail saura éveiller votre intérêt.
In english, thank you for agreeing to review my work. I hope you find it interesting. And I would like to add a special thank you to Brett for agreeing to get up before 5am for the defence.

Merci **Rémi** et **Roger**, déjà pour m'avoir pris pour faire cette thèse alors même que je n'avais jamais vraiment entendu parler d'écologie. **Roger**, merci pour tes précieux conseils et ta capacité à trouver les erreurs qui ont pu se glisser par ci, par là. **Rémi**, merci aussi pour tes conseils et merci, pour ta présence et ton soutient. Merci à vous deux pour avoir supporté les relectures douloureuses que j'ai pu vous infliger, et merci pour m'avoir permis d'élever la qualité de ce que j'ai pu produire (j'espère).

Thanks **Daniel Turek** for being so quick to react and so efficient in your answers. The help you gave me mainly concerned details and a few bugs that are incomprehensible to ordinary mortals, but without you, some (most ?) codes and models would have never worked as they did.

Je vais évidemment dire un grand merci à **Titouan**. Malgré ton air grognon je sais bien que tu n'es que gentillesse. Je le vois bien à chaque fois que tu veux m'aider à me laver ou à ranger quelque chose. Et comment mentionner **Titouan** sans mentionner ma **Princesse** ? Même si tu es bien plus indépendante, et que je te sens parfois un peu intéressée, un grand merci à toi d'être un rayon de lumière dans chacune de mes journées. Vous êtes tous les deux un support émotionnel exceptionnel, et ce sans rien faire de plus que d'exister. Merci le sang.

J'ai dit que la découverte du CEFE et des gens avait été la touche finale pour bien démarrer. Du côté du CEFE je vais commencer par remercier l'ancien bureau 202 qui m'a accueilli, **Valentin**, **Okasana**, **Maud** et **Gilles**, merci pour la bonne ambiance qui a régné. **Jan Perret**, **Jan Perret**, **Jan Perret**, hâte de refaire une fondue et de discuter de comment faire de la CMR propre sur des plantes. **Marwan**, hâte de me re-promener avec toi en manif, et n'hésite pas à continuer à poser des questions, même si je ne pense pas pouvoir accélérer plus que ça tes modèles (remember calculate=F). **Maëlis**, ravi qu'on soit allé voir des lions en Afrique du Sud. **Mellina**, merci d'avoir partagé le bureau avec moi, j'aurai été bien seul dedans vers la fin en l'absence de Gilles. Aussi, **Killian**, **Tom**, **Lise**, **Soumaya**, **Adé**, **Coline**, **Théo**, **Thibault**, **Déborah**, **Javi**, **Cyrielle**, **Thierry**, **Simon**, **Patricia**, **Sarah**, **Sarah**, **Nicolas** et ceux qui sont passés dans l'équipe et ceux que j'ai pu oublier (désolé, vraiment), merci à toute l'équipe de me faire dire "zut je vais pas voir les gens" quand il pleut et que je reste en télétravail. Et merci à **Olivier** et **Aurélien** de faire vivre cette équipe comme ça et toujours chercher à améliorer ce qui peut l'être.
Merci aussi au comité d'animation du CEFE qui a organisé les apéros réguliers, les journées des doctorants, les barbecues au TE et les week-ends de cohésion.

Lucas, **Louise** et **Lou** (ou **Lou** et **Louise**?), merci pour ces belles après-midi et soirées passées ensemble. Je retiens particulièrement la décoration, chaque année, du sapin de noël tous ensemble et les parties de Terraforming Mars ou Wonder Woods.

Merci à la **team pipistrelle**, **Lou(ise?)**, **Nadège**, **Romain**, **Guilhem** et **Léa**. Ça a été de superbes week-ends les uns chez les autres, de superbes découvertes musicales (Prince Ali chanté par Johnny ???) et de superbes dégustations de pâtés, de génépi et rhums arrangés. Hâte de la prochaine fois !

Merci aux ITX, famille d'adoption de l'agro qui montre si bien qu'on peut être à la fois intolérant et super sympa. D'abord, **Lou(ise?)** et **Nadège** les mamans de la famille, sans vous bien peu d'événements s'organiseraient, merci pour tous ces week-ends de folie qui nous ont regroupé. **Valentin**, **Gaëlle**, **Pierre**, **Léa** et **Cyril**, **Alban** et **Aymerick** merci pour tous ces bons moments passés et à venir, pour les randonnées à écouter les histoires de Pierre sur le paysage, pour les pâtés et épaules de sanglier à 1h du matin. Merci aux **viti-oenos**, qui, dans une tradition française bien retrouvée, nous abreuvent de (bon) vin comme si c'était de l'eau. Enfin, **Alban**, merci pour les invitations dans ta ferme-château fortifiée !

Merci à tous les copains du Tarn, **Félicien**, **Nicolas**, **Joël**, **Lucas**, **Sébastien**, **Dorian** et **Antoine**. Je ne vous vois pas si souvent mais c'est toujours beaucoup de plaisir et de rires.

Et comment ne pas remercier tous ceux qui me soutiennent dans mon addiction au jeu ?
D'abord ceux de la troupe de spectacle de Kob et ses Amis, aussi à la recherche de la Shattered Star (l'étoile fragmentée ?) et d'un caillou volé en Outreterre, **Guilhem**, **Antoine**, **Anton** et feu **Kévin**, merci à vous pour toutes ces sessions et histoires alternant épique, drama et wholesome dans le rire et la bonne humeure. **Antoine**, **Anton** et **Ketbi**, promis je jouerai un personnage moins chaotique mauvais pour remplacer Svalnäs. Et enfin j'espère que les aventures des Fiers de bières qui ont des haches avec **Mumu**, **Val** et **Joël** continueront encore longtemps, on ne change pas une équipe qui gagne !
Ensuite, je remercie la team jdr Tarn & Dadou, **Félicien**, **Nicolas**, **Lucas** et **Joël**. J'espère que les aventures de Nathaniel vous plaisent et j'avoue avoir un peu hâte de voir comment se déroulera la fin. Et même si j'ai parfois l'impression que vous faites juste n'importe quoi, c'est peut-être ça le mieux finalement.
Ensuite merci à tous ceux qui ont participé aux trois éditions des jeux de rôle **grandeur nature** qui ont déjà été organisés. Ce sont vraiment des expériences uniques et des souvenirs incroyables (sauf le Suzard, ça c'est vraiment dégueulasse).
Mumu, **Val** et **Thomas**, merci pour cette campagne Gloomhaven qu'on aimerait parfois continuer plus longtemps malgé l'heure. Méfiez-vous quand même de pas traiter Stal Rim comme un stagiaire trop fort. Et merci **Anton** d'avoir commencé la campagne avec nous.
Enfin merci à la team tarot, **Laurine**, **Jérémy**, **Lisa**, **Wakinian**, **Théo**, **Pablo**, **Nicolas** et **Nicolas**. Ça a eu le don fou de me faire oublier que j'étais en train de rédiger cette thèse pendant une demi-heure tous les jours et ça, ça n'a pas de prix. Mention spéciale à l'Hydre, on peut lui couper un roi mais il en repoussera bien deux.

Quelques remerciements en vrac, d'abord à **SciHub**, sans qui la science irait moins vite, ensuite à tous ceux qui ont conçu le template LaTeX **Kaobook** que j'ai utilisé pour cette thèse (il est bô ce template hein ?) et enfin à la chouette maléfique **Duo**, qui m'a permis de découvrir le chinois sans même menacer ma famille quand je me rappelais à 23h que je n'avais pas fait de leçon de la journée.

Avant de finir, un grand merci à ma famille. **Papa**, **maman**, grâce vous et à l'éducation que vous m'avez donnée, le fils de deux ingénieurs agronomes que je suis a pu lui-même devenir ingénieur agronome (quelle belle reproduction sociale), puis se dépasser et arriver dans le monde académique. Merci pour votre soutien sans faille et l'ouverture d'esprit que vous m'avez apporté. **Julie** et **Christian** merci, les vacances qu'on passe tous ensemble sont d'excellents souvenir parce que vous êtes là. Merci aussi **Brigitte** de répondre toujours présente quand je viens vers toi ! Et merci à la belle-famille **Anne**, **Olivier** et **Léna**.

Et pour finir, merci à toi **Lou** d'avoir fait partie de ma vie depuis déjà 5 ans et pour encore bien des décennies. Merci pour le support que tu m'as apporté. Je n'ai exprimé que peu de besoin de soutient mais c'est bien parce que tu étais toujours là. Merci d'être là et d'être toi.

Merci à tous !

Contents

List of Figures

List of Tables

Introduction

Introduction

Population dynamics

The importance of population dynamics

Worldwide, biodiversity is in decline. Animal populations face many threats due to human activities that affect their environment, or even impact the population directly [1–3]. For example, overfishing directly affects some species, but also disturbs the marine ecosystem equilibrium [4]. Deforestation directly results in habitat reduction [5] and contributes to the propagation of vector-borne diseases by creating new habitats for vectors of these diseases [6]. Pollution, by making the environment toxic or modifying physico-chemical variables such as acidity, can be responsible for higher mortality [7]. Moreover, some species that have been introduced (voluntarily or not) can become invasive and have economic [8] and ecological [9] impacts. Reintroduced or protected species can be involved in conflicts with humans when they share territory, such as the wolf in France [10] or the elephant in Africa [11].

For these reasons, conservation management programmes are developed. These programmes must be based on scientific knowledge determined by both qualitative and quantitative data [12]. Ecological knowledge allows us to understand how ecosystems function and the mechanisms responsible for biodiversity loss [13, 14]. Population dynamics in particular is an active area of research in which statistical modelling tries to answer questions such as 'How many individuals are there in a population?', 'Where do they live?', 'What is their survival rate?' and many others. The need for modelling arises from the impossibility of directly observing the elements that would answer these questions. For populations in the wild, it is often impossible to carry out exhaustive monitoring of individuals in a specific area. (Even plants in a 1m x 1m square cannot be counted easily [15].) The failure to account for detectability issues then leads to inaccurate results, such as underestimating the population size [16] or the distribution area of the studied species [17]. In general, the imperfect surveying of individuals is one of the main difficulties that must be accounted for in ecological data [18, 19]. Methods have been developed to try to address this issue. For instance, capture-recapture (CR) models are powerful methods that can estimate population size, survival rate, state transition between physiological states, or migration rates, while taking into account imperfect detection [20].

The basics of capture-recapture

The concept of capture-recapture is quite simple – even intuitive. Consider a population in a specific home range where it is assumed that no births, deaths, immigration or emigration occur. This population remains the same from the beginning to the end of the experiment and is qualified as a **closed population**. We want to carry out longitudinal monitoring of

[1]: Dirzo et al. (2014), 'Defaunation in the Anthropocene'
[2]: McCauley et al. (2015), 'Marine defaunation'
[3]: Payne et al. (2016), 'Ecological selectivity of the emerging mass extinction in the oceans'
[4]: Jackson et al. (2001), 'Historical overfishing and the recent collapse of coastal ecosystems'
[5]: Giam (2017), 'Global biodiversity loss from tropical deforestation'
[6]: Walsh et al. (1993), 'Deforestation'
[7]: McNeely (1992), 'The sinking ark'
[8]: Lovell et al. (2006), 'The Economic Impacts of Aquatic Invasive Species'
[9]: Gallardo et al. (2016), 'Global ecological impacts of invasive species in aquatic ecosystems'
[10]: Grente et al. (2022), 'Wolf depredation hotspots in France'
[11]: Shaffer et al. (2019), 'Human-Elephant Conflict'
[12]: Sutherland et al. (2004), 'The need for evidence-based conservation'
[13]: Moussy et al. (2022), 'A quantitative global review of species population monitoring'
[14]: Nichols et al. (2006), 'Monitoring for conservation'
[15]: Perret et al. (2023), 'Plants stand still but hide'
[16]: Cubaynes et al. (2010), 'Importance of Accounting for Detection Heterogeneity When Estimating Abundance'
[17]: Comte et al. (2013), 'Species distribution modelling and imperfect detection'
[18]: Yoccoz et al. (2001), 'Monitoring of biological diversity in space and time'
[19]: Guillera-Arroita et al. (2010), 'Design of occupancy studies with imperfect detection'
[20]: Williams et al. (2002), *Analysis and Management of Animal Populations*

this population at regular time intervals called occasions (at least two) to estimate the size N of the population. On a first occasion, M individuals are captured, tagged and released. On a second occasion, n individuals are captured from the whole population. Of these n individuals, m were already marked on the first occasion. Assuming that all individuals have the same probability of being captured, we expect that the proportion of marked individuals in the second sample (m/n) is approximately the proportion of marked individuals in the whole population (M/N):

$$\frac{m}{n} \approx \frac{M}{N}. \tag{0.1}$$

This concept can be traced back to the 16th century, but the detailed explanation is attributed to Laplace (1786) [21]. He proposed estimating the population N of France using birth registers as the marked population M and major cities of known population size as the sample n, in which m are births from these major cities.

[21]: Laplace (1786), 'Sur les naissances, les mariages et les morts'

Marking animals originated with Petersen (1894) [22], who marked fish to estimate mortality rate. Although he did not estimate population size, the estimator derived from Equation 0.1, $N \approx n \times M/m$ took his name, along with Lincoln's – who used it to estimate the size of a duck population in America [23]. It is known as the Lincoln-Petersen estimator.

[22]: Petersen (1986), 'The yearly immigration of young plaice into the Limfjord from the German Sea, ect'

[23]: Lincoln (1930), 'Calculating waterfowl abundance on the basis of banding returns'

Later, the capture-recapture method was extended for closed populations to multiple capture-recapture. Models were developed to account for more than two occasions [24–26]. Compared to two-occasion experiments, this allowed more complex cases to be treated. These models allow the capture probability to vary due to time, individual heterogeneity, behavioural response, or even several of these combined.

[24]: Chapman (1952), 'Inverse, Multiple and Sequential Sample Censuses'
[25]: Darroch (1958), 'The Multiple-Recapture Census'
[26]: Otis et al. (1978), 'Statistical inference from capture data on closed animal populations'

In 1964, Cormack used multiple capture-recapture to estimate the survival rate in an **open population** (i.e. individuals may enter or leave the population of interest through birth, death, immigration or emigration) [27]. In the following year, Jolly and Seber simultaneously published two papers to 'complete' Cormack's model [28, 29]. These three papers led to the model now called Cormack-Jolly-Seber (CJS) and several extensions of this.

[27]: Cormack (1964), 'Estimates of Survival from the Sighting of Marked Animals'
[28]: Jolly (1965), 'Explicit estimates from capture-recapture data with both death and immigration-stochastic model'
[29]: Seber (1965), 'A note on the multiple-recapture census'

To estimate survival in several areas occupied by a population subject to migration from one area to another, Arnason published two papers in 1972 and 1973 [30, 31]. With the later paper of Schwarz in 1993 [32], the model came to be known as the Arnason-Schwarz model. Its focus is both the survival rate and the migration rate (or the transition rate between different states).

[30]: Neil Arnason (1972), 'Parameter estimates from mark-recapture experiments on two populations subject to migration and death'
[31]: Neil Arnason (1973), 'The estimation of population size, migration rates and survival in a stratified population'
[32]: Schwarz et al. (1993), 'Estimating migration rates using tag-recovery data'

Bayesian statistics

CR model parameters are either estimated using frequentist or Bayesian inference. In our case, Bayesian inference is required. In Bayesian inference, parameters are considered random variables. This does not refer to the variability of the parameters (as they are typically fixed unknown quantities), but to uncertainty about their true value. As such, Bayesian inference estimates the density of probability of the value that a parameter can take. This density is calculated accounting for prior knowledge about the parameter that is not represented in the likelihood for the data.

This contrasts with frequentist inference, which aims to estimate the most likely value of the parameters conditional on the data. Another interest of Bayesian inference is that it can be applied to complex models, with the downside that the computations can be time-consuming.

Let us assume that we observe data X that gives information about a certain parameter θ through a model with likelihood $\mathcal{L}(\theta|X)$. The Bayesian paradigm aims to estimate the posterior probability of θ, denoted by $\pi(\theta|X)$. Using Bayes' rule , the posterior density is computed as:

$$\pi(\theta|X) = \frac{\mathcal{L}(\theta|X)\pi(\theta)}{\int \mathcal{L}(\theta|X)\pi(\theta)}$$

where $\pi(\theta)$ is the density of the parameter distribution before the data is used to inform it. It is called the prior distribution. The denominator (called predictive probability) is unique because the whole fraction should integrate to 1. Therefore, we often ignore it and write the posterior probability as being proportional to the likelihood and the prior.

$$\pi(\theta|X) \propto \mathcal{L}(\theta|X)\pi(\theta)$$

When choosing a model, two key questions are what are the parameters θ and what is the likelihood. Contrary to frequentist statistics, before estimating the parameters, Bayesian statistics also needs to define the prior $\pi(\theta)$. This is a complicated matter that has no definite answer. The prior distribution can be chosen to integrate a priori information about the parameters, either obtained from previous experiments or elicited from subjective estimation. In these cases, we talk about an informative prior. If the prior does not contain information, and is uniform over the parameter space, it is called non-informative or uninformative. Historically, mainly *conjugate priors* were used because they lead to known posterior distribution families. For example, if $X \sim Bin(N, p)$ and N is known, and if we choose a beta prior over p, $p \sim \beta(a, b)$, then the posterior distribution of p is also a beta distribution $p|x \sim \beta(a + x, b + N - x)$. Since powerful methods such as Monte Carlo Markov Chains (MCMC) have become more accessible, other common priors are used. The choice of prior is not to be taken lightly. Some uninformative priors can be '*improper*' if they do not integrate to 1, like an unbounded uniform distribution. It is important to verify that the resulting posterior distribution is proper. Another point to consider is that a non-informative prior on a parameter may not be uninformative on a function of this parameter. For example, let α be a probability parameter modelled as a probit function of a parameter a. Then an uninformative prior on a (e.g. a normal distribution with a large standard error) will result in a prior on α, indicating that either $\alpha = 0$ or $\alpha = 1$.

When the selected priors do not lead to a known distribution, computing summary statistics for the posterior distribution may prove difficult. One solution is to sample from the posterior distribution several times, then use the samples to approximate the posterior distribution. The most common method to sample from complex distributions is MCMC. An MCMC algorithm is constructed by defining the transition probabilities of the Markov chain in a way that the distribution in which we sample at the n^{th} iteration converges to the distribution of interest when n increases, independently of the initial value of the chain. Various algorithms can be

Likelihood

Let X be a random variable with density (or mass) function f depending on the parameter θ.

$$x \mapsto f(x|\theta)$$

where x is a realisation of X. The likelihood function is f when it is viewed as a function of θ with x fixed:

$$\theta \mapsto f(\theta|x)$$

Noted $\mathcal{L}(\theta|X)$

Bayes' formula

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Markov chain

Markov chains are stochastic models describing a sequence of random events in which the probability of the outcome of each event depends only on the previous outcome. Say we have the sequence of random variable $X_1, ...X_n$. The probability that $X_i = x \forall i = 2, ..., n$ only depends on the value that took X_{i-1}: $P(X_i = x|X_1, ..., X_{i-1}) = P(X_i = x|X_{i-1})$. $P(X_i = x|X_{i-1})$ is the transition probability of the Markov chain.

used to construct an MCMC, but we will use two specific algorithms: *Gibbs sampling* and the Metropolis-Hastings (MH) algorithm. Gibbs sampling consists of sampling successively from the posterior distribution of each parameter conditionally on the others. It is used when the conditional posterior is a known distribution from which it is easy to sample. For example, assuming that we have a model (for data X) with two parameters c and d, and we assign the following conjugate priors: $c \sim \beta(a_0, b_0)$ and $d \sim \mathcal{N}(\mu_0, \sigma_0)$, leading to full conditional posterior beta and normal distributions. Although we cannot jointly sample c and d, it is easy to sample them successively from their conditional posterior distributions: $c|(X, d) \sim \beta(a, b)$ and $d|(X, c) \sim \mathcal{N}(\mu, \sigma)$. More details about Gibbs sampling can be found in various references such as Tierney (1994) [33] and Gelfand (2000) [34]. Metropolis-Hastings is used when the conditional posterior distribution cannot be sampled easily. As with Gibbs sampling, the parameter is updated conditionally on the current value of all other parameters, and possibly on its current value. First, a proposal θ' for a new parameter value is drawn by sampling in a given proposal distribution $q(\theta)$. Then this proposal is accepted with probability

$$r = min\left(1, \frac{\pi(\theta'|X)q(\theta|\theta')}{\pi(\theta|X)q(\theta'|\theta)}\right) \quad .$$

If the proposal θ' is rejected, then the previous value θ is kept and the algorithm proceeds with sampling the following parameters. More details about the MH algorithm can be found in several books and reviews (see [35–37]). When using MH, the proposed value is often dependent on the current value, and there is high auto-correlation in the chains. The consequence is that many iterations may be needed to reach the point where the distribution in which we are sampling has converged to the distribution of interest. Thus, many more iterations are necessary to approximate the distribution of interest.

Assessing the convergence of the chains can be done in various ways. By running several chains with different starting points and plotting them together, it can be seen if the chains converge to the same distribution and if they mix well with each other. Various convergence statistics can also be used. We will use the convergence diagnostic from Gelman, Rubin and Brooks (GRB, [38, 39]). The GRB also consists of running several chains with different starting points. Then it compares the variance within and between chains. If there is no difference, the computed $\hat{R}$ statistic should be 1. The authors recommend a threshold of 1.1, with lower values indicating convergence. Another useful metric is the effective sample size. Since the samples are not independent, the metric estimates how many independent samples they are worth. In the case of complete independence, the effective sample size is the same as the sample size. High auto-correlation will reduce the effective sample size, making it smaller than the sample size. For example, a sample size of 5000 iterations with high auto-correlation could be worth 100 independent samples. So the effective sample size would be 100.

[33]: Tierney (1994), 'Markov Chains for Exploring Posterior Distributions'

[34]: Gelfand (2000), 'Gibbs Sampling'

[35]: Chib et al. (1995), 'Understanding the Metropolis-Hastings Algorithm'

[36]: Chen et al. (2000), *Monte Carlo Methods in Bayesian Computation*

[37]: Gelman et al. (2015), *Bayesian Data Analysis*

[38]: Gelman et al. (1992), 'Inference from Iterative Simulation Using Multiple Sequences'

[39]: Brooks et al. (1998), 'General Methods for Monitoring Convergence of Iterative Simulations'

Tags for CR

Types of tags

Several kinds of methods can be used for marking animals, which can be referred to as 'tags'. Two categories of tags can be distinguished: *man-made tags* and *natural tags*. There are many kinds of man-made tags. They can be simple externally mounted tags such as rings for birds. Different materials can be used, including plastic, aluminium or stainless steel. When these tags are used for individual identification, they are engraved or marked with a unique identifier. These tags may require the physical recapture of an individual or its visual observation in order to read the tags. This observing event is called 'recapture'. Some tags can also be used remotely, such as radiotelemetry tags. These offer the advantage of being detected without the need to actually see the individual and its tag. They can be applied externally (with collars, for example) or internally (mainly for mammals and fish). Weight, colour, material, method of attachment are among many variables to consider when choosing a tag, because they may affect the individual. For example, heavy tags will make it hard for birds to fly, aluminium rings are known to harm the legs of some bird species [40], and band colour can influence bird behaviour ([41]). For a more extended review on tags and their impact, see Murray & Fuller (2000) [42]. Using man-made tags has many advantages and allows a list of tagged individuals. However, they all require the physical capture of the animal at least once to apply the tag. This can be a problem when studying elusive or difficult-to-catch species. Handling the animal can also cause stress or harm.

In contrast, natural tags are existing features of animals that can be used to identify them. In this category, two types of tags can be distinguished: *photographic tags* and *genetic tags*. Photographic tags are distinctive marks or patterns on visible parts of the animal. These marks can be present at birth, like colour spots on leopards ([43]) or patterns on beetles (*Nicrophorus orbicollis* and *Nicrophorus americanus* [44]). Photographic tags can also be acquired, such as scars on dolphins (*Tursiops truncatus* [45]) or whales (*Ziphius cavirostris* [46]). For terrestrial animals, photographic tags have the advantage that observers are not needed to spot the individuals in the field. Camera traps can be installed that automatically take pictures of an individual triggering the movement captor. For marine animals, a picture removes the need to get too close to the individual. Recognising the individuals from photographs can be done manually, but with the development of deep learning, algorithms can now match individuals from photographs with good reliability [47].

Genetic tags, generally known as DNA fingerprints, allow the identification of individuals through the genotyping of individuals. Historically done through blood samples collected from individuals, the improvement of genotyping technology now allows much lower quality samples to be used. Today, many studies focus on DNA collected through non-invasive genetic sampling. The DNA left by an animal, in scat or hair, can be collected without having to catch or disturb the animal [48, 49]. Studies using such DNA have been carried out on bears ([50]), bobcats ([51, 52]), pronghorns ([53]), and elephants ([54]), for example. Environmental samples such as scat are very convenient to sample. However, the sample

[40]: Meyers (1994), 'Leg bands cause injuries to parakeets and parrots'

[41]: Burley (1985), 'Leg-band color and mortality patterns in captive breeding populations of Zebra finches'

[42]: Murray et al. (2000), 'A critical review of the effects of marking on the biology of vertebrates'

[43]: Swanepoel et al. (2015), 'Density of leopards Panthera pardus on protected and non-protected land in the Waterberg Biosphere, South Africa'

[44]: Quinby et al. (2021), 'Estimating population abundance of burying beetles using photo-identification and mark-recapture methods'

[45]: Labach et al. (2022), 'Distribution and abundance of common bottlenose dolphin (Tursiops truncatus) over the French Mediterranean continental shelf'

[46]: Curtis et al. (2021), 'Abundance, survival, and annual rate of change of Cuvier's beaked whales (Ziphius cavirostris) on a Navy sonar range'

[47]: Crall et al. (2013), 'HotSpotter - Patterned species instance recognition'

[48]: Taberlet et al. (1999), 'Non-invasive genetic sampling and individual identification'

[49]: Taberlet et al. (1999), 'Noninvasive genetic sampling'

[50]: Dreher et al. (2007), 'Noninvasive estimation of black bear abundance incorporating genotyping errors and harvested bear'

[51]: Ruell et al. (2009), 'Estimating bobcat population sizes and densities in a fragmented urban landscape using noninvasive capture–recapture sampling'

[52]: Morin et al. (2018), 'Efficient single-survey estimation of carnivore density using fecal DNA and spatial capture-recapture'

[53]: Woodruff et al. (2016), 'Estimating Sonoran pronghorn abundance and survival with fecal DNA and capture—recapture methods'

[54]: Laguardia et al. (2021), 'Nationwide abundance and distribution of African forest elephants across Gabon using non-invasive SNP genotyping'

must not be too degraded to allow the DNA to be replicated properly for a correct identification of the genotype. In addition, a panel of markers must be developed that is complex enough to differentiate between individuals. Taberlet & Luikart (1999) [48] give the probability PI that two individuals in a population have the same genotype (i.e. probability of identity) for given allele frequencies:

[48]: Taberlet et al. (1999), 'Non-invasive genetic sampling and individual identification'

$$PI = \sum p_i^4 + \sum (2p_i p_j)^2$$

where p_i and p_j are the frequencies of the i^{th} and j^{th} alleles. PI is the probability of identity. A PI_{sib} exists that is a corrected statistic taking into account the substructure of a population in which many siblings can be found. The genetic markers are mainly microsatellites and single nucleotid polymorphisms (SNPs).

Microsatellites and SNPs for genetic tagging

Historically, the genetic markers used to distinguish individuals were microsatellites. Microsatellites are repetitive sequences of nucleotides in the DNA, such as 'ATATATAT'. The number of repetitions of the fragment (here 'AT') varies between individuals, leading to multiple alleles being available at a single locus within a population. The high number of alleles available makes microsatellites useful for differentiating between individuals; usually around 10 to 20 loci are needed to identify individuals.

A recent and noteworthy shift in the field of markers is occurring as microsatellites are being replaced by SNP markers. SNPs represent genetic variations at a specific locus, with the theoretical possibility of up to four alleles, although in most cases, the majority are bi-allelic due to low mutation rates. These markers have many advantages compared to microsatellites. They change in a way that is well described by simple mutation models such as the infinite site model [55]. Since they only present two alleles, the data produced is very easy to standardise, independently of the laboratory or the methods used. In addition, false alleles are easily detected since only two are expected. Lastly, the sequences of interest are very short (around 50–70 base pairs) compared to microsatellites (around 80–300 base pairs). The shorter length makes it easier to amplify when using degraded DNA [56]. This is important when using low-quality DNA. However, being bi-allelic, SNPs have a much reduced identifying power compared to microsatellites. Many more SNPs than microsatellites are required to obtain the same power: around two to six times more [57]. Using SNPs, around 50 to 100 loci are required to differentiate between individuals, whereas only 10 loci are needed for microsatellites.

[55]: Vignal et al. (2002), 'A review on SNP and other types of molecular markers and their use in animal genetics'

[56]: Morin et al. (2004), 'SNPs in ecology, evolution and conservation'

[57]: Schopen et al. (2008), 'Comparison of information content for microsatellites and SNPs in poultry and cattle'

Assumptions related to tags

Traditionally, several assumptions have been made about tags so that the data can be used by a conventional CR model. These assumptions are not always stated explicitly. Conventional CR models assume that:

1. All the individuals in the population of interest can be tagged.
2. The tagging method does not affect survival.

3. A tag is unique to the individual.
4. The tag is permanent, i.e. cannot be lost or modified.
5. **The tag allows the individual to be identified.**

These assumptions are necessary to ensure unbiased estimates, along with the other assumptions made when using conventional capture-recapture models. The first hypothesis is usually met in the sense that the population of interest is made up of the markable individuals. This can often exclude young individuals, either because they are too small for man-made tags or because they do not have natural identification features such as scars yet. 'What is the population of interest?' is a crucial question related to this hypothesis.

The second hypothesis can be tested for man-made tags, and there is no reason for natural tags to affect survival.

The third hypothesis is easily met with man-made tags. It should also be met with photographic tags. To ensure that genetic tags are unique, enough markers must be used to lower the probability that several individuals have the same genotype by chance (PI) under a given threshold.

Although man-made tags such as rings can be lost, this is unlikely, and an individual's genotype will not change. However, photographic tags based on scars could change. New scars can appear on top of the old ones and potentially change the pattern. If the pattern changes slowly enough compared to the frequency of recapture, this may not be a problem, since the individual would be recognised. However, if the pattern changes drastically between two captures, then assumption four cannot be made. This may happen either because of a specific event that led to larger scars hiding the older marks, or because small modifications accumulated to change the pattern drastically between two occasions far apart in time. Birth patterns could potentially also be hidden by scarring.

Individual identification

The last hypothesis is the subject of particular attention in this thesis. It is assumed that, with genetic tags, if the genotype of an individual is fully observed without error, then the individual is identified without error. In the same way, for photographic tags, if a photograph is taken in a way that fully and clearly shows the pattern, then it allows the identification of the individual without error.

But in some cases, a photograph is blurry or does not cover the entirety of the pattern, so the individual might not be correctly recognised. Genetic tags have similar problems. One prevalent issue is *allelic dropout*, in which the PCR fails to amplify one of the alleles in a heterozygous individual [58], resulting in the incorrect inference of homozygosity. Furthermore, allelic dropout can result in missing data when both alleles fail to amplify. Another source of error is the creation of *false alleles* during PCR amplification. These artificial alleles can lead to erroneous genotype calls, misidentifying homozygous individuals as heterozygous.

In order to mitigate these errors, PCR replicates are carried out [48, 59]. One sample is amplified in multiple independent PCRs, and a genotype is called based on the consensus of several replicates, reducing the impact of allelic dropout and false alleles. Yet even with replicates, missing

[58]: Taberlet et al. (1996), 'Reliable genotyping of samples with very low DNA quantities using PCR'

[48]: Taberlet et al. (1999), 'Non-invasive genetic sampling and individual identification'
[59]: Navidi et al. (1992), 'A multiple-tubes approach for accurate genotyping of very small DNA samples by using PCR'

data can still occur if too many replicates fail to amplify a locus or if a robust consensus cannot be reached. In these cases, the probability of identifying an individual decreases, so these issues should not be neglected [49, 60].

[49]: Taberlet et al. (1999), 'Noninvasive genetic sampling'
[60]: Waits et al. (2005), 'Noninvasive Genetic Sampling Tools for Wildlife Biologists'

Another problem that can arise with photographic tags is that the markings of many species are on their sides. If an individual is never photographed from both sides simultaneously, there is no reliable way of matching the marks from both sides, and these marks cannot be recognised as belonging to the same individual. This specific problem has been tackled in several papers [61, 62], and a method and a package have been developed to analyse such data [63].

[61]: Bonner et al. (2013), 'Mark-recapture with multiple, non-invasive marks'
[62]: McClintock et al. (2013), 'Integrated modeling of bilateral photo-identification data in mark–recapture analyses'
[63]: McClintock (2015), 'multimark'

The following section focuses on the problem of individual misidentification, describing what this consists of, the consequences of ignoring it, and the different ways of dealing with it.

The challenge of misidentification in CR studies

Misidentification

Misidentifying an individual is not recognising it when it is captured, thus not assigning the capture to the true individual. Three different cases of misidentification can be distinguished, but this thesis only considers the third:

1. misidentification due to evolving marks
2. misidentification in which two real individuals are confused with each other
3. misidentification in which an individual that does not exist is identified.

1) In contrast to man-made tags, natural tags cannot be matched to a known bank of tags. Thus, when an individual is misidentified because its tag has changed, a new individual is 'identified' that does not exist in the population, and subsequent captures will be identified as this new individual (or another if the tag changes again) [64]. These false individuals that are created in the dataset but which do not actually exist are referred to as 'ghosts'. This problem is discussed at the end of this thesis.

[64]: Yoshizaki et al. (2009), 'Modeling misidentification errors in capture–recapture studies using photographic identification of evolving marks'

2) If the photographs do not show the full pattern clearly or if the genotype is not observed perfectly (missing loci or genotyping error), then the identification is uncertain. It may happen that the marks are so unclear that a sample cannot be identified as belonging to an existing individual. This would occur if the percentage of loci available for the identification of a sample is too low. Then the PI corresponding to the available loci would be high, meaning that several individuals could likely share the same genotype for these loci. We can assume that this case will not happen, as such samples represent the lowest quality data and can be filtered out beforehand.

3) The data quality may be high enough to not confuse the individuals, but too low to ensure that the individual is recognised. This would lead to a sample being misidentified as another non-existing individual,

creating a ghost. It is similar to the changing tags problem, but the original individual can still be correctly identified on later occasions. The below shows an example for an individual captured at occasions 1 and 3, but misidentified (as a ghost) on one of these occasions:

True history	Observed histories
101	100
	001

Ignoring misidentification may have serious consequences on estimating parameters with traditional CR models. Creel et al. (2003)[65] showed that this can result in the population size being greatly overestimated. In addition, Winiarski et al. (2016) [66] found that survival probability can be underestimated by the models. The intuitive explanation is that misidentification adds a lot of unobserved events in histories, thus lowering the capture rate estimates, with an impact on the other parameters of the model such as survival and population size. The conventional models have multinomial distributions, with the outcomes being the possible output histories. They cannot be modified to accommodate misidentification and still maintain the multinomial structure [67].

[65]: Creel et al. (2003), 'Population size estimation in Yellowstone wolves with error-prone noninvasive microsatellite genotypes'

[66]: Winiarski et al. (2016), 'Effects of photo and genotype-based misidentification error on estimates of survival, detection and state transition using multistate survival models'

[67]: Yoshizaki et al. (2011), 'Modeling misidentification errors that result from use of genetic tags in capture–recapture studies'

This thesis focuses on the third kind of misidentification, and it also assumes that misidentifications are unique. That is, each misidentification will produce a different ghost. This hypothesis is more realistic with SNPs as genetic markers than with microsatellites, where only a few loci are examined.

Misidentification hypotheses

This thesis considers the following assumptions about misidentification:

1. A real individual cannot be misidentified as another real individual.
2. Misidentification always results in the creation of a new individual or 'ghost'.
3. Misidentifications are unique, and two misidentifications cannot be matched to the same ghost.

To reduce misidentification, several studies have proposed solutions when using genetic tags. These solutions range from field methods and improved laboratory techniques for genetic analysis [60, 68] to pre-analysis software that helps filter out data likely to contain errors [69]. Regarding visual pattern recognition, computer-aided image matching techniques [47, 70] have been developed to aid identification. However, improved genetic methods often come at an increased cost and workload because replicates must be carried out. In addition, whether with genetic tags or computer-assisted methods, when using low-quality samples, there will still be a risk of misidentification.

[60]: Waits et al. (2005), 'Noninvasive Genetic Sampling Tools for Wildlife Biologists'

[68]: Paetkau (2003), 'An empirical exploration of data quality in DNA-based population inventories'

[69]: McKelvey et al. (2005), 'dropout'

[47]: Crall et al. (2013), 'HotSpotter - Patterned species instance recognition'

[70]: Bolger et al. (2012), 'A computer-assisted system for photographic mark–recapture analysis'

Modelling misidentification

Most studies simply remove the problematic data and assume that no misidentification has occurred in the data they retain. The proportion of

removed data can be quite large. For example, Laguardia et al. (2021) [54] used only 58% of the collected samples. Lukacs and Burnham (2005) [71] point out that it may be beneficial to allow a small degree of uncertainty in identification (2–5%) by modelling it, if the cost is offset by the number of additional samples that can be kept. Two types of models have been developed that model uncertainty in identification.

[54]: Laguardia et al. (2021), 'Nationwide abundance and distribution of African forest elephants across Gabon using non-invasive SNP genotyping'

[71]: Lukacs et al. (2005), 'Estimating population size from DNA-based closed capture-recapture data incorporating genotyping error'

The first type of model integrates *genotype uncertainty*. Wright et al. (2009) [72] developed a model that uses replicates of genotypes to model the genotype observation process before modelling capture-recapture. Let G^{obs} be the observed genotypes, G^{true} the true genotypes, X the capture matrix, d the probability of allelic dropout, N the population size, θ the CR parameters, and γ the allele frequencies. They proposed a model with three components: the genotype observation process $[G^{obs}|G^{true}, X, p]$, the genotype distribution $[G^{true}|N, \gamma]$, and the sampling process $[X|N, \theta]$.

[72]: Wright et al. (2009), 'Incorporating genotype uncertainty into mark-recapture-type models for estimating abundance using DNA samples'

Another model, developed by Knapp et al. (2009) [73], uses the likelihood of the observed genotype $P(G^{obs} = g)$ to compute the probability that two samples i and j from the same individual lead to the observed genotypes g and g', $P(G_i^{obs} = g, G_j^{obs} = g'|G_i^{true} = G_j^{true})$. Then they reverse the probability using Bayes' rule to get $P(G_i^{true} = G_j^{true}|G_i^{obs} = g, G_j^{obs} = g')$. The computation uses the population size N, allowing its estimation.

[73]: Knapp et al. (2009), 'Incorporating genotyping error into non-Invasive DNA-based mark—recapture population estimates'

However, these two models can only be used for genetic tags, and they have only been developed to estimate population size, not survival. (Wright et al.'s model could be extended to open populations.)

The second type of model deals directly with potential misidentification in the capture-recapture history matrix. A first model was developed by Lukacs and Burnham (2005) [71] for closed populations. Let p be the probability of capturing an individual, and α the probability that the first encounter of a history be correctly identified. They give the probability of a history $h = (h_1, \ldots, h_T)$ with the first encounter at l and subsequent encounters to be

$$P(h) = \left[\prod_{1}^{t=l-1}(1-p)\right] p\alpha \left[\prod_{t=l+1}^{T} p^{h_t}(1-p)^{1-h_t}\right],$$

and the probability of a history h with the first encounter at l and no recapture to be

$$P(h) = \left[\prod_{1}^{t=l-1}(1-p)\right]\left[p\alpha\left(\prod_{t=l+1}^{T}(1-p)\right) + p(1-\alpha)\right].$$

Yoshizaki et al. (2011) [67] remark that Lukacs and Burnham "do not present a rigorous development of the cell probabilities" and that the model "ignores the complete dependence between the pair of histories created whenever a genotype is incorrectly identified". They modify the model, giving the probability of a history with several encounters:

[67]: Yoshizaki et al. (2011), 'Modeling misidentification errors that result from use of genetic tags in capture–recapture studies'

$$P(h) = \prod_{t=1}^{T}(p_t\alpha)^{h_t}(1-p_t\alpha)^{1-h_t},$$

and the probability of a history h with a unique encounter:

$$P(h) = \prod_{t=1}^{T} (p_t \alpha)^{h_t} (1 - p_t \alpha)^{1-h_t} + p_{t_1(h)}(1 - \alpha)$$

where $t_1(h)$ is the occasion of the first capture of h. The sum of the probabilities of the possible histories is greater than 1. Thus, a multinomial model cannot be used. To estimate the parameters, the authors use a least-squares method.

Based on Yoshizaki's model, Link et al. (2010) [74] developed a latent multinomial model (LMM) that they named $M_{t,\alpha}$. In the $M_{t,\alpha}$ model, a misidentification is coded by a 2 in the latent history:

[74]: Link et al. (2010), 'Uncovering a Latent Multinomial'

Latent history	Observed histories
102	100
	001

And the probability of latent history h is

$$P(h) = \prod_{t=1}^{T} (1 - p_t)^{I(h_t=0)} (p_t \alpha)^{I(h_t=1)} (p_t(1 - \alpha))^{I(h_t=2)}.$$

The LMM is the framework that I chose to use in this thesis.

Another interesting model is presented by Yoshizaki et al. (2011) [67]. The idea of the model is to simply remove histories likely to be ghosts and apportion a multinomial law to all the other observable histories. Let h' be an observed history with two or more captures, and $\mathcal{H}'$ the set of possible histories with two or more captures. Let the random variable $y_{h'}$ represent the number of observed histories h'. Then $\mathbf{y}' = (y_{h'_1}, ...)$ follows a multinomial of index $N' = \sum \mathbf{y}'$ and probabilities

[67]: Yoshizaki et al. (2011), 'Modeling misidentification errors that result from use of genetic tags in capture–recapture studies'

$$\pi_{h'} / \pi*$$

where $\pi_{h'} = \prod_{t=1}^{T} p_t^{h'_{i,t}} (1 - p_t)^{1-h'_{i,t}}$, and

$$\pi* = \sum_{h \in \mathcal{H}'} \pi_h.$$

Finally, N is estimated by

$$N' / \hat{\pi}*$$

As this model is very simple, I will refer to it as 'Yoshizaki's model' and use it in comparison with the LMM and LMM extensions described in Chapter 2 and Chapter 3.

Developments around the latent multinomial model

Since its development in 2010, the latent multinomial model (LMM, [74]) has received some interest. Bonner and Holmberg (2013) [61] adapted it to photo identification with multiple "non-invasive" marks. They modelled individual misidentification coming from left-side and right-side photographs that might have been taken of the same individual. In the same year, McClintock et al. (2013) [62] developed a similar model and implemented it in the R package `multimark` [63]. McClintock et al.

[61]: Bonner et al. (2013), 'Mark-recapture with multiple, non-invasive marks'

[62]: McClintock et al. (2013), 'Integrated modeling of bilateral photo-identification data in mark–recapture analyses'

(2014) [75] also extended the LMM to individual heterogeneity in capture or identification probability.

When Link et al. published their model, they explained how maximum likelihood estimation was not realistic and showed how to construct a Markov chain to estimate the parameters (details in Section 1.2.4). In 2014, Vale et al. used ADMB [76] to implement maximum likelihood estimation [77]. The downside of Vale et al.'s method is that it is limited to a single-state closed population, with no individual heterogeneity. This means that transitions and survival cannot be estimated with this method. They tested the model on simulations with low capture rates (down to 0.1) and showed that the model was highly biased, underestimating the population size. In 2015 and 2016, Schofield & Bonner [78] and then Bonner et al. [79] corrected and improved the construction of the MCMC for faster computing. They also extended the model to a different case of misidentification. They considered the case in which two individuals can be confused. That is, an individual A who is seen can be misidentified as another existing and previously tagged individual B, even though B is not seen.

[75]: McClintock et al. (2014), 'Probit models for capture-recapture data subject to imperfect detection, individual heterogeneity and misidentification'

[76]: Fournier et al. (2012), 'AD Model Builder'

[77]: Vale et al. (2014), 'Maximum likelihood estimation for model $M_{t,\alpha}$ for capture–recapture data with misidentification'

[78]: Schofield et al. (2015), 'Connecting the latent multinomial'

[79]: Bonner et al. (2015), 'Extending the latent multinomial model with complex error processes and dynamic markov bases'

The thesis project

Context and objectives

This PhD project was funded by the French National Research Agency (ANR) as part of the project **MoVe=>ADAPT** – *Mosquito Vectors Adaptation in a Changing World*. The latter project aims 'to study how adaptation of mosquito vectors to environmental modifications associated with global change impacts their fitness and the life-history traits influencing vectorial capacity, in order to predict more accurately the epidemiological consequences of niche expansions and the spread of mosquito-borne pathogens'. The thesis project involved applying the capture-recapture analytical approach, which has been instrumental to obtain unbiased field estimates of vertebrate population demographic parameters, to the study of natural mosquito populations. The focus was on mosquito larvae. The original plan was to collect and model data to assess the cost of adaptation by comparing fitness trade-offs in reciprocally transplanted natural populations occurring in contrasting environments. Since the data quality was expected to be very poor, the model needed to account for possible misidentification.

At the beginning of the PhD, the protocol for gathering the data on larvae was an untested idea. The objective of the PhD was to develop the model needed to analyse the data in parallel to the development of the protocol. Some preliminary data was expected to be available by the beginning of the first year. With this perspective, the PhD was started with the purpose of delivering a model able to analyse this mosquito larvae data. The plan was to select a model dealing with misidentification, implement it and test it, extending it step by step towards the final model. Since the selected model [74] was for a single-state closed population, these steps were extensions to a multi-state closed population, single-state open population and multi-state open population, adapting these to specific

[74]: Link et al. (2010), 'Uncovering a Latent Multinomial'

data requirements, incorporating environmental survival covariates, and improvement of the estimates.

Around the midway point of the PhD, it became clear that the data would not be available before the end of the thesis. It was judged that the complete lack of data for a study aiming to specifically model this data would greatly diminish the overall interest of the project. Hence, while keeping the objective of the mosquito larvae model in mind, it was decided to make the study more general. This was quite easy since most of the steps already worked on were not specific to the mosquito model. The thesis ended up considering the mosquito model simply as an illustration of the potential of the LMM developed and extended in the thesis.

Structure and content of the thesis

The thesis has five chapters, each describing different extensions of the model.

Chapter 1: Closed population modelling of misidentification. This chapter describes the original model from Link et al. (2010) [74] and the algorithm used to construct an MCMC to estimate the parameters. The chapter also presents the extension of the model to multi-state observations. Finally, the chapter tests the model under various scenarios and examines whether transition rates between states are well estimated or not in the presence of misidentification.

Chapter 2: Using a covariate of identification. This chapter presents an extension of the LMM that incorporates an identification quality covariate in order to better discriminate which history with a unique capture is a potential ghost. A comparative study of my model with the original model and with Yoshizaki's model is also made.

Chapter 3: Repeated observations on an occasion. This chapter presents an extension of the model that accounts for multiple captures of the same individual on a given occasion. The new model is compared to Yoshizaki's model, which could also be used for these cases. I also applied the model to a real dataset from a study of the Eurasian otter (*Lutra lutra*).

Chapter 4: Open population modelling of misidentification. The fourth chapter presents the extension of the LMM to open populations. First the single-state case is treated, followed by its extension including an identification covariate, and lastly the multi-state open population case. A simulation study shows the performance of the single-state open population model.

Chapter 5: Using the LMM to study mosquito larvae survival. In the last chapter, two fieldwork protocols are put forward for the mosquito larvae study in order to obtain unbiased estimates with the multi-state open population LMM. A simulation study compared the two protocols, which differ in their way of dealing with pupae.

The chapters are relatively independent, but any reader not already familiar with the LMM should read the first chapter. Chapter 1 can be seen as an introduction to all the subsequent chapters, as it presents the

LMM, its limitations, and how to extend it in different directions. Chapters 2, 3 and 4 are meant to be read independently by those interested in specific extensions of the model. Thus, some parts are repeated, especially the algorithms used to estimate the parameters.

At the very end of the thesis, I added a section with most of the notations used in this work. It is presented twice so that one can be detached to accompany the reading of a paper version.

Chapter one

Closed population modelling of misidentification

1

1.1 Aim of the chapter

Despite the interest that the LMM received and its potential, there is no easy implementation. As a result it is not really used. Researchers tend to simply delete the poor quality data and assume that no misidentifications remain. In addition, Vale et al. (2014) [77] used simulations to show that the LMM leads to very biased and uncertain estimates when recapture is low. They tested the model with capture probabilities varying between 0.05 and 0.5 and identification probabilities between 0.9 and 0.99. It would be interesting to know how the model performs with a lower identification rate. Finally, the scope of the model is limited to single-state closed-population experiments. It is even unknown how misidentifications affect transition estimates if they are not accounted for in a multi-state experiment.

[77]: Vale et al. (2014), 'Maximum likelihood estimation for model $M_{t,\alpha}$ for capture–recapture data with misidentification'

In this chapter, I implement the LMM, using the algorithm proposed by Bonner et al. (2016) [79]. I code it in the R language, using the NIMBLE package [80]. Since R and NIMBLE are widely used by researchers doing CMR, such an implementation will be easier to share and use.
I replicate the simulation study of Vale et al. (2014) [77]. I extend the range of parameters to include lower identification rates, down to 80% good identifications. I also test the effect of an informative prior on the identification rate as a first way of correcting the bias from low capture probabilities.
In the next step I extend the LMM to multi-state observations, while still estimating population size in closed population experiments. This will allow the additional estimates of migration between studied sites or transition between states.
I test the effect of ignoring the misidentifications on the transition estimates in the case of state dependant capture and identification probabilities, and see how the multi-state LMM performs in estimating both the transition probabilities and the population size.

[79]: Bonner et al. (2015), 'Extending the latent multinomial model with complex error processes and dynamic markov bases'
[80]: Valpine et al. (2017), 'Programming With Models'

1.2 Models of capture-recapture for single-state in closed-population

1.2.1 Notations

Parameters

- N: Population size,
- p_t: Probability that an individual is captured at time t,
- α: Probability that a captured individual correctly identified
- $\psi_{r,s}$: in multistate, probability that an individual transition from state r to state s between two consecutive occasions.

Data and Latent variables

- ω_i: Observed history i
- ν_j: Latent error history j (in which misidentification are noted down)
- ξ_k: Latent capture history k (real capture history)

Statistics

- y_i: number of observed history i
- x_j: number of latent error history j
- z_k: number of latent capture history k
- $\mathbf{y} = (y_1, ..., y_{2^T-1})$): vector of counts of observed histories
- $\mathbf{x} = (x_1, ..., x_{3^T})$): vector of counts of latent error histories
- $\mathbf{z} = (z_1, ..., z_{2^T})$): vector of counts of latent capture histories

1.2.2 The "classical" model (M_t)

When estimating population size N in a closed capture-recapture experiment (i.e. the population is assumed not to change), with the model known as M_t ([25, 26]), individuals are assumed to be observed ('captured') with probability p_t at occasion t for $t = 1, 2, ..., T$ and identified individually through tags/tracking devices or natural markings. Capture events are assumed independent between individuals and over time.

[25]: Darroch (1958), 'The Multiple-Recapture Census'
[26]: Otis et al. (1978), 'Statistical inference from capture data on closed animal populations'

For each occasion, an individual is assigned 0 if it was not captured, or 1 if it was. This leads to 2^T possible distinct histories, including the unobservable all-zero history. They are represented by the sequence $\boldsymbol{\omega}_i = (\omega_{i,1}, ..., \omega_{i,T})$ where $\omega_{i,t}$ is 0 or 1. Here, y_i is the number of individuals with history ω_i and $\mathbf{y} = (y_1, y_2, ..., y_{2^T-1})$. The likelihood of history i is

$$\pi_i = \prod_{t=1}^{T} \left[p_t^{I(\omega_{i,t}=1)} (1 - p_t)^{I(\omega_{i,t}=0)} \right] \tag{1.1}$$

where $I(test)$ is 1 if *test* is true, and is 0 otherwise. Then, $\mathbf{y}$ follows a multinomial distribution

$$[\mathbf{y}|N, \mathbf{p}] = \frac{N!}{\prod_i y_i!} \prod_i \pi_i^{y_i} \tag{1.2}$$

1.2.3 The Latent Multinomial Model, $M_{t,\alpha}$

To account for individual misidentifications, Yoshizaki et al. [67] proposed an $M_{t,\alpha}$ model in which captured individuals are correctly identified with probability α. Misidentifications are assumed to always create a new individual (a "ghost"). An individual cannot be mistaken as another and two errors cannot create the same ghost. Essentially what will happen is what is described on Figure 1.1.

[67]: Yoshizaki et al. (2011), 'Modeling misidentification errors that result from use of genetic tags in capture–recapture studies'

To estimate the parameters of the model, Link et al. (2010) [74] developed a latent structure for the $M_{t,\alpha}$ model, which allows for Bayesian parameter estimation. In this structure, misidentifications are represented by 2 in the latent error histories $\boldsymbol{\nu}_j = (\nu_{j,1}, ..., \nu_{j,T})$. The frequency of the latent error history ν_j is noted x_j, and the vector of all latent error frequencies is $\mathbf{x} = (x_1, ..., x_{3^T})$. To make future developments of the

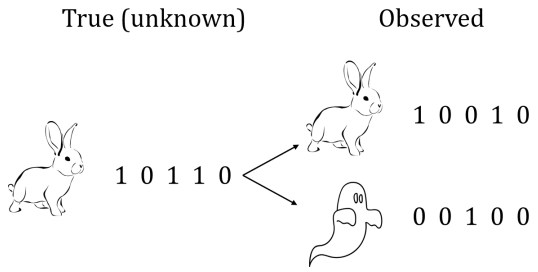

Figure 1.1: Misidentification process. On occasion 3, the cute rabbit was misidentified so we do not realise that we have captured it (a 0 is registered) and we add a ghost in the dataset.

model easier, the observation process of the likelihood is broken down into two parts, the capture process and the identification process. We followed Bonner et al. (2015) [79] by introducing latent capture histories $\boldsymbol{\xi}_k = (\xi_{k,1}, ..., \xi_{k,T})$. These are the true capture histories, i.e. in the absence of individual misidentifications, composed of 0 and 1. The frequency of the latent capture history ξ_k is noted z_k, and the vector of all latent capture frequencies is $\mathbf{z} = (z_1, ..., z_{2^T})$.

[79]: Bonner et al. (2015), 'Extending the latent multinomial model with complex error processes and dynamic markov bases'

In this model framework, the observed frequencies vector $\mathbf{y}$ is a linear transformation $\mathbf{y} = \mathbf{A}\mathbf{x}$ of the latent error frequencies $\mathbf{x}$ for a known matrix $\mathbf{A}$. The constraint matrix $\mathbf{A}$ is $(2^T - 1) \times 3^T$ with a 1 in row i and column j if the latent error history j gives rise to the observed i. All the other entries are zeros. The latent capture frequencies vector $\mathbf{z}$ is also a linear transformation of the same $\mathbf{x}$, $\mathbf{z} = \mathbf{B}\mathbf{x}$, for another known matrix $\mathbf{B}$. $\mathbf{B}$ is $2^T \times 3^T$ with 1 at row k and column j if the latent capture history ξ_k and the latent error history v_j have the same capture pattern.

The joint likelihood of $\mathbf{y}$, $\mathbf{x}$ and $\mathbf{z}$ is

$$[\mathbf{y}, \mathbf{x}, \mathbf{z}|N, p, \alpha] = I(\mathbf{y} = \mathbf{A}\mathbf{x})\,[\mathbf{x}|\mathbf{z}, \alpha]\,[\mathbf{z}|N, \mathbf{p}] \tag{1.3}$$

The probability of the capture process $[\mathbf{z}|N, \mathbf{p}]$ is the same as the closed CMR M_t model, using histories ξ and frequencies $\mathbf{z}$. The capture likelihood is the following multinomial product where π_k is computed as in Equation 1.2, using history ξ instead of ω:

$$[\mathbf{z}|N, \mathbf{p}] = \frac{N!}{\prod_k z_k!} \prod_k \pi_k^{z_k} \tag{1.4}$$

Bonner et al. (2015) [79] gives the likelihood of the identification process, knowing the real captures:

$$[\mathbf{x}|\mathbf{z}, \alpha] = I(\mathbf{z} = \mathbf{B}\mathbf{x}) \frac{\prod_k z_k!}{\prod_j x_j!} \prod_j \left[\prod_{t=1}^{T} \alpha^{I(v_{j,t}=1)} (1 - \alpha)^{I(v_{j,t}=2)} \right]^{x_j} \tag{1.5}$$

The ratio of factorials accounts for the many relabelling of the marked individuals that would produce the same counts in $\mathbf{x}$ and $\mathbf{z}$.

The full likelihood is obtained by summing $[\mathbf{y}, \mathbf{x}, \mathbf{z}|N, p, \alpha]$ over all values of $\mathbf{x}$ belonging to the set $\mathcal{F}_\mathbf{y} = \{\mathbf{x}|\mathbf{y} = \mathbf{A}\mathbf{x}\}$:

$$[\mathbf{y}|N, p, \alpha] = \sum_{x \in \mathcal{F}_\mathbf{y}} [\mathbf{y}|\mathbf{x}, \mathbf{z}, N, p, \alpha] \tag{1.6}$$

Note that there is no need to sum over **z** because **z** is defined by **x**

1.2.4 Estimating parameters of the LMM

The feasible set $\mathcal{F}_{\mathbf{y}}$ is complicated to enumerate, which makes the likelihood (Equation 1.6) almost untractable in terms of computation. Maximum likelihood estimation (MLE) is therefore not practical. Conveniently, Link et al. [74] show how a Markov Chain Monte Carlo (MCMC) can be constructed in a Bayesian analysis. The Markov chain allows for the estimation of the posterior density:

[74]: Link et al. (2010), 'Uncovering a Latent Multinomial'

$$[N, \mathbf{p}, \alpha|\mathbf{y}] \propto [\mathbf{y}|N, \mathbf{p}, \alpha]\,[N]\,[\mathbf{p}]\,[\alpha], \tag{1.7}$$

where $[N]$, $[\mathbf{p}]$ and $[\alpha]$ denote the priors on population size, capture probability and identification probability. The algorithm presented in this section is the one developed by Link et al. (2010) [74], with the improvements from Bonner et al. (2015) [79]. The modification they made are explained where they occur.

[79]: Bonner et al. (2015), 'Extending the latent multinomial model with complex error processes and dynamic markov bases'

The MCMC is constructed following six steps:

1. Let $\beta(a_0^t, b_0^t)$ denote the beta prior on p_t and $\beta(a_0^\alpha, b_0^\alpha)$ denote the beta prior on α.

2. Initialize all parameters as well as a set of latent histories satisfying **y**=**Ax**. Such a set can be obtained by assuming that no mistakes were made. The latent frequencies of the histories containing 2's are 0 and all the other match the observed frequencies one-to-one. In order to run several chains with different initialisations, one can take the previous initialisation of **x** and follow the later steps of (5) by only adding misidentification to the set and always accepting the proposed ones without going though the Metropolis-Hasting acceptance. In the initial latent set, fix the number of unseen individual to a random realistic number.

3. Sample the capture rate with Gibbs sampling as shown by Link et al. [74]. The likelihood being multinomial, it follows that the beta priors lead to full conditional beta posterior distribution:

$$p_t|\mathbf{Z} \sim \beta(a_0^t + a^t, b_0^t + b^t)$$

   where a^t [1] is the number of captured individuals at time t and b^t [2] is the number of unseen individuals at time t (including those never seen).

   1: $a^t = \sum_n I(\xi_{n,t} = 1)$
   2: $b^t = \sum_n I(\xi_{n,t} = 0)$

4. Sample the identification rate using Gibbs sampling. Similar to the capture rate, it has a full conditional beta posterior distribution:

$$\alpha|\mathbf{x} \sim \beta(a_0^\alpha + a^\alpha, b_0^\alpha + b^\alpha)$$

   where a^α [3] is the total number of correct identifications and b^α [4] is the total number of misidentifications.

   3: $a^\alpha = \sum_j \sum_t x_j I(\nu_{j,t} = 1)$
   4: $b^\alpha = \sum_j \sum_t x_j I(\nu_{j,t} = 2)$

5. Sample jointly N and **x** since the number of errors in **x** changes the population size. Sampling **x** requires to be able to sample from

$\mathcal{F}_{\mathbf{y}}$. Link et al. (2010) [74] proposed sampling moves from the null space of matrix $\mathbf{A}$ (that is from the set of $\mathbf{x}$ such as $\mathbf{Ax} = 0$) [5], and adding or subtracting them to the current $\mathbf{x}$ in the MCMC.
Schofield and Bonner [78] showed that if the basis of $Ker_{\mathbb{Z}}(\mathbf{A})$ was not carefully selected, some parts of the space $\mathcal{F}_{\mathbf{y}}$ could be disconnected from the others and the Markov chain would only explore sub-spaces, depending on the initial $\mathbf{x}$, possibly leading to biased estimations. They proposed to sample moves from the Markov basis of $\mathbf{A}$ [81], a set in $Ker_{\mathbb{Z}}(A)$ that connect all $\mathcal{F}_{\mathbf{y}}$ irrespective of the values in $\mathbf{y}$. Such a basis ensures that the whole set $\mathcal{F}_{\mathbf{y}}$ is connected by single moves and that no move will get out of the set. The drawback is that the computation of that Markov basis is heavy and algebraic software such as 4ti2 [82] will not be able to calculate it for $T \geq 5$. Bonner et al. [79] proposed a mechanism to avoid computing that basis. It consists in sampling from dynamic Markov basis [83] which is the set of moves $M(x)$ that connect each $\mathbf{x}$ to some neighbours. $M(x)$ is a subset of the complete Markov base that contains only the moves with a positive acceptance probability. It is easy to infer which sets $\mathbf{x}'$ are next to a given $\mathbf{x}$, making the dynamic base much more simple to use than the Markov base.
The algorithm is randomly adding or removing an error from the set of latent histories. To add an error, the authors choose a history that may have generated a ghost (i.e. a history containing a 0), and "merge" it with a potential ghost (i.e. replace the 0 by a 2 and remove the ghost history). To remove an error, they choose a history containing a 2, replace it by a 0 and add a history with a unique capture (coded 1) at that time.
More formally, follow the steps:

a) Define:
   - $\boldsymbol{\nu}^{(1t)}$ the history with a unique capture at time t (potential ghost),
   - $\chi_{0,t}(\mathbf{x}) = \{\boldsymbol{\nu} | \nu_t = 0, x_\nu > 0, x_{\nu^{(1t)}} > 0\}$ the set of histories having *potentially* generated a ghost at time t, for the given $\mathbf{x}$,
   - $\chi_{2,t}(\mathbf{x}) = \{\boldsymbol{\nu} | \nu_t = 2, x_\nu > 0\}$ the set of histories *containing* a ghost at time t, for the given $\mathbf{x}$.

b) With probability 0.5, go to (i), otherwise go to (ii).
   i. Add a misidentification (i.e. a ghost) to the latent set.
      - Sample $\boldsymbol{\nu}^{(0)} \in \chi_{0.}(x) = \bigcup_t \chi_{0,t}(x)$.
      - Sample $t \in \{t | \nu_t^{(0)} = 0, x_{\nu_{1t}} > 0\}$.
      - Define $\boldsymbol{\nu}^{(2)} = \boldsymbol{\nu}^{(0)} + 2\boldsymbol{\nu}^{(1t)}$.
      - Define the move $b_{\boldsymbol{\nu}^{(0)}, \boldsymbol{\nu}^{(1t)}, \boldsymbol{\nu}^{(2)}} = (-1, -1, +1)$, and $b_\nu = 0$ for all other latent histories.
   ii. Remove a misidentification from the latent set.
      - Sample $\boldsymbol{\nu}^{(2)} \in \chi_{2.}(x) = \bigcup_t \chi_{2,t}(x)$.
      - Sample $t \in \{t | \boldsymbol{\nu}_t^{(2)} = 2\}$.
      - Define $\boldsymbol{\nu}^{(0)} = \boldsymbol{\nu}^{(2)} - 2\boldsymbol{\nu}^{(1t)}$.
      - Define the move $b_{\boldsymbol{\nu}^{(0)}, \boldsymbol{\nu}^{(1t)}, \boldsymbol{\nu}^{(2)}} = (+1, +1, -1)$, and $b_\nu = 0$ for all other latent histories.

c) Define $\mathbf{x}' = \mathbf{x}^{(k-1)} + \mathbf{b}$.
d) Calculate $\mathbf{z}' = \mathbf{B}\mathbf{x}'$ and $N' = \sum \mathbf{x}'$.

5:

$$\begin{aligned} Ker_{\mathbb{Z}}(\mathbf{A}) &= Ker(\mathbf{A}) \cap \mathbb{Z}^d \\ &= \{\mathbf{x} \in \mathbb{Z}^d | \mathbf{Ax} = 0\} \end{aligned}$$

[78]: Schofield et al. (2015), 'Connecting the latent multinomial'

[81]: Diaconis et al. (1998), 'Algebraic algorithms for sampling from conditional distributions'

[82]: team (), *4ti2—A software package for algebraic, geometric and combinatorial problems on linear spaces*

[79]: Bonner et al. (2015), 'Extending the latent multinomial model with complex error processes and dynamic markov bases'

[83]: Dobra (2012), 'Dynamic markov bases'

e) Let r_1 be

$$r_1 = min\left(1, \frac{[\mathbf{y}, \mathbf{x}', \mathbf{z}'|N', \mathbf{p}, \alpha]}{[\mathbf{y}, \mathbf{x}^{(k-1)}, \mathbf{z}^{(k-1)}|N, \mathbf{p}, \alpha]} \frac{q(\mathbf{x}^{(k-1)}|\mathbf{x}')}{q(\mathbf{x}'|\mathbf{x}^{(k-1)})}\right) \quad (1.8)$$

With probability r_1, set $\mathbf{x}^{(k)} = \mathbf{x}'$, $\mathbf{z}^{(k)} = \mathbf{z}'$ and $N^{(k)} = N'$. Otherwise set $\mathbf{x}^{(k)} = \mathbf{x}^{(k-1)}$, $\mathbf{z}^{(k)} = \mathbf{z}^{(k-1)}$ and $N^{(k)} = N^{(k-1)}$.

See Appendix A for the detail of how the likelihood ratio part of r_1 simplifies.

6. Sample the number of unseen individuals x_1:
   a) set $\mathbf{x}' = \mathbf{x}$, $\mathbf{z}' = \mathbf{z}$ and x_0 the number of unseen individual in $\mathbf{z}$ (and $\mathbf{x}$),
   b) sample a move $c \in [-D, D]$ where D is fixed integer,
   c) define $x_0' = x_0 + c$,
   d) set the number of unseen individuals in $\mathbf{x}'$ and $\mathbf{z}'$ to x_0',
   e) accept it with probability r_2 with:

$$r_2 = min\left(1, \frac{[\mathbf{z}'|N', p]}{[\mathbf{z}|N, p]}\right). \quad (1.13)$$

7. repeat steps 3 to 6 as much as needed.

1.2.5 Proof of convergence

Bonner et al. (2015) [79] give the proof of convergence of the algorithm they developed for their band read error model in the supplementary material. For the sake of completeness of this manuscript, we **copy** the proof of convergence they gave, using the notations from this thesis. Despite the model of Bonner et al. (2015) being different, the proof is the same for the algorithm of the $M_{t,\alpha}$ model.

To prove that chains generated from the algorithm of Section 1.2.4 converge to the correct distribution, we need to satisfy four conditions:

1. that Step 3 and 6 produce chains which converge to $\pi(N, \mathbf{p}|\mathbf{z})$ for any $\mathbf{z}$ such that $\mathbf{z} = \mathbf{B}\mathbf{x}$ for some $\mathbf{x} \in \mathcal{F}_y$,
2. that step 4 produces chains which converge to $\pi(\alpha|\mathbf{x})$,
3. that Step 5 produces chains which converge to $\pi(\mathbf{x}|\mathbf{y}, N, \mathbf{p}, \alpha)$ for any N, $\mathbf{p}$ and α in the parameter space,
4. that the joint posterior distribution $\pi(\mathbf{x}, \mathbf{z}, N, \mathbf{p}, \alpha|\mathbf{y})$ satisfies the positivity condition of Robert and Casella (2010, pg. 345 [84]).

Sampling from $\pi(N, \mathbf{p}|\mathbf{z})$ is equivalent to sampling from the posterior distribution for a simple M_t model. This is now standard, and so we conclude that Condition 1 is satisfied. It is also trivial to show that Condition 2 is satisfied since it is the proof that the β prior is conjugate with the product of binomial in the likelihood. It is also simple to show that the positivity constraint is satisfied given that the prior distributions for $\mathbf{p}$ is positive over all of $[0, 1]^T$, as assumed. It remains to show that Condition 3 holds.

We assume here that $\mathcal{F}_y$ contains at least two elements. The fibre always contains at least one element with no errors which we denote by $\mathbf{x}^{\emptyset}$. The

Likelihood ratio simplification

When adding an error, we have

$$\frac{[\mathbf{x}'|\mathbf{z}', \alpha]}{[\mathbf{x}^{(k-1)}|\mathbf{z}^{(k-1)}, \alpha]} \frac{[\mathbf{z}'|N', \mathbf{p}]}{[\mathbf{z}^{(k-1)}|N, \mathbf{p}]} = \frac{x_{\nu_0} x_{\nu_1} (1-\alpha)}{x'_{\nu_2} \alpha} \frac{1}{N \prod_{t=1}^{T} p_t} \quad (1.9)$$

and when removing an error,

$$\frac{[\mathbf{x}'|\mathbf{z}', \alpha]}{[\mathbf{x}^{(k-1)}|\mathbf{z}^{(k-1)}, \alpha]} \frac{[\mathbf{z}'|N', \mathbf{p}]}{[\mathbf{z}^{(k-1)}|N, \mathbf{p}]} = \frac{x_{\nu_2} \alpha}{x'_{\nu_0} x'_{\nu_1} (1-\alpha)} N' \prod_{t=1}^{T} p_t \quad (1.10)$$

Proposal density

The proposal densities are calculated by multiplying the probabilities of each sampling step used for defining the move. They are successively: the probability of adding (or removing) an error, the probability of choosing the ν_0 (or ν_2) and the probability of choosing the t knowing the sampled ν. When adding an error, the proposal density q is:

$$q(x'|x^{(k-1)}) = \frac{0.5}{\#\chi_0.\#\{t|\nu_{0,t} = 0, x_{\nu_{1t}} > 0\}} \quad (1.11)$$

and when removing an error, is:

$$q(x'|x^{(k-1)}) = \frac{0.5}{\#\chi_2.\#\{t|\nu_{2,t} = 2\}} \quad (1.12)$$

where $\#S$ denotes the cardinality of S.

[79]: Bonner et al. (2015), 'Extending the latent multinomial model with complex error processes and dynamic markov bases'

[84]: Robert et al. (2004), *Monte Carlo Statistical Methods*

entries of this element are

$$\mathbf{x}_\nu^\emptyset = \begin{cases} \mathbf{y}_\nu & \text{if } \nu \text{ is observable,} \\ 0 & \text{otherwise} \end{cases} .$$

Cases in which $\mathcal{F}_y = \mathbf{x}^\emptyset$ arise when no errors could have occurred, for example, if no individuals were ever recaptured. These situations are easily identified and there is no need to sample from the joint posterior of both $\mathbf{x}$ and θ in such cases since $\mathbf{x} = \mathbf{x}^\emptyset$ with probability one.

Some useful results that are easy to prove are:

1. that any configuration of the latent error histories within the fibre has a positive probability under the conditional posterior for all values of the parameters in the parameter space,

   Lemma 1.2.1 *If $\mathbf{x} \in \mathcal{F}_y$, then $\pi(\mathbf{x}|\mathbf{y}, N, \mathbf{p}, \alpha) > 0$ for all values of N and $\mathbf{p}$ in the parameter space.*

2. that the local sets within the dynamic Markov basis are symmetric,

   Lemma 1.2.2 *Let $\mathcal{M}_1(\mathbf{x})$ and $\mathcal{M}_2(\mathbf{x})$ be, respectively the sets of moves that add and remove an error to $\mathbf{x}$, and let $\mathbf{x} \in \mathcal{F}_y$. If $\mathbf{b}^+ \in \mathcal{M}_1(x)$, then $-\mathbf{b}^+ \in \mathcal{M}_2(x + \mathbf{b}^+)$, and if $\mathbf{b}^- \in \mathcal{M}_2(x)$, then $-\mathbf{b}^- \in \mathcal{M}_1(x + \mathbf{b}^-)$.*

3. that all proposals remain inside $\mathcal{F}_y$,

   Lemma 1.2.3 *Let $\mathbf{x} \in \mathcal{F}_y$. If $\mathbf{b} \in \mathcal{M}(x) = \mathcal{M}_1(x) \bigcup \mathcal{M}_2(x)$ then $x + b \in \mathcal{F}_y$.*

4. that there is a unique element $\mathbf{x}^\emptyset$ in $\mathcal{F}_y$ with no errors.

   Lemma 1.2.4 *Let $e_t(\mathbf{x})$ be the number of misidentification in $\mathbf{x}$ at occasion t. Suppose that $\mathbf{x}^\emptyset \in \mathcal{F}_y$. Then $e_t(\mathbf{x}^\emptyset) = 0 \; \forall t = 1, \ldots, T$ if and only if*

   $$\mathbf{x}_\nu^\emptyset = \begin{cases} \mathbf{y}_\nu & \text{if } \nu \text{ is observable,} \\ 0 & \text{otherwise} \end{cases}$$

First we establish irreducibility. Proposition 1.2.5 implies that there is a path connecting any two elements in the fibre while Proposition 1.2.6 implies that each step, and hence the entire path, has positive probability under the transition kernel. Together, these show that that the chains are irreducible.

Proposition 1.2.5 *For any distinct $\mathbf{x}_1, \mathbf{x}_2 \in \mathcal{F}_y$ there exists a sequence of moves $\mathbf{b}_1, \ldots, \mathbf{b}_L$ such that:*

1. $\mathbf{b}_{L'} \in \mathcal{M}(\mathbf{x}_1 + \sum_{l=1}^{L'-1} \mathbf{b}_l)$ *for all* $L' = 1, \ldots, L$,
2. $\mathbf{x}_1 + \sum_{l=1}^{L'-1} \mathbf{b}_l \in \mathcal{F}_y$ *for all* $L' = 1, \ldots, L-1$, *and*
3. $\mathbf{x}_2 = \mathbf{x}_1 + \sum_{l=1}^{L} \mathbf{b}_l$,

where we take $\mathbf{x}_1 + \sum_{l=1}^{0} \mathbf{b}_l = \mathbf{x}_1$.

Proof. The proof follows by (reverse) induction on the number of errors. Suppose that $e_t(\mathbf{x}_1) > 0$ for some t. Then $\chi_{2,t}(\mathbf{x}_1)$ is non-empty and $\exists \mathbf{b}^-_{11} \in \mathscr{M}_2(\mathbf{x}_1)$. Then $e_t(\mathbf{x}_1 + \mathbf{b}^-_{11}) = e_t(\mathbf{x}_1) - 1$ and $\mathbf{x}_1 + \mathbf{b}^-_{11} \in \mathscr{F}_y$ by Lemma 1.2.3. Repeating this procedure $L_1 = \sum_{t=1}^{T} e_t(\mathbf{x}_1)$ times, we find $\mathbf{b}^-_{11}, \ldots, \mathbf{b}^-_{1L_1}$ such that

1. $\mathbf{b}^-_{1L'} \in \mathscr{M}_2(\mathbf{x}_1 + \sum_{l=1}^{L'-1} \mathbf{b}^-_{1l})$ for $L' = 1\ldots, L_1$,
2. $\mathbf{x}_1 + \sum_{l=1}^{L'} \mathbf{b}^-_{1l} \in \mathscr{F}_y$ for all $L' = 1\ldots, L_1$, and
3. $e_t(\mathbf{x}_1 + \sum_{l=1}^{L_1} \mathbf{b}^-_{1l}) = 0$.

It follows by Lemma 1.2.4 that $\mathbf{x}_1 + \sum_{l=1}^{L_1} \mathbf{b}^-_{1l} = \mathbf{x}^\emptyset$. By the same argument, $\exists \mathbf{b}^-_{21}, \ldots, \mathbf{b}^-_{2L_2}$ such that

1. $\mathbf{b}^-_{2L'} \in \mathscr{M}_2(\mathbf{x}_2 + \sum_{l=1}^{L'-1} \mathbf{b}^-_{2l})$ for $L' = 1\ldots, L_2$,
2. $\mathbf{x}_2 + \sum_{l=1}^{L'} \mathbf{b}^-_{2l} \in \mathscr{F}_y$ for all $L' = 1\ldots, L_2$, and
3. $e_t(\mathbf{x}_2 + \sum_{l=1}^{L_2} \mathbf{b}^-_{2l}) = 0$.

Moreover, $-\mathbf{b}_{2,L_2-l+1} \in \mathscr{M}_1(\mathbf{x}^\emptyset + \sum_{l=0}^{L'-1} -\mathbf{b}^-_{2,L_2-l})$ for all $L' = 1, \ldots, L_2$ by Lemma 1.2.2. Then the sequence $\mathbf{b}_1, \ldots, \mathbf{b}_L$ where $L = L_1 + L_2$, $\mathbf{b}_l = \mathbf{b}^-_{1l}$ for $l = 1, \ldots, L_1$ and $\mathbf{b}_{L_1+l} = -\mathbf{b}^-_{2,L_2-l+1}$ for $l = 1, \ldots, L_2$ satisfies the conditions of the proposition. Note that half of this argument suffices if either $\mathbf{x}_1 = \mathbf{x}^\emptyset$ or $\mathbf{x}_2 = \mathbf{x}^\emptyset$. □

Notations reminder

- $\mathscr{F}_y = \{\mathbf{x} : \mathbf{y} = \mathbf{A}\mathbf{x}\}$
- $\mathscr{M}(\mathbf{x})$: dynamic Markov base (i.e. set of moves available from $\mathbf{x}$),
- $\chi_{2,t}(\mathbf{x})$: set of histories that contain misidentifications at t and whose count in $\mathbf{x}$ is positive.
- $e_t(\mathbf{x})$: number of misidentifications at t in $\mathbf{x}$.

Proposition 1.2.6 *Let* $\mathbf{x} \in \mathscr{F}_y$. *If* $\mathbf{b} \in \mathscr{M}(\mathbf{x})$ *then* $P(\mathbf{x}^{(k+1)} = \mathbf{x} + \mathbf{b} | \mathbf{x}^{(x)} = \mathbf{x}) > 0$.

Proof. Suppose that $\mathbf{b} \in \mathscr{M}_1(\mathbf{x})$ and let $\mathbf{x}' = \mathbf{x} + \mathbf{b}$. Then $-\mathbf{b} \in \mathscr{M}_2(\mathbf{x}')$ by Lemma 1.2.2. Direct calculation of Equation 1.11 and Equation 1.12 shows that both $q(\mathbf{x}'|\mathbf{x}) > 0$ and $q(\mathbf{x}|\mathbf{x}') > 0$. Combined with Lemma 1.2.1 it follows that $P(\mathbf{x}^{(k+1)} = \mathbf{x}' | \mathbf{x}^{(k)} = \mathbf{x}) = r_1 > 0$ (from Equation 1.8). A similar argument shows that $P(\mathbf{x}^{(k+1)} = \mathbf{x} + \mathbf{b} | \mathbf{x}^{(k)} = \mathbf{x}) > 0$ for all $\mathbf{b} \in \mathscr{M}_2(\mathbf{x})$. □

We establish aperiodicity by showing that there is positive probability of holding at $\mathbf{x}^\emptyset$.

Proposition 1.2.7 *If* $\mathbf{x}^{(k)} = \mathbf{x}^\emptyset$ *then* $P(\mathbf{x}^{(k+1)} = \mathbf{x}^\emptyset) \geq 0.5$.

Proof. The set $\mathscr{M}_2(\mathbf{x}^\emptyset)$ is empty since there are no errors to remove from $\mathbf{x}^\emptyset$. However, the algorithm still proposes to draw a move from $\mathscr{M}_2(\mathbf{x}^\emptyset)$ with probability 0.5. When this occurs we set $\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)}$ so that $P(\mathbf{x}^{(k+1)} = \mathbf{x}^\emptyset) > 0.5$. □

This shows that $\mathbf{x}^\emptyset$ is an aperiodic state and hence that the entire chain is aperiodic (Cinlar, 1975, pg. 125 [85]). Since the fibre is finite, irreducibility and aperiodicty are sufficient to ensure that the chains have a unique stationary distribution which is also the limiting distribution (see Cinlar, 1975, Corollary 2.11 [85]). That this distribution is equal to the target distribution is guaranteed by the detailed balance condition of the MH algorithm which holds under Proposition 1.2.8 (Liu, 2004, pg. 111 [86]).

[85]: Çınlar (2013), *Introduction to stochastic processes*

[86]: Liu (2004), *Monte Carlo Strategies in Scientific Computing*

Proposition 1.2.8 *If* $q(\mathbf{x}'|\mathbf{x}) > 0$ *then* $q(\mathbf{x}|\mathbf{x}') > 0$ *for all* $\mathbf{x}, \mathbf{x}' \in \mathcal{F}_y$.

Proof. Suppose that $q(\mathbf{x}'|\mathbf{x}) > 0$. Then either $\mathbf{x}' - \mathbf{x} \in \mathcal{M}_1(x)$ or $\mathbf{x}' - \mathbf{x} \in \mathcal{M}_2(x)$. If $\mathbf{x}' - \mathbf{x} \in \mathcal{M}_1(x)$ then $\mathbf{x} - \mathbf{x}' \in \mathcal{M}_2(x')$ by Lemma 1.2.2 and $q(\mathbf{x}|\mathbf{x}') > 0$. Similarly if $\mathbf{x}' - \mathbf{x} \in \mathcal{M}_2(x)$ then $\mathbf{x} - \mathbf{x}' \in \mathcal{M}_1(x')$ and $q(\mathbf{x}|\mathbf{x}') > 0$. □

This complete Bonner et al. (2015)'s proof that the Markov chains produced by the algorithm of Section 1.2.4 have unique limiting distribution $\pi(\mathbf{x}, N, \mathbf{p}|\mathbf{y}, \alpha)$ so that realisations from the tail of a converged chain can be used to approximate properties of the joint posterior distribution of $\mathbf{x}$, N and $\mathbf{p}$.

1.3 Extending the Latent Multinomial Model

1.3.1 Additional notations

Parameters

- $\psi_{r,s}$: Probability that an individual transition from state r to state s between two consecutive occasions.
- δ_s: Probability that an individual is in state s at $t = 1$.

1.3.2 Multi-state capture-recapture model

The time-dependent multistate model assumes individuals to move independently over a finite set of S states, $E = \{e_1, ...e_S\}$. These states are not observed at each occasion for every individual but only when they are captured. Capture histories ω_i are now composed of $S + 1$ values. The $1, ..., S$ are used when the individuals are seen in states $e_1, ...e_S$ and the 0 when the individuals are not seen. We assume that the state is always correctly identified on capture. We now have $p_{s,t}$, the detection probabilities that vary both in time (denoted as before t) and in states (denoted s). We note

- $\psi_{s,r}$ the probability of being in state e_r at time $t + 1$ if in state e_s at time t (i.e. the transition probability),
- δ_s the probability of being in states e_s at $t = 1$.

To compute the probability of history ω_i, define

$$\pi_i^{(1)}(s) = \begin{cases} \delta_s(1 - p_{s,1}) & \text{if} \quad \omega_{i,1} = 0 \\ \delta_s(p_{s,1}) & \text{if} \quad \omega_{i,1} = s \end{cases} \tag{1.14}$$

Then for $t = 1, ..., T - 1$,

$$\pi_i^{(t+1)}(s) = \begin{cases} \left[\sum_{r=1}^{S} \pi_i^{(t)}(r)\psi_{r,s}\right](1 - p_{s,t+1}) & \text{if} \quad \omega_{i,t+1} = 0 \\ \left[\sum_{r=1}^{S} \pi_i^{(t)}(r)\psi_{r,s}\right] p_{s,t+1} & \text{if} \quad \omega_{i,t+1} = s \\ 0 & \text{if} \quad \omega_{i,t+1} = r \neq s \end{cases} \tag{1.15}$$

Note that $\sum_{s=1}^{S} \pi_i^{(t)}(s)$ is the probability of the history ω_i until time t. Then, the likelihood of history ω_i is

$$\pi_i = \sum_{s=1}^{S} \pi_i^{(T)}(s)\,. \tag{1.16}$$

As for the M_t model, conditioned on the population size, the vector $\mathbf{y}$ follows a multinomial with cell probabilities π_i:

$$[\mathbf{y}|N, \boldsymbol{\delta}, \boldsymbol{\psi}, \mathbf{p}] = \frac{N!}{\prod_i y_i!} \prod_i \pi_i^{y_i} \tag{1.17}$$

1.3.3 Multistate LMM

LMM likelihood reminder

Capture process:

$$[\mathbf{z}|N, \mathbf{p}] = \frac{N!}{\prod_k z_k!} \prod_k \pi_k^{z_k}$$

With single-state histories probabilities:

$$\pi_k = \prod_1^T p_t^{I(\xi_{k,t}=1)}(1 - p_t)^{I(\xi_{k,t}=0)}$$

Identification process:
If $I(\mathbf{z} = \mathbf{B}\mathbf{x})$, then

$$[\mathbf{x}|\mathbf{z}, \alpha] = \frac{\prod_k z_k!}{\prod_j x_j!} \prod_j \left[\prod_{t=1}^{T} A_{j,t}\right]^{x_j}$$

$$A_{j,t} = \alpha^{I(\nu_{k,t}=1)}(1 - \alpha)^{I(\nu_{k,t}=2)}$$

In Section 1.2.3 we broke the likelihood of the LMM into two components: the capture one and the identification one. Thanks to that we can modify each part independently. The global likelihood given by Equation 1.3 is still valid:

$$[\mathbf{y}, \mathbf{x}, \mathbf{z}|N, p, \boldsymbol{\delta}, \boldsymbol{\psi}, \alpha] = I(\mathbf{y} = \mathbf{A}\mathbf{x})\,[\mathbf{x}|\mathbf{z}, \alpha]\,[\mathbf{z}|N, \boldsymbol{\delta}, \boldsymbol{\psi}, \mathbf{p}] \tag{1.18}$$

For the detection part, the process is the same as for the multi-state capture-recapture model (Section 1.3.2). Thus, the likelihood is computed with Equation 1.4. The probabilities π_k are calculated using Equations 1.14 to 1.16 replacing observed histories ω by the latent capture histories ξ.

If we consider that the probability of correctly identifying an individual is the same for every state, then the likelihood doesn't change much compared to the single-state model. To account for possible misidentifications, latent error histories ν_j have to include other values to denote misidentifications on the different stages. They now include $2S + 1$ different values (0 for the unseen, S values for the S seen states and S values for misidentifications on the S states). There are $(2S + 1)^T$ latent error histories. The likelihood of the identification process is computed with Equation 1.5, rewriting $A_{j,t} = \alpha^{I(\nu_{j,t} \in [1,S])}(1 - \alpha)^{I(\nu_{j,t} > S)}$. For example, if three states are considered, the identification likelihood of latent history $(1, 4, 0, 2, 6)$ is

$$A_{(1,4,0,2,6)} = \alpha \times (1 - \alpha) \times 1 \times \alpha \times (1 - \alpha) = \alpha^2(1 - \alpha)^2$$

State heterogeneity could be considered for the identification process by setting $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_S)$. In that case the previous example likelihood would simply be

$$A_{(1,4,0,2,6)} = \alpha_1 \times (1 - \alpha_1) \times \alpha_2 \times (1 - \alpha_3)$$

1.3.4 Estimating parameters with multi-state observation

The algorithm to construct the MCMC is pretty much the same as in Section 1.2.4. We need to add samplers for the initial state probability $\boldsymbol{\delta}$ and for the transition probabilities $\boldsymbol{\psi}$. In addition, the way of proposing a $\mathbf{x}'$ must be adapted. The MCMC is constructed this way:

1. Let $\beta(a_0^t, b_0^t)$ denote the beta prior on p_t, $\beta(a_0^\alpha, b_0^\alpha)$ denote the beta prior on α, $Dir({\mathbf{a_0}}^\delta)$ the Dirichlet prior on δ and $Dir({\mathbf{a_0}}^{\psi_{s,\cdot}})$ the Dirichlet prior on $\boldsymbol{\psi}_{s,\cdot}$.

2. Initialize all parameters as well as a set of latent histories satisfying $\mathbf{y}$=$\mathbf{A}\mathbf{x}$. Such a set can be obtained by assuming that no mistakes were made. The latent frequencies of the histories containing errors are 0 and all the other match the observed frequencies one-to-one. In order to run several chains with different initialisations, one can take the previous initialisation of $\mathbf{x}$ and follow the later steps of (7) by only adding misidentification to the set and always accepting the proposed ones without going though the Metropolis-Hasting acceptance. In the initial latent set, fix the number of unseen individual to a random realistic number.

3. Sample the capture rate with Gibbs sampling as shown by Link et al. (2010) [74]. The likelihood being multinomial, it follows that the beta priors lead to full conditional beta posterior distribution:

   $$p_t|\mathbf{Z}, \boldsymbol{\psi}, \boldsymbol{\delta} \sim \beta(a_0^t + a^t, b_0^t + b^t)$$

   where a^t [6] is the number of captured individuals at time t and b^t [7] is the number of unseen individuals at time t (including those never seen).

   [74]: Link et al. (2010), 'Uncovering a Latent Multinomial'

   6: $a^t = \sum_n I(\xi_{n,t} > 0)$

   7: $b^t = \sum_n I(\xi_{n,t} = 0)$

4. Sample $\boldsymbol{\delta}$, the initial states probabilities with Gibbs sampling. It has a full conditional dirichlet posterior distribution.

   $$\boldsymbol{\delta}|\mathbf{z}, \mathbf{p}, \boldsymbol{\psi} \sim Dir({\mathbf{a_0}}^\delta + \mathbf{a}^\delta)$$

   where $\mathbf{a}^\delta$ is the number of individual in each state at occasion 1. For the unseen individual at occasion 1, the state they were in is unknown but can be sampled [8].

   8: This is done using the forward and backward algorithms to sample the initial state in its full posterior distribution. More details in Appendix B.1.

5. Sample the transition rates with Gibbs sampling. They have a full conditional beta posterior distribution.

   $$\psi_{s,\cdot}|\mathbf{z}, \mathbf{p}\boldsymbol{\delta} \sim Dir({\mathbf{a_0}}^{\psi_{s,\cdot}} + \mathbf{a}^{\psi_{s,\cdot}})$$

   where $\mathbf{a}^{\psi_{s,\cdot}}$ is the number of of times an individual transitioned from state s to the others. Just like for the initial states, we cannot know what transition occurred for an unseen individual but a transition can be sampled. [9].

   9: This again is done using the forward and backward algorithms to sample the realised transitions in their full posterior distribution. More details in Appendix B.2.

6. Sample the identification rate using Gibbs sampling. Similar to the capture rate, it has a full conditional beta posterior distribution:

   $$\alpha|\mathbf{x} \sim \beta(a_0^\alpha + a^\alpha, b_0^\alpha + b^\alpha)$$

   where a^α [10] is the total number of correct identifications and b^α [11]

   10: $a^\alpha = \sum_j \sum_t x_j I(\nu_{j,t} \in [1, S])$

   11: $b^\alpha = \sum_j \sum_t x_j I(\nu_{j,t} = \in [S+1, 2S])$

is the total number of misidentifications.

7. Sample jointly N and $\mathbf{x}$ since the number of errors in $\mathbf{x}$ changes the population size.
   a) Define:
      - $\boldsymbol{\nu}^{(1st)}$ the history with a unique capture at occasion t in state e_s (potential ghost),
      - $\chi_{0,s,t}(x) = \{\boldsymbol{\nu} | \nu_t = 0, x_\nu > 0, x_{\nu^{(1st)}} > 0\}$ the set of histories having *potentially* generated a ghost in state e_s at occasion t, for the given $\mathbf{x}$,
      - $\chi_{2,s,t}(x) = \{\boldsymbol{\nu} | \nu_t = s + S, x_\nu > 0\}$ the set of histories *containing* a ghost in state e_s at occasion t, for the given $\mathbf{x}$.
   b) Sample a state s uniformly from $1, ..., S$.
   c) With probability 0.5, go to (i), otherwise go to (ii).
      i. Add a misidentification (i.e. a ghost) to the latent set.
         - Sample $\boldsymbol{\nu}^{(0)} \in \chi_{0,s,.}(x) = \bigcup_t \chi_{0,s,t}(x)$.
         - Sample $t \in \{t | \boldsymbol{\nu}_t^{(0)} = 0, x_{\nu^{(1st)}} > 0\}$.
         - Set $\boldsymbol{\nu}^{(2)} = \boldsymbol{\nu}^{(0)}$ and then $\nu_t^{(2)} = s + S$.
         - Define the move $b_{\boldsymbol{\nu}^{(0)}, \boldsymbol{\nu}^{(1st)}, \boldsymbol{\nu}^{(2)}} = (-1, -1, +1)$ and $b_\nu = 0$ for all other latent histories.
      ii. Remove a misidentification from the latent set.
         - Sample $\boldsymbol{\nu}^{(2)} \in \chi_{2,s,.}(x) = \bigcup_t \chi_{2,s,t}(x)$.
         - Sample $t \in \{t | \boldsymbol{\nu}_{s,t}^{(2)} = S + s\}$.
         - Define $\boldsymbol{\nu}^{(0)} = \boldsymbol{\nu}^{(2)}$ and then $\nu_t^{(0)} = 0$.
         - Define the move $b_{\boldsymbol{\nu}^{(0)}, \boldsymbol{\nu}^{(1st)}, \boldsymbol{\nu}^{(2)}} = (+1, +1, -1)$ and $b_\nu = 0$ for all other latent histories.
   d) Define $\mathbf{x}' = \mathbf{x}^{(k-1)} + b$.
   e) Calculate $\mathbf{z}' = \mathbf{B}\mathbf{x}'$ and $N' = \sum \mathbf{x}'$.
   f) With probability r_1, set $\mathbf{x}^k = \mathbf{x}'$, $\mathbf{z}^k = \mathbf{z}'$ and $N^(k) = N'$. Otherwise set $\mathbf{x}^k = \mathbf{x}^{k-1}$, $\mathbf{z}^k = \mathbf{z}^{k-1}$ and $N^(k) = N^{(k-1)}$.

$$r_1 = min\left(1, \frac{[\mathbf{y}, \mathbf{x}', \mathbf{z}' | N', p, \boldsymbol{\psi}, \boldsymbol{\delta}, \alpha]}{[\mathbf{y}, \mathbf{x}^{(k-1)}, \mathbf{z}^{(k-1)} | N, p, \boldsymbol{\psi}, \boldsymbol{\delta}, \alpha]} \frac{q(\mathbf{x}^{(k-1)} | \mathbf{x}')}{q(\mathbf{x}' | \mathbf{x}^{(k-1)})}\right). \quad (1.21)$$

8. Sample the number of unseen individuals x_1:
   a) sample a move $c \in [-D, D]$ where D is fixed integer,
   b) define $x_1' = x_1 + c$,
   c) if $x_1' \geq 0$, accept it with probability r_2 [12]:

$$r_2 = min\left(1, \frac{[\mathbf{z}' | N', p, \boldsymbol{\psi}, \boldsymbol{\delta}]}{[\mathbf{z} | N, p, \boldsymbol{\psi}, \boldsymbol{\delta}]}\right). \quad (1.22)$$

9. repeat steps 3 to 6 as much as needed.

In the case where we consider state heterogeneity in identification probability, after having updated $\boldsymbol{\alpha}$ from its full conditional posterior, the sampling step of $\mathbf{x}$ should be repeated once for each state instead of just once for a random state.

Only a few details change the proof of convergence of the algorithm compared to Section 1.2.5. The first condition becomes that Step 3 to 5 and 7 produces chains which converge to $\pi(N, \mathbf{p}, \boldsymbol{\delta}, \boldsymbol{\psi} | \mathbf{z})$ for any $\mathbf{z}$ such

Proposal density

The proposal densities are calculated by multiplying the probabilities of each sampling step used for defining the move. They are successively: the probability of adding (or removing) an error, the probability of sampling the state, the probability of choosing the ν_0 (or ν_2) and the probability of choosing the t knowing the sampled ν. When adding an error, the proposal density q is:

$$q(x' | x^{k-1}) = \frac{0.5/S}{\#\chi_{0.}\#\{t | \nu_{0,s,t} = 0, x_{\nu_{1st}} > 0\}} \quad (1.19)$$

and when removing an error, is:

$$q(x' | x^{k-1}) = \frac{0.5/S}{\#\chi_{2,s,.}.\#\{t | \nu_{2,s,t} = 2\}} \quad (1.20)$$

where $\#S$ denotes the cardinality of S.

12: r_2 takes the same form as r_1 but only the number of unseen individuals changes and the proposal density is symmetric. Thus the ratio parts involving $[\mathbf{x}|\mathbf{z}, \alpha]$ and $q(x' | x^{k-1})$ simplify to 1 and we are left with Equation 1.22.

that $\mathbf{z} = \mathbf{B}\mathbf{x}$ for some $\mathbf{x} \in \mathcal{F}_y$. The validity of the condition is as trivial as it was in the original proof. The second and third conditions do not change, and their proof remain the same.

1.4 Simulation studies

1.4.1 Simulation design for single-state model

Link et al. [74] have shown that the $M_{t,\alpha}$ model was effective on one simulation with 5 capture occasions, $\alpha = 0.9$ and $\mathbf{p} = (0.3, 0.4, 0.5, 0.6, 0.7)$ over a population of 400 individuals. Vale et al. tested it for numbers of occasions between 4 and 12, capture rates between 0.05 and 0.5, identification rates between 0.9 and 0.99 and with population size of 400 and 1000. They show that for low recapture rate the parameter α becomes unidentifiable. Thus we expect the model to be weakly identifiable for simulations with low capture rate, making the posterior density unidentifiable. Gartett and Zeger (2000) [87] define weak identification as the situation where the technical conditions for identifiability are met but the data provides little information about the particular parameters so that their posterior and prior distributions are similar. Cole and Mc Crea (2016) [88] say that using informative priors can result in an identifiable posterior when the model is weakly identifiable. We ran the model using three different priors for parameter α.

[74]: Link et al. (2010), 'Uncovering a Latent Multinomial'

[87]: Garrett et al. (2000), 'Latent class model diagnosis'

[88]: Cole et al. (2016), 'Parameter redundancy in discrete state-space and integrated models'

Thus, we will test the model Mt, α on parameters range similar as Vale et al. (2014) [77] and compare the result for three different priors. We simulated observation data for 5, 7 and 9 occasions, on population size of 500 and 1000 with identification rate of 0.8, 0.9 and 0.95 and capture rates of 0.1, 0.2, 0.3 and 0.4. It makes 24 parameters combinations for each of the three different number of occasions. For the sake of simplicity, we considered the time-dependent $M_{t,\alpha}$ model, even though the capture rate was held constant over time in the simulations. For the priors, the first one is a non-informative Beta prior. The other two are informative such as might have been obtained through an evaluation of the identification protocol. Assume that the protocol is run on n known individuals and results in n_a correct identifications and n_b errors. The prior is then $\alpha \sim \beta(n_a, n_b)$. We used $n = 100$ because it is a convenient value to use and it is very close to the capacity of a 96-well PCR plate. The first informative prior was centered on the value used for simulation: for α simulated at 0.8, we have $\alpha \sim \beta(80, 20)$, for α simulated at 0.9, $\alpha \sim \beta(90, 10)$ and for α simulated at 0.95, $\alpha \sim \beta(95, 5)$.
The second informative prior is similar to the first one but it is centered on a value that underestimates α. It is represented on Figure 1.2. We chose to underestimate alpha because the model has a tendency to do as such when the capture rate gets too low. By reproducing the observed bias in the prior, we test the sensitivity to the prior in an unfavourable case.. The values of n_a and n_b were chosen such that the true value used for the simulation lies around the 95th percentile of the prior distribution (dashed line on Figure 1.2). They are as following: $\alpha_{simulated(0.8)} \sim \beta(74, 26)$, $\alpha_{simulated(0.9)} \sim \beta(85, 15)$ and $\alpha_{simulated(0.95)} \sim \beta(91, 9)$. These priors have respective means of 0.74, 0.85 and 0.91. In order to study the effect of the prior on α over the model, we calculated the overlap τ between this

prior and the estimated posterior as suggested by Garrett and Zeger (2000)[87].

[87]: Garrett et al. (2000), 'Latent class model diagnosis'

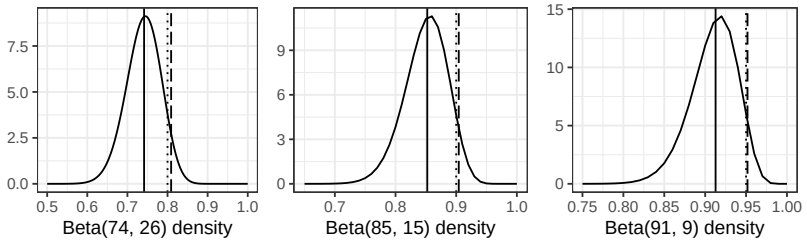

Figure 1.2: Beta densities for biased priors on identification probability for the three values used in simulations. The dashed line represents the 95th percentile, the black line the median of the prior and the dotted line the true value of the simulation.

1.4.2 Simulation design for multistate model

For the multistate model, we tested the estimation of population size with the same design as for single-state was used (i.e. for the same capture and identification probability scenarios). We considered three states with possibility of transition between all states. For the sake of comparison, the transition matrix used is taken from Worthington et al. (2019) [89] as:

[89]: Worthington et al. (2019), 'Estimation of population size when capture probability depends on individual states'

$$\boldsymbol{\phi} = \begin{pmatrix} 0.76 & 0.12 & 0.12 \\ 0.1 & 0.8 & 0.1 \\ 0.15 & 0.15 & 0.7 \end{pmatrix}$$

and the initial states are fixed to its equilibrium distribution, that is $\boldsymbol{\delta} = (0.33, 0.4, 0.27)$.

Additionally, we tested the impact of errors on the estimates of transition rates. For that part, we made ten simulations for each of the four following scenarios:

	constant probabilities	state dependant probabilities
1	identification, capture	-
2	capture	identification
3	identification	capture
4	-	identification, capture

When the capture was state dependent, it was simulated as $\mathbf{p} = (0.3, 0.4, 0.5)$. When the identification was state dependent, it was simulated as $\boldsymbol{\alpha} = (0.8, 0.9, 0.95)$. Otherwise they were $p = 0.4$ and $\alpha = 0.9$. The transitions were the same as previously. For each of the scenarios, we ran the model M that does not take into account the misidentification and the model M_α that incorporates misidentifications. The models were adapted to the state dependence of the simulations.

1.4.3 Implementation

We used NIMBLE [80] to implement the model. Unlike Jags (for example),

[80]: Valpine et al. (2017), 'Programming With Models'

NIMBLE allows new distributions as well as all samplers for the MCMC to be written as we need. We needed it to code the likelihood of the model and to code the sampler of **x**. We were also able to write all the Gibbs samplers previously detailed for a maximum computational efficiency. In order to improve efficiency, all observable histories which had zero count were not considered, i.e. their corresponding rows and columns in matrices **A** and **B** were deleted as suggested in Schofield & Bonner (2015) [78]. For the single-state simulations, the MCMC was run over 1E6 iterations after a burn-in period of 20,000 iterations (30,000 for $\alpha = 0.8$) and the chains were thinned by a factor of $1/200$ in order to limit memory usage. For the multistate simulations, the computational cost per iteration is much higher so we only ran 500,000 iterations with a thinning of $1/100$ and an additional burn-in of 60,000 iterations. For most simulation scenarios, this proves to be enough. For simulations where $T = 5$ as well as where $T = 7$, $p \leq 0.2$, we instead had to use more iterations. We used 1E6 iteration with a thinning of $1/200$ and an additional burn-in of 100,000 iterations. We ran two chains for each simulation with two different starting points. For the first one, **x** was initialised as the set of observed histories, as if there was no error. In the second one, we arbitrarily added 40 errors randomly.

[78]: Schofield et al. (2015), 'Connecting the latent multinomial'

1.4.4 Results single-state

Running two chains with 1,030,000 iterations on a 3.0GHz Intel processor took less than five minutes, even with $T = 9$. Convergence was assessed by looking at the N chains. It is necessary as the $\hat{R}$ can be under 1.1 and the effective sample size over 100 while the chains clearly show convergence problems. With the uninformative prior on α, convergence was achieved for all simulations with a capture rate of 0.3 or above. For $T = 5$, with $p = 0.2$ some chains did not converge while with $p = 0.1$ none did. Increasing T to 7 did allow for a better convergence with $p = 0.2$ but not with $p = 0.1$. Finally, $T = 9$ resulted in good convergence for more than half the simulations with $p = 0.1$. In addition, convergence was slower for lower values of α and for $N = 1000$. There is a high autocorrelation for the N-chains that makes some of them have an effective sampling size less than 100. When an informative prior on α is used, the chains always converge and the effective sampling size is always over 75 (average is over 200).

The population size estimation with a single-state and $N = 500$ is shown in Figure 1.3. Using the uninformative prior, no bias was observed for $p \geq 0.3$. When $p = 0.2$, the average relative bias goes from 3% (when T=9) to 14% (when T=5). When convergence was reached for simulations with $p = 0.1$, the average relative bias was over 30% when T=9 and over 40% when T=7.
When adding the unbiased prior, for $p = 0.1$, the population size is underestimated by about 10% on average but this bias rises to 40% for some simulations. Also, for 80% of the simulations with $p = 0.1$, the real population size lies in the estimated 95% interval. The use of the biased prior does not affect the estimations for $p = 0.4$. But as p decreases, the population size gets more underestimated.
At the lowest, the average bias goes down to 32% for the lowest values of p and α with 9 capture sessions. Higher values of α lead to a reduced

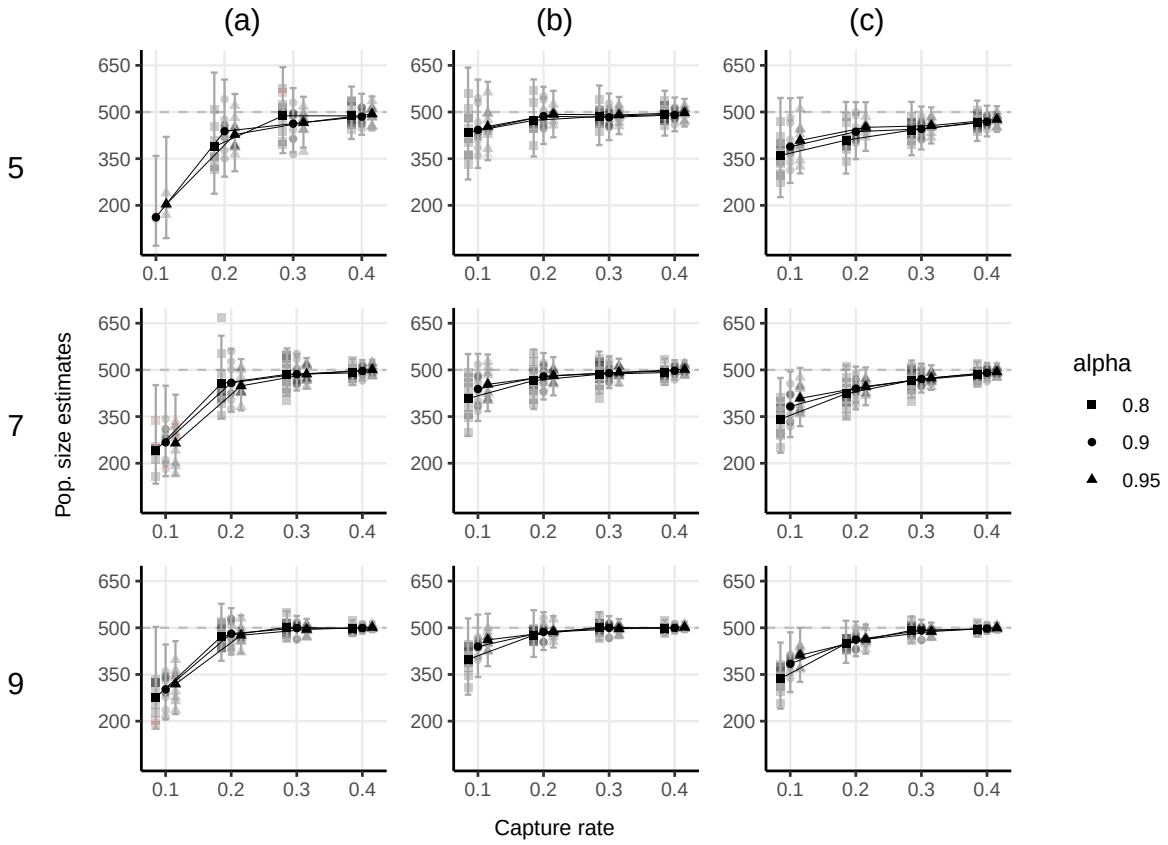

Figure 1.3: Single-state population size estimations (y axis) depending on capture probability (x axis), identification probability and number of capture occasion (on the left). Columns are for various priors on the identification probability, (a) uninformative, (b) informative centered on true value, (c) informative centered on a lower value. Horizontal dashed lines indicate true population size. Grey and red symbols show simulation-specific estimates of the population-size posterior mean, red ones indicating that the N chains had $\hat{R} > 1.1$ or an effective sample size under 50. Black symbols connected by lines and error bars show, respectively, the estimates of the mean and the 95% credible intervals of the posterior distribution of population-size averaged across simulations.

bias when it occurs and a reduced confidence interval. The results are very similar for $N = 1000$ only slightly better.

When looking at the overlaps between a prior and a posterior, Garrett et al. (2000) [87] give the value of 0.35 as a guide, over which a model is weakly identified. We show the overlaps between prior and posterior of α in Figure 1.4. With the uninformative prior, all simulations with $p \geq 0.3$ and most of the ones with $p = 0.2$ result in an overlap between prior and posterior for α that is lower than 0.35. With the informative priors, for most of the simulations, the prior and posterior of α are highly overlapping and almost confounded for low recaptures. The informative priors overlap less with their corresponding posterior for $p \geq 0.3$.

[87]: Garrett et al. (2000), 'Latent class model diagnosis'

1.4.5 Results multi-state, population size estimates

Running two chains of 1,100,000 iterations, on the same processor as for single-state, took around 4 hours for $T = 9$. With the uninformative prior on α, almost no chain converged when $p = 0.1$. Where $T = 5$ and $p = 0.2$, the MCMC converged for around half of the simulations. Convergence was reached for all the other scenarios. For the scenarios where convergence was not reached, adding an informative prior led to

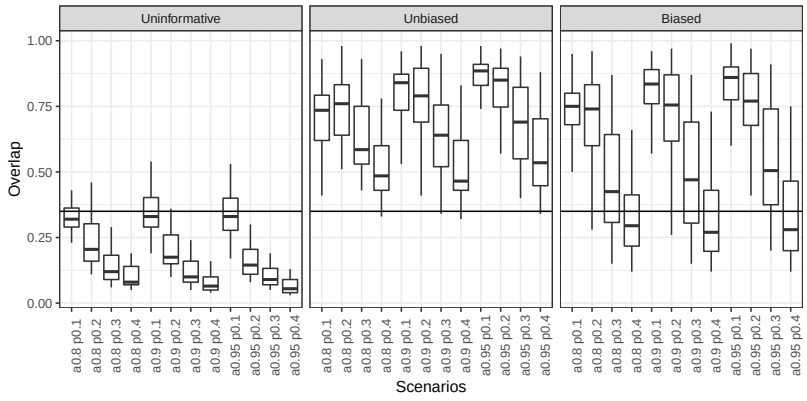

Figure 1.4: Boxplots of the overlap value beween the prior and posterior of the identification probability. The horizontal line is at 0.35 (see [87]). On the x-axis legend, the letter 'a' stands for the identification probability and the letter 'p' for the capture probability, the corresponding values simulated following.

proper convergence. Some more iterations are needed for $N = 1000$ as a lot of chains have an effective sampling size under 100.

Multistate population size estimation for $N = 500$ are shown on Figure 1.5. Using the uninformative prior, no bias is observed for $p \geq 0.4$. For $T = 5$, the estimates are biased as soon as $p \leq 0.3$. The average relative bias ranges from 10% (for $p = 0.3$, $\alpha = 0.95$) to 50% (for $p = 0.2$, $\alpha = 0.8$). When $T = 7$, the estimates are slightly biased (5% at most) for $p = 0.3$. Results show more bias for lower capture rates, bias ranging between 16% and 30% for $p = 0.2$. When $T = 9$, the estimates are biased only for $p \leq 0.2$, bias ranging between 9% and 14%. When adding the unbiased prior, the average relative bias is reduced. For $p = 0.2$, $\alpha = 0.8$, it is reduced to 10% for $T = 9$ and to 17% for $T = 7$ and $T = 5$. The results for $N = 1000$ are similar although the bias is reduced.

The estimations of transitions probabilities are globally unbiased for $p \geq 0.3$ or for $T = 9$. Some transitions have an average bias that is always under 0.1. The relative bias can be quite high for low probability transition but the estimation always lies in the 95% interval. For $p = 0.2$ the size of this interval is around 0.4, the estimates are thus very imprecise. Finally adding an informative prior on α does not change the estimates of the transitions probabilities nor the size of the estimated intervals.

1.4.6 Results multi-state, transition estimates

The estimates of states transition rates for all scenarios are shown in Figure 1.6. When both the capture and identification probabilities are constant, no bias is observed and there is no difference in the estimates between models M and M_α. When only the identification probability is state dependent, if the misidentifications are ignored, we only see a small bias for the probability to stay in state one, the one that had the more misidentifications. A greater bias is observed if the capture probability depends on the state and model M is used. Transitions towards the least observed state (the 1) are overestimated while transitions towards the most observed state (the 3) are underestimated. If both probabilities depend on the state, the bias is amplified. Transitions toward state 1 (least observed state and higher rate of misidentifications) are systematically overestimated while it is the contrary for state 3. In all scenarios where

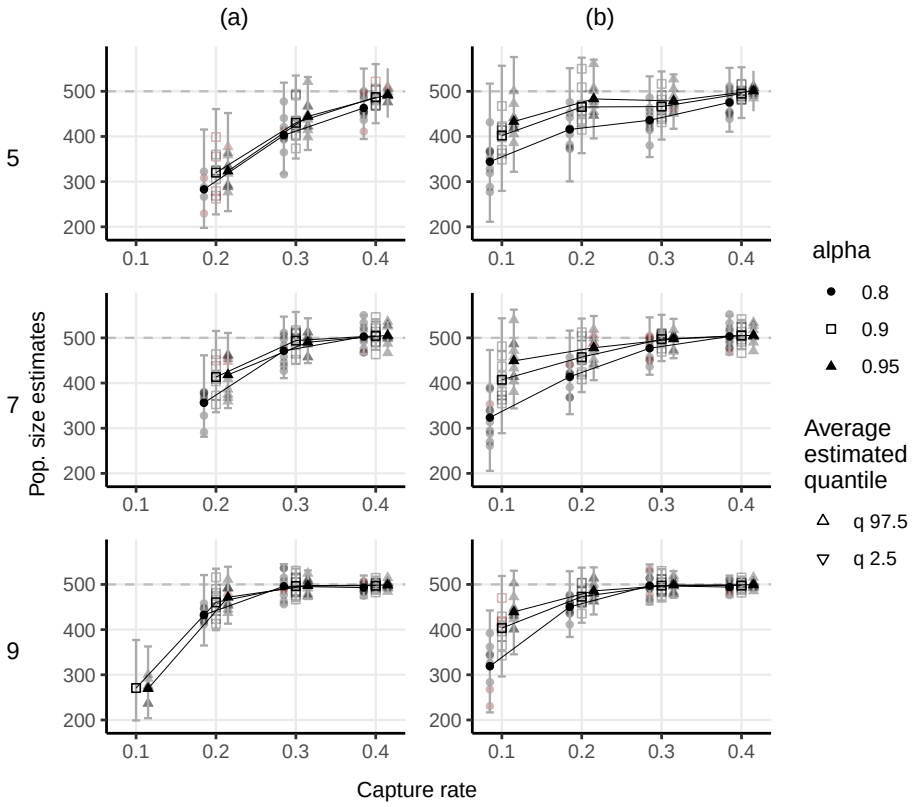

Figure 1.5: Multistate population size estimations (y axis) depending on capture probability (x axis), identification probability (point shape) and number of capture occasions (on the left). Columns are for various priors on the identification probability, (a) uninformative, (b) informative centered on true value. Grey and red symbols show simulation-specific estimates of the population-size posterior mean, red ones indicating that the N chains had $\hat{R} > 1.1$ or an effective sample size under 50. Black symbols connected by lines and error bars show, respectively, the estimates of the mean and the 95% credible intervals of the posterior distribution of population-size averaged across simulations.

bias was observed when using model M, the use of model M_α led to unbiased estimates.

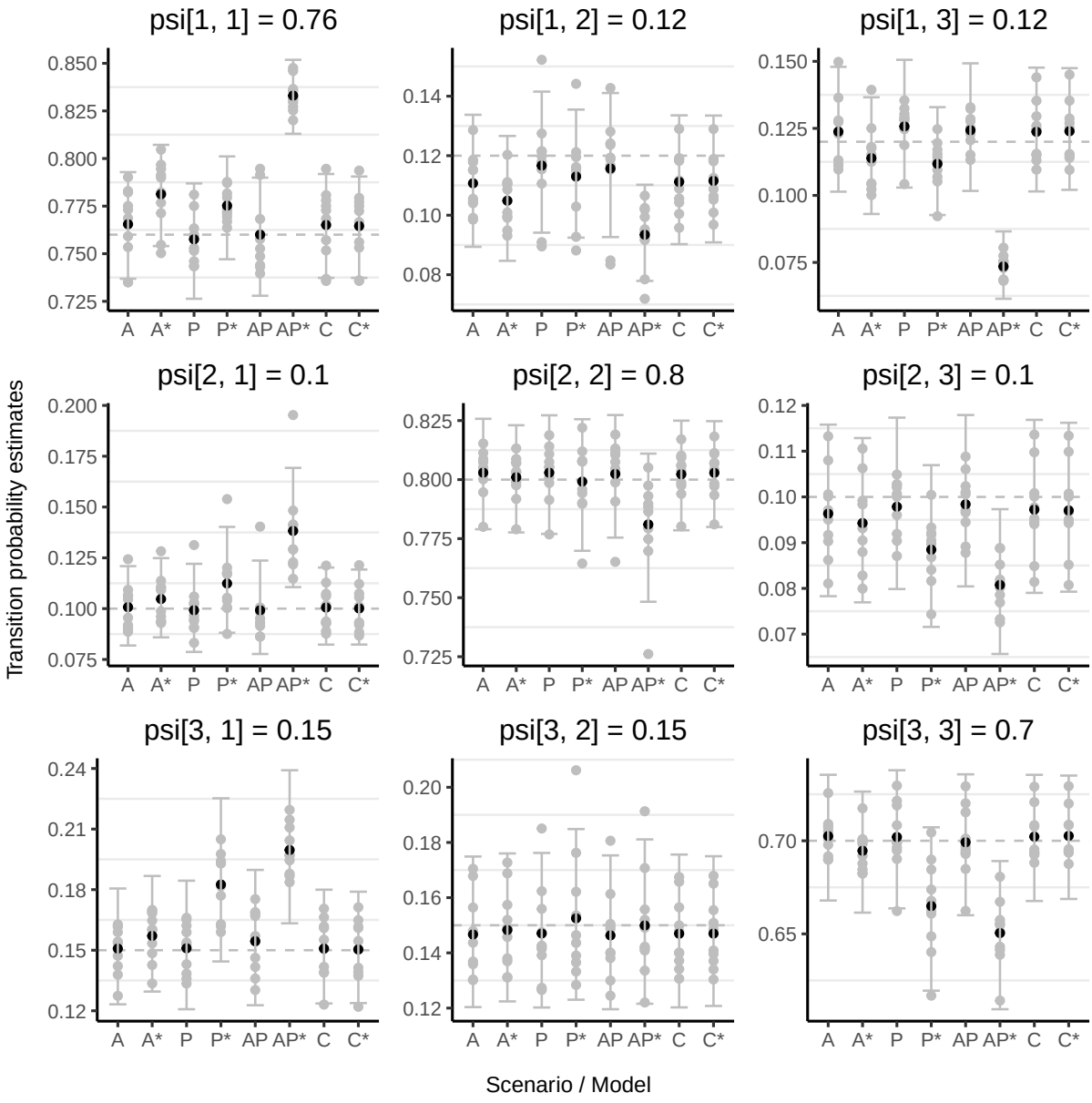

Figure 1.6: Estimates of states transition rates depending of state dependence and model used. The scenarios are named as follow: A stands for alpha state dependent, P for capture state dependant, AP for both state dependant and C for both constant. The star indicates that misidentifications were ignored. The black dots are the average estimates, the error bar are the limits of the average 95% interval.

1.5 Discussion

We conducted a simulation analysis to help design CMR experiments where low-quality DNA is to be used for identification and where low capture rates are expected. We have shown, in single and multi-state experiments, the range of parameters over which the LMM can be safely used to estimate population size and transition probabilities in a closed population. For experiments with fewer occasions than those tested here, a higher capture probability must be achieved. For four occasions, some simulations suggested that a capture probability around 0.4 was a minimum for good estimates. For three occasions, a capture probability of around 0.5 could lead to either good estimates or high bias, while 0.6 seems to lead to more reliable estimates. We have also shown how transition estimates are biased by misidentification when there is state heterogeneity in the probability of capture or identification. We have shown how the use of the LMM allows for good estimates. If transition rates are the main interest of a study and there is evidence of state dependence of the capture and identification probabilities, then it is necessary to model the potential misidentifications. The greater the differences between states in the capture and identification probabilities, the greater the bias will be.

When the capture rates and the number of capture occasions are too low, the model is weakly identified. Carlin and Louis (1996) [90] say that, in this case, there is a high cross-correlation between the parameters which leads to very slow convergence. As the probability of identification α decreases, this problem is amplified and the estimates become less precise. This demonstrates that the use of the LMM does not completely solve the identification problem, but should be used in parallel with experimental error reduction. Although the use of an informative prior does not guarantee the identifiability of a weakly identified model, this appears to be the case for our simulations since convergence is always achieved when using one, even a biased one. Considering that the priors we used were highly informative, in cases with low recaptures, where the data do not inform on α, it may seem reasonable to remove the parameter by fixing its value, rather than trying to estimate it. Finally, sensitivity to this prior should be tested, since with enough captures the biased prior will lead to slightly biased estimates compared to using an uninformative prior. An important consideration in the decision to use such a prior is the cost involved. If it is more expensive to test the identification protocol on known individuals than it is to increase the capture effort, then it may not be worthwhile. The limiting factor is having access to known individuals from which we can collect DNA. For very rare species, this may not even be possible.

[90]: Carlin et al. (1996), *Bayes and empirical Bayes methods for data analysis*

In this chapter, we implemented the LMM of Link et al. (2010) [74] using Nimble and the sampling algorithm of Bonner et al. (2015) [79]. This allowed for much faster MCMC than what Link et al. (2010) [74] reported. This work is a first step toward the accessibility of the model at a larger scale. The nimble implementation will allow others to use the model for their own needs. However, for cases different from what has been presented here, some changes must be done, both in the model code and in the samplers code.

[74]: Link et al. (2010), 'Uncovering a Latent Multinomial'

[79]: Bonner et al. (2015), 'Extending the latent multinomial model with complex error processes and dynamic markov bases'

It is important to keep in mind that the model assumes that ghosts can only be seen once. This hypothesis may not be true in some cases, so the model is not applicable to them. It is also useful to note that the framework of the LMM is not limited to closed populations and can be modified to estimate survival. This is achieved by replacing the likelihood of the capture process $[\mathbf{z}|N, \mathbf{p}]$ with the likelihood of an open population model, such as the Cormack-Jolly-Seber (CJS) model $[\mathbf{z}|\phi, \mathbf{p}]$. Bonner et al. (2015) [79] developed such a model with a different kind of misidentification (an individual is misidentified as another one that has been seen at least once before) and we are currently working on a multistate open population with misidentifications such as in this paper. The model can also be extended using data augmentation, in order to account for capture heterogeneity between individuals as in McClintock et al. (2014) [75].

[79]: Bonner et al. (2015), 'Extending the latent multinomial model with complex error processes and dynamic markov bases'

[75]: McClintock et al. (2014), 'Probit models for capture-recapture data subject to imperfect detection, individual heterogeneity and misidentification'

In extreme cases where $p \approx 1$, the sampling efficiency of misidentifications (step 5 in Section 1.2.4) can be very low. This is because when sampling a history possibly responsible for a misidentification (containing a 0), the probability of sampling another ghost (i.e. history with a single capture) is high. Nevertheless if such a problem were to occur, it is very likely that keeping only good quality data would be a viable option, hence making the LMM unnecessary.

For studies using low-quality DNA in order to identify individuals, the simulation study in this chapter shows that more samples could be kept or even collected. The LMM makes it possible to allow for about 5 to 10% of misidentifications and have good estimates of the parameters. Higher error rates are not a real problem, but the uncertainty in the estimates will be much higher. A low capture probability can be compensated for if prior information about the misidentifications is available. The LMM is especially promising for the study of large populations or very elusive species, where increasing the capture effort could then be expensive compared to keeping samples. In addition, there is potential for new experiments where lower quality samples can be obtained, provided DNA can be sampled. An example of such a study would be on insects such as mosquitoes, as in the project that motivated this paper.

Chapter two

Using a covariate of misidentification in closed population

2

2.1 Aim of the chapter

Experiments using low-quality DNA for identification of the individuals involve a series of distinct steps. First, samples are collected, from which DNA is extracted and amplified according to a given panel, made such that genotypes should be unique. Finally, individuals are genotyped based on the amplified DNA. Traditionally, the identification panel has relied on microsatellites. Amplification of DNA by Polymerase Chain Reaction (PCR) or quantitative PCR allows for the resolution of the alleles through electrophoresis. However, a recent and noteable shift is occurring, as microsatellites are being replaced by Single Nucleotide Polymorphism (SNP) markers. SNP genotyping is primarily conducted through SNP chips or next-generation sequencing (NGS). SNP chips, are microarrays with thousands to millions of specific SNP probes attached to their surface. These chips allow high-throughput analysis, enabling the simultaneous genotyping of a large number of SNPs from a single sample. The choice between microsatellites and SNPs as marker types profoundly influences the processing of samples and consequently on the resulting data and the underlying assumptions made about them.

Assessing the quality of genotyping data is critical to ensuring the reliability of the results. An important measure of data quality is the call rate, which assesses the percentage of successfully genotyped loci in a sample. A high call rate indicates that a large proportion of loci have been successfully genotyped, reflecting a more complete data-set. For PCR-based genotyping, the reproducibility of genotypes across replicates can be used as an indicator of data quality. Consistent genotype calls across replicates provide confidence in the accuracy of the results, whereas discrepancies between replicates may indicate potential problems. An example of a measure of quality when using multitube PCR is given by Miquel et al. (2006) [91].
For next-generation sequencing (NGS), the depth of coverage is a critical parameter for assessing data quality. A higher coverage depth ensures that more reads are allocated to a locus, increasing the confidence in genotype calls. Conversely, low coverage depth can introduce uncertainty and false positives or negatives in genotype identification. NGS technologies typically provide a measurement of the sequencing quality, such as the Q-score for Illumina [92].
Both measurement cited above are continuous values. The quality measurement given by Miquel et al. (2006) [91] is comprised between 0 and 1. The Q-score output from Illumina sequencing is a log probability. In this chapter, we will consider a quantitative continuous covariate.

[91]: Miquel et al. (2006), 'Quality indexes to assess the reliability of genotypes in studies using noninvasive sampling and multiple-tube approach'

[92]: Illumina (2011), 'Quality scores for next-generation sequencing'

In addition to the genotypes and quality measures, the relatedness between the genotypes of all samples is also calculated. Information about quality and relatedness are used to construct the capture histories. In the context of the LMM, the quality is currently used to screen out data of too low quality and remove the samples that would break the

misidentification hypothesis (ensuring that the only misidentifications are the creation of ghosts). The relatedness between all genotypes is used to match those from the same individuals. However this information is only used to construct the capture histories and is not used in the model to detect potential ghosts. In this chapter, I propose a new model that integrates the quality data as a covariate.

McClintock et al. (2014) [75] developed a probit model based on the LMM that allows for individual heterogeneity in either the capture probability or the identification probability. This model can account for individual covariates in the identification probability, but it doesn't allow the use of a covariate that is not at the individual level, such as the quality of an observed genotype. However, the genotyping quality of each sample is a factor that greatly affects the probability of correctly identifying which individual the sample comes from. Therefore, in this chapter I extend the LMM to include the quality of the genotypes used to identify the individuals. This should allow for a better understanding of which history with a unique capture may be the result of an error and improve the estimates.

[75]: McClintock et al. (2014), 'Probit models for capture-recapture data subject to imperfect detection, individual heterogeneity and misidentification'

2.2 A probit extension of the LMM

2.2.1 Additional notations

Parameters

- $\alpha_{n,t}$: Probability that the individual n is correctly identified at t
- $\theta_\alpha = (a, b)$: regression parameters of the probit model.

Data and Latent variables

- **Y**: set of observed histories
- **X**: set of latent error histories
- **Z**: set of latent capture histories

2.2.2 The probit model

The $M_{t,\alpha}$ model presented earlier uses only capture histories: i.e. the data are histories composed of 1s and 0s (0: not observed, 1: observed). If there are multistate observation, histories are composed of as many number as there are states plus the 0. For simplicity reasons, we will present the model here for single state observation. In addition to the histories, we now have a quality measurement for each sample. This will be used as a covariate for identification probability. Compared to the previous model, the use of a covariate involves some changes. The probability of correct identification α is no longer constant. Thus, we use the detailed matrices of histories **Y**, **X** and **Z**. Note that the notations **x** and **z** are still used to account for the number of matrices identical, with the exception of individual labelling. We define $\alpha_{n,t}$ as the probability of correctly identifying individual n at time t, given that it was captured at that time. If n was not captured at time t, then $\alpha_{n,t}$ is not needed. For the sake of completion, in these cases we set $\alpha_{n,t} = 1$ [1]. We note $\boldsymbol{\alpha} = (\alpha_{n,t})_{n\in[1,N],t\in[1,T]:\nu_{n,t}>0}$, the vector of identification probabilities

1: As such, it does not change the likelihood to multiply by $\alpha_{n,t}$

associated with each realised capture. In order to model $\alpha_{n,t}$ as a function of a covariate, we introduce $\boldsymbol{\theta}_\alpha$, the set of parameters defining $\boldsymbol{\alpha}$. We write $[\boldsymbol{\alpha}|\boldsymbol{\theta}_\alpha]$. Since α depends on n, we refer to this model as $M_{t,\alpha n}$.

Following the above, the likelihood is as follows:

$$[\mathbf{Y}, \mathbf{X}, \mathbf{Z}|N, \mathbf{p}, \boldsymbol{\alpha}, \boldsymbol{\theta}_\alpha] = I(\mathbf{Y}|\mathbf{X})[\mathbf{X}|\mathbf{Z}, \boldsymbol{\alpha}][\mathbf{Z}|N, \mathbf{p}][\boldsymbol{\alpha}|\boldsymbol{\theta}_\alpha]. \tag{2.1}$$

We then modify the various parts of the likelihood.

- $I(\mathbf{Y}|\mathbf{X})$

Let's define a function f such that, for a latent error history v_j, it results in the corresponding set of observed histories (ω_i): $f(v_j) = (\omega_i)$. For example, $f((1,1,2)) = \{(1,1,0),(0,0,1)\}$. If we apply f to all the latent histories in $\mathbf{X}$, the resulting set of histories (ω_i) must be equal to $\mathbf{Y}$, except for index inversions. An example is given on Figure 2.1. Thus, $I(\mathbf{Y}|\mathbf{X})$ is 1 if $f(\mathbf{X}) = \cup_j f(v_j) = \mathbf{Y}$ and 0 otherwise. We write this as $I(\mathbf{Y} = f(\mathbf{X}))$.

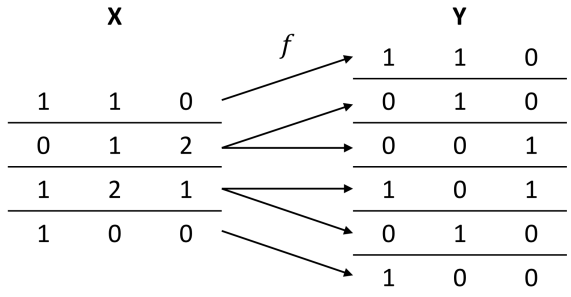

Figure 2.1: Example of $f(\mathbf{X}) = \mathbf{Y}$

- $[\mathbf{X}|\mathbf{Z}, \boldsymbol{\alpha}]$

First, we rewrite the part $I(\mathbf{z} = \mathbf{Bx})$ of the Equation 1.5. As for $I(\mathbf{Y}|\mathbf{X})$, let's define a function g that, for a latent error history v_j, results in the corresponding latent capture history ξ_j: $g(v_j) = \xi_j$. For example $g((1,1,2)) = (1,1,1)$. If we apply g to all histories in $\mathbf{X}$, the resulting set of histories (ξ_j) must be equal to $\mathbf{Z}$. An example is given in Figure 2.2. Thus, $I(\mathbf{X}|\mathbf{Z})$ is 1 if $g(\mathbf{X}) = \cup_j g(v_j) = \mathbf{Z}$ and 0 otherwise. We write this as $I(\mathbf{Z} = g(\mathbf{X}))$.

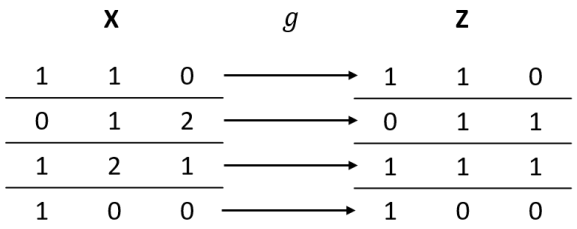

Figure 2.2: Example of $g(\mathbf{X}) = \mathbf{Z}$

The identification likelihood is a product of Bernoulli trials for all captured individuals at all times they were captured. All the different $\alpha_{n,t}$ cannot be summarised and must be multiplied according to the identification result.

$$[\mathbf{X}|\mathbf{Z}, \boldsymbol{\alpha}] = I(\mathbf{X}|\mathbf{Z}) \frac{\prod_k z_k!}{\prod_j x_j!} \prod_{n=1}^{N} \prod_{t=1}^{T} \alpha_{n,t}^{I(v_{n,t}=1)} (1-\alpha_{n,t})^{I(v_{n,t}=2)} . \tag{2.2}$$

- $[\mathbf{Z}|N, \mathbf{p}]$

The capture likelihood is a product of categorical trials, that results in each individual having a well-defined history, and a factorial accounting for the possible reordering:

$$[\mathbf{Z}|N, \mathbf{p}] = \frac{N!}{\prod_k z_k!} \prod_{n=1}^{N} \prod_{t=1}^{T} p_t^{I(\xi_{n,t}=1)} (1 - p_t)^{I(\xi_{n,t}=0)}, \tag{2.3}$$

Since the capture probability is constant across individuals, it can be simplified as Equation 1.4.

- $[\boldsymbol{\alpha} \mid \boldsymbol{\theta}]$

For this part, similar to McClintock et. al (2014) [75], we chose to develop a probit model. Other links could be used, especially since there are no missing covariates. The probit model gives us

[75]: McClintock et al. (2014), 'Probit models for capture-recapture data subject to imperfect detection, individual heterogeneity and misidentification'

$$\alpha_{n,t} = \boldsymbol{\phi}(a \cdot \tau_{n,t} + b)$$

where $\boldsymbol{\phi}$ is the standard normal cumulative distribution function. Thus, $\boldsymbol{\theta} = (a, b)$. We propose a model where $b \neq 0$. To understand why, let's consider what would happen if we kept a sample for which $\tau = 0$ (i.e. having observed no loci at all for that sample). In that case, we could only randomly assign the sample in an already existing history or in a new one. The probability of putting it in the right history would be very low. On the other hand, if τ is large enough (i.e. having observed most loci with good confidence), the probability of misidentifying the sample would be very small. Thus we want $\alpha = \boldsymbol{\phi}(a \cdot 0 + b) \approx 0$, so $b < 0$.

To fully specify the probit model, we define $u_{n,t}$ as a binary indicator of the success of the identification of the capture of individual n at occasion t. That is, $u_{n,t} = 1$ if the sample n, t resulted in a correct individual identification, and 0 otherwise. We also define $\tilde{u}_{n,t}$, a continuous latent process of $u_{n,t}$. We set $\tilde{u}_{n,t} \sim \mathcal{N}(a\tau_{n,t} + b, 1)$ and if $\tilde{u}_{n,t} < 0$ then $u_{n,t} = 0$, or else if $\tilde{u}_{n,t} > 0$ then $u_{n,t} = 1$.

We note $\mathbf{u} = (u_{n,t})_{n\in[1,N],t\in[1,T]|v_{n,t}>0}$ and $\tilde{\mathbf{u}} = (\tilde{u}_{n,t})_{n\in[1,N],t\in[1,T]}$ Since all covariates are known, conditional on $\mathbf{X}$, all the $u_{n,t}$ are known. So the definition of $\tilde{u}_{n,t}$ is not really needed, but it does allow for Gibbs sampling of a and b (see Section 2.2.3).

We only have left to specify priors for a and b:

$$a \sim \mathcal{N}(\mu_a, \sigma_a^2),$$

$$b \sim \mathcal{N}(\mu_b, \sigma_b^2).$$

2.2.3 Estimating the parameters

In order to construct the Markov chain to estimate the parameters, some changes have to be done to the algorithm. The capture probability can be updated as for the $M_{t,\alpha}$ model, but the MCMC has to sample the latent values $\tilde{u}$ and the parameters a and b instead of α. To account for the individual level, the algorithm for sampling the latent histories $\mathbf{X}$ also needs to change.

The MCMC is constructed this way:

1. Let $\beta(a_0^t, b_0^t)$ denote the beta prior on p_t, $\mathcal{N}(\mu_a, \sigma_a^2)$ the normal prior on a and $\mathcal{N}(\mu_b, \sigma_b^2)$ the normal prior on b.

2. Initialize all parameters as well as a set of latent histories satisfying $I(\mathbf{Y} = f(\mathbf{X}))$. Such a set can be obtained by assuming that no mistakes were made (i.e. the exact set of observed histories). In order to run several chains with different initialisations, one can take the previous initialisation of $\mathbf{X}$ and follow the later steps of (5) by only adding misidentifications to the set and always accepting the proposed ones without going though the Metropolis-Hasting acceptance. In the initial latent set, fix a random realistic number of all-zero histories to be part of the population.

3. Sample the capture rate with Gibbs sampling. Nothing changes compared to Section 1.2.4, sample from

$$p_t|\mathbf{Z} \sim \beta(a_0^t + a^t, b_0^t + b^t)$$

where a^t [2] is the number of captured individuals at time t and b^t [3] is the number of unseen individuals at time t (including those never seen).

2: $a^t = \sum_n I(\xi_{n,t} = 1)$
3: $b^t = \sum_n I(\xi_{n,t} = 0)$

4. Sample the identification rate probit parameters with Gibbs sampling. With the probit model, the parameters a and b can be sampled in their full conditional posterior. First, the $\tilde{\mathbf{u}}$ also need to be updated by sampling the $\mathbf{u}_{n,t}$ in their full conditional posterior using the values of $\boldsymbol{\tau}$. For simplification of the notation, we introduce notation L denoting the total number of capture realised and $l = 1, \ldots, L$ the indexes of each capture. Then:

$$\begin{aligned} \tilde{u}_l|\cdot &\sim \begin{cases} \mathcal{TN}_{(0,+\infty)}(a\tau_l + b, 1) & \text{if} \quad \nu_l = 1, \\ \mathcal{TN}_{(-\infty,0)}(a\tau_l + b, 1) & \text{if} \quad \nu_l = 2, \end{cases} \\ a|\cdot &\sim \mathcal{N}(\mu'_a, \sigma'^2_a), \\ b|\cdot &\sim \mathcal{N}(\mu'_b, \sigma'^2_b), \end{aligned} \tag{2.4}$$

where $\mathcal{TN}$ is the truncated normal distribution and

$$\begin{cases} \sigma'^2_a = \left(\frac{1}{\sigma_a^2} + \sum_{l=1}^L \tau_l^2\right)^{-1}, \\ \mu'_a = \sigma'^2_a \left(\frac{\mu_a}{\sigma_a^2} + \sum_{l=1}^L \tau_l(\tilde{u}_l - b)\right), \end{cases} \tag{2.5}$$

and

$$\begin{cases} \sigma'^2_b &= \frac{\sigma_b^2}{L\sigma_b^2+1}, \\ \mu'_b &= \sigma'^2_b \left(\frac{\mu_b}{\sigma_b^2} + \sum_{l=1}^L (\tilde{u}_l - a\tau_l)\right). \end{cases} \tag{2.6}$$

5. Sample $\mathbf{X}$ with Metroplis Hastings. In order to propose an $\mathbf{X}'$, the definitions of sets that where sampled ($\chi_{0,t}(x)$ and $\chi_{2,t}(x)$) need to be changed. Let $\chi_{0,t}(\mathbf{X}) = \{i|\nu_{i,t} = 0\}$ be the set of individual which were unseen at time t in their latent error history if at least one individual is seen only at occasion t. Otherwise $\chi_{0,t}(\mathbf{X}) = \emptyset$. Let $\chi_{2,t}(\mathbf{X}) = \{i|\nu_{i,t} = 2\}$ be the set of individual which were misidentified at time t in their latent error history. Finally, let $\chi_{1,t}(\mathbf{X}) = \{i|\nu_{i,t} = 1, (\nu_{i,s})_{s\neq t} = 0\}$ be the set of individual which

were only seen once, at time t, in their latent error history. Then, to generate $\mathbf{X}^{(k)}$, use the following steps:

a) Set $\mathbf{X}' = \mathbf{X}^{(k-1)}$.
b) With probability 0.5 go to step c, otherwise go to step d.
c) Add a ghost to the proposal set of latent histories $\mathbf{X}'$.
   i. Sample uniformly $t \in \{t|\chi_{0,t}(\mathbf{X}) \neq \emptyset\}$, the set of occasions for which at least one individual is unseen and one individual is only seen at that occasion.
   ii. Sample uniformly $i_0 \in \chi_{0,t}(\mathbf{X})$ the set of unseen individuals at occasion t.
   iii. Sample $i_1 \in \chi_{1,t}(\mathbf{X})$ proportionally to $1 - \alpha_{i_1,t}$.
   iv. Set $v'_{i_0,t} = 2$, with the covariate of identification associated to $v_{i_1,t}$.
   v. Remove v_{i_1} from $\mathbf{X}'$.
   vi. Go to step e.
d) Remove a ghost from the proposal set of latent histories $\mathbf{X}'$.
   i. Sample uniformly $t \in \{t|\chi_{2,t}(\mathbf{X}) \neq \emptyset\}$, the set of occasions where at least one misidentification is present.
   ii. Sample $i_2 \in \chi_{2,t}(\mathbf{X})$ proportionally to $\alpha_{i_2,t}$.
   iii. Add to $\mathbf{X}'$ an individual with a single capture at time t with the covariate of identification that is associated to $v_{i_2,t}$.
   iv. Set $v'_{i_2,t} = 0$.
e) Compute $\mathbf{Z}' = g(\mathbf{X}')$ Set N' as the number of individuals in $\mathbf{X}'$.
f) With probability r_1, set $\mathbf{X}^{(k)} = \mathbf{X}'$, $\mathbf{Z}^{(k)} = \mathbf{Z}'$ and $N^{(k)} = N'$. Otherwise set $\mathbf{X}^{(k)} = \mathbf{X}^{(k-1)}$, $\mathbf{Z}^{(k)} = \mathbf{Z}^{(k-1)}$ and $N^{(k)} = N^{(k-1)}$.

$$r_1 = min\left(1, \frac{N'! \prod_{i=1}^{N'} \pi_i}{N! \prod_{i=1}^{N} \pi_i} \frac{q(\mathbf{X}^{(k-1)}|\mathbf{X}', \boldsymbol{\alpha})}{q(\mathbf{X}'|\mathbf{X}^{(k-1)}, \boldsymbol{\alpha})}\right), \tag{2.7}$$

where $\pi_i = \prod_{t=1}^{T} p_t^{I(v_{i,t}>0)}(1-p_t)^{I(v_{i,t}=0)}\alpha_{i,t}^{I(v_{i,t}=1)}(1-\alpha_{i,t})^{I(v_{i,t}=2)}$ and $[\mathbf{X}'|\mathbf{X}]$ is the proposal density for $\mathbf{X}'$. When adding a ghost:

$$[\mathbf{X}'|\mathbf{X}, \boldsymbol{\alpha}] = \frac{0.5(1-\alpha_{v_1,t})}{\sum_{i\in\chi_{1,t}(\mathbf{X})}(1-\alpha_{i,t})\#\{t|\chi_{0,t}(\mathbf{X}) \neq \emptyset\}\#\chi_{0,t}} \tag{2.8}$$

and when removing a ghost:

$$[\mathbf{X}'|\mathbf{X}, \boldsymbol{\alpha}] = \frac{0.5\alpha_{v_2,t}}{\sum_{i\in\chi_{2,t}(\mathbf{X})}(\alpha_{i,t})\#\{t|\chi_{2,t}(\mathbf{X}) \neq \emptyset\}} \tag{2.9}$$

where $\#S$ denotes the cardinal of ensemble S.

6. Sample the number of unseen individuals:
   a) set $\mathbf{X}' = \mathbf{X}$, $\mathbf{Z}' = \mathbf{Z}$ and x_0 the number of unseen individual in $\mathbf{Z}$ (and $\mathbf{X}$),
   b) sample a move $c \in [-D, D]$ where D is fixed integer,
   c) define $x'_0 = x_0 + c$,
   d) set the number of unseen individuals in $\mathbf{X}'$ and $\mathbf{Z}'$ to x'_0,
   e) accept $\mathbf{X}'$ and $\mathbf{Z}'$ with probability r_2:

$$r_2 = min\left(1, \frac{[\mathbf{Z}'|N', \mathbf{p}]}{[\mathbf{Z}|N, \mathbf{p}]}\right). \tag{2.10}$$

7. repeat steps 3 to 6 as much as needed.

The proof of convergence of this algorithm is the same as in Section 1.2.4. Some notations change, due to the moves not being written as vectors. In the proof, we replace the definition of moves as vectors by functions of a set of latent histories: $b(\mathbf{X}) \in \mathcal{M}(\mathbf{X})$. The reverse move of $b(\mathbf{X})$ is $b^{-1}(b(\mathbf{X}))$. The full proof can be re-written with these new notation. It will remain the same so the proof is still valid.

2.3 Simulation study

2.3.1 Scenarios

In this section, we conducted a simulation study in which we evaluated the models $M_{t,\alpha}$ and M_{t,α_n}, along with the model proposed by [67] , in which all histories with a single capture are excluded. We also compared the models with the standard capture-recapture M_t model to demonstrate the danger of disregarding the misidentifications.

[67]: Yoshizaki et al. (2011), 'Modeling misidentification errors that result from use of genetic tags in capture–recapture studies'

Yoshizaki's conditional model

If ω is any observable capture history, let ω'' denote capture histories with 2 or more captures. Let the random variable $y_{\omega''}$ represent the number of observed histories ω''. Then $\mathbf{y}'' = (y_{\omega''_1}, \ldots)$ follows a multinomial of index $N'' = \sum \mathbf{y}''$ and probabilities

$$\pi_{\omega''} = \pi_\omega / \pi*$$

where π_ω is the same as Equation 1.2 and

$$\pi* = \sum_{\omega''} \pi_\omega.$$

Finally, N is estimated by

$$N''/\hat{\pi}*$$

We simulated capture-recapture according to the model $Mt, \alpha n$. We simulated observation data for $T = 5, 7, 9$, $N = 500$, $p = 0.1, 0.2, 0.3, 0.4$ and $\bar{\alpha} = 0.8, 0.9, 0.95$. Values of parameters a and b were computed automatically, depending on the number of captured individuals, to achieve the wanted $\bar{\alpha}$. There were 12 parameters combinations for each of the three different number of occasions. For the sake of simplicity, we always used the time-dependent models (M_t, $M_{t,\alpha}$ and $Mt, \alpha n$), although the capture rate was kept constant over time in the simulations. For each scenario, we simulated 100 data-sets.

We utilised uninformative priors for all parameters. Setting uninformative priors is straightforward for all the parameters of the models M_t, $M_{t,\alpha}$ and Yoshizaki's model, as well as for the parameter $\mathbf{p}$ of the M_{t,α_n} model. For the parameters a and b of the M_{t,α_n} model, we estimated both parameters in the case where no misidentification occurred and in the case where all single-capture histories resulted from misidentifications (arbitrarily limited to a maximum of 30% misidentifications). The mean of the prior was then set as the mean of the two estimates, and the standard error as the difference between both estimates divided by four. We compared these priors with those obtained by estimating a and b in all cases of numbers of misidentifications, between zero misidentifications and the maximum of misidentifications (or resulting 30% of the captures) and fitting a normal to all estimates. Since the inverse-probit function is not linear, the obtained prior is not really normal, but the one computed using only the minimum and maximum number of misidentifications turns out to be very close to being absolutely uninformative.

2.3.2 Implementation

As previously, we used NIMBLE ([80]) to implement the models. The advantage of NIMBLE is that it allows writing all the samplers of the MCMC (mandatory here) and new distributions. Thus, we wrote the likelihood of the model and the sampler of $\mathbf{X}$. We were also able to

[80]: Valpine et al. (2017), 'Programming With Models'

rewrite all the Gibbs samplers of the MCMC for maximum computational efficiency.

For the $M_{t,\alpha}$, the MCMC was run over 1E6 iterations after a burn-in period of 200,000 iterations and the chains were thinned by a factor of 1/100 in order to limit memory usage. For the M_{t,α_n}, the MCMC was run over 1E6 iterations after a burn-in period of 40,000 iterations and the chains were thinned by a factor of 1/100 in order to limit memory usage. These numbers of iterations were to ensure that, when it can be achieved, all chains converge, but less iteration are needed in most cases. We ran two chains for each simulation with two different starting points. For the first one, **X** was initialised as the set of observed histories, as if there was no error. In the second one, we arbitrarily added 40 errors randomly.

The M_t model was run using the same scripts as the $M_{t,\alpha}$ but without the MCMC sampler relatives to **x** and α. Yoshizaki's conditional model was also implemented with NIMBLE. No sampler were rewritten for that model as it doesn't need too many iteration to be run.

2.3.3 Compared results of models

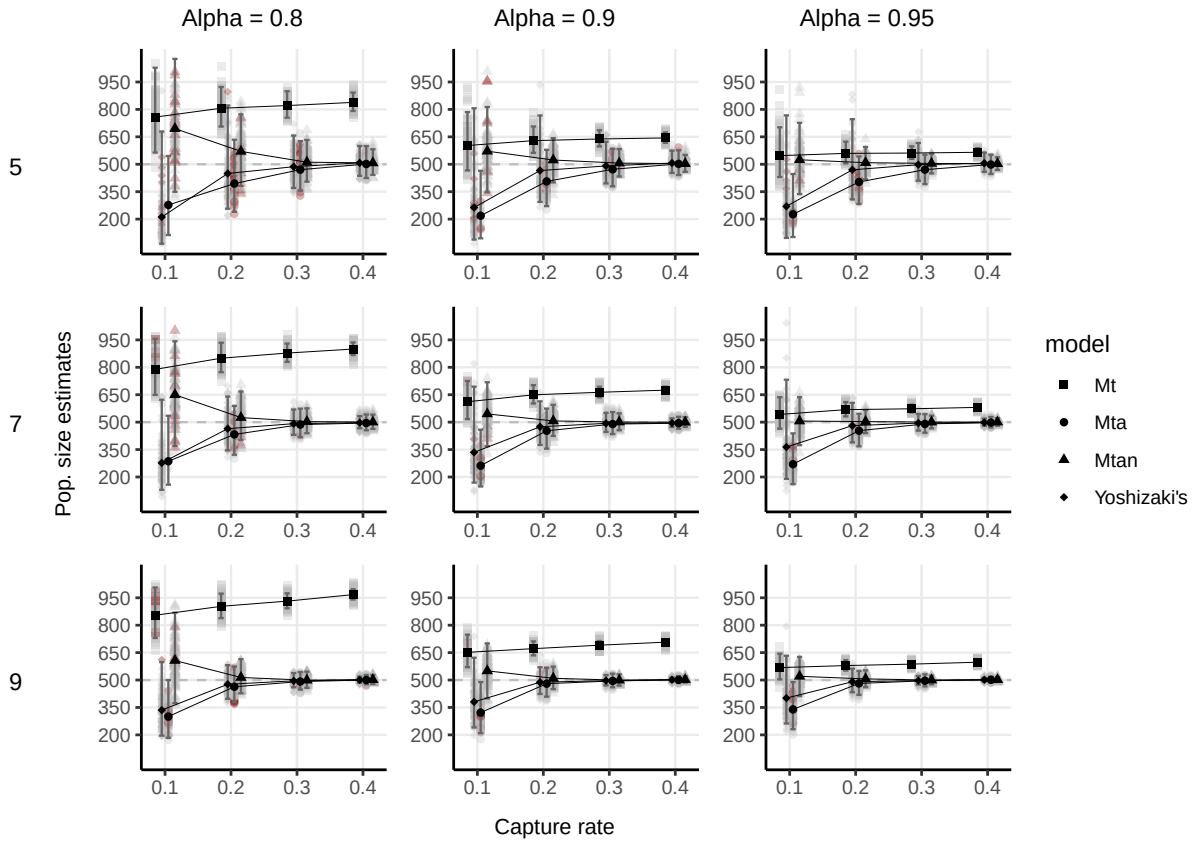

Figure 2.3: Single state population size estimations (y axis) depending on capture probability (x axis), identification probability (columns), number of capture occasions (lines) and model used (dot shape). Horizontal dashed lines indicate true population size. Grey and red symbols show simulation-specific estimates of the population-size posterior mean, red ones indicating that the N chains had $\hat{R} > 1.1$ or an effective sample size under 50. Black symbols connected by lines and error bars show, respectively, the estimates of the mean and the 95% credible intervals of the posterior distribution of population-size averaged across simulations.

Running two chains of 1,200,000 iteration on a 3.0GHz Intel processor took about 1 hour 40 minutes. We checked the convergence graphically

by looking at the N chains because N is the slowest moving parameter. We also looked at the $\hat{R}$ and the resulting effective sample size. The M_t model always converged perfectly and there was no indication of any problems.Yoshizaki's model converged in all scenarios. However, for $p = 0.1$, the chains appeared to be constrained by a lower bound. For the $M_{t,\alpha}$, the MCMC converged on all simulations where $p \geq 0.2$ but when $p = 0.1$, it only converged for simulations with nine capture occasions. In comparison, with the M_{t,α_n}, the MCMC almost always converged: only a few simulations with five occasions and $p = 0.1$ failed to converge. Convergence was always faster for higher identification rates.

The population size estimates of all models are shown on Figure 2.3. The M_t model always overestimated the population size, up to a factor of two. Lower identification rates and higher capture rate both led to greater overestimation.
Yoshizaki's model had the same trend of estimates as the $M_{t,\alpha}$ model. For very low capture rates, when convergence was achieved, the chains seemed to minimize the population size, resulting in an unlikely very low identification probability estimate. For capture rates greater than 0.2 there was no bias and with nine occasions or more, $p = 0.2$ is sufficient to obtain unbiased estimates. For capture rates that were too low, when the convergence was acceptable, the confidence interval estimated by Yoshizaki's model averaged over the simulations included the true value. This was not the case with $M_{t,\alpha}$. In addition, in cases with no bias, the confidence interval of Yoshizaki's model was smaller than that of $M_{t,\alpha}$, making the former more accurate. However, despite the zero average bias, with a high number of occasions and high capture rates, Yoshizaki's model resulted in a lower percentage of simulations with the true value of N in the confidence interval than the $M_{t,\alpha}$ [4]: with nine occasions and $p = 0.4$, the percentage of simulations with the true population size in the confidence interval was 77% using Yoshizaki's model, and 94% using the $M_{t,\alpha}$ model.
The M_{t,α_n} model performed slightly better than the others: with $p = 0.1$, the average bias was lower than with Yoshizaki's model, and more than 90% had the true population size in the confidence interval (against 60% to 70% for the conditional). With $p \geq 0.2$, there was no bias on average, and the confidence interval was smaller than that produced by $M_{t,\alpha}$. The confidence interval estimated with M_{t,α_n} was smaller than that obtained with $M_{t,\alpha}$: from 5% up to 35% on average when the $M_{t,\alpha}$ model converged. It was also smaller than that produced by Yoshizaki's model with $p \leq 0.3$. Also, the confidence interval contained the true value of the population size more than 90% of the time in all scenarios. The identification rate played a significant role in the estimates. The lower it was, the higher the uncertainty in the estimates.

4: With 9 occasions and $p = 0.4$, percent of simulations with true population size in the confidence interval:

- yoshizaki's model: 77%
- $M_{t,\alpha}$: 94%

A table in Appendix C gives these percentages for all scenarios.

2.4 Discussion

In this chapter, I developed and implemented an extension of the LMM that uses quality covariates to model the probability of identification. This new M_{t,α_n} model proved to be better than the $M_{t,\alpha}$. Specifically, provided that misidentification rate is kept below a maximum of 5%, even with only five occasions and a capture rate as low as 0.1, accurate

estimates of the parameters can be obtained. In addition, although the computation time for the probit model is drastically increased compared to the original model (from 5 minutes to 2 hours for 9 occasions with a population size of 500 individuals), it is still very reasonable. We also showed that Yoshizaki's model gives equivalent estimates to the $M_{t,\alpha}$ model. The estimates are biased in the same scenarios and the bias is of similar magnitude. Since Yoshizaki's model is very easy to implement and fast to run, we suggest to run it for any closed CMR experiment. The comparison with a classical model will confirm whether the data contains misidentifications or not. Then, if there are misidentifications, run the M_{t,α_n} model. We give the NIMBLE code for running Yoshizaki's model in Appendix D.

With very low capture rates (≤ 0.1), misidentification rates higher than 0.05 will require more occasions to produce good estimates. Even with 9 occasions, the estimates are slightly biased for 10% misidentifications, but the bias is reduced compared to scenarios with fewer occasions.
To make the M_{t,α_n} model worth using, the total number of identifications realised must be enough to allow for a good estimate of the probit parameters. If the population is small (e.g. less than a hundred individuals for example) and the capture rate does not compensate it, the total number of capture will be low. The probit may be difficult to estimate, and higher confidence interval may result from the M_{t,α_n} model compared to the $M_{t,\alpha}$. Thus, in such a case, the $M_{t,\alpha}$ model could lead to a better fit.

In some cases, even if the $\hat{R}$ statistic is below 1.1, poor convergence of the chain can be observed. Therefore, the convergence must be visually checked in the chains of parameter N, the slowest parameter to converge.

The extension we propose here may also have applications other than simply improving the estimates. For example, in a multistate CMR experiment where the identification probability is expected to vary between states, a model M_{t,α_s} (without covariate) could be built where $\boldsymbol{\alpha} = (\alpha_1, ..., \alpha_S)$, that is estimating as many different α_s as there are states. In this case, using the M_{t,α_n} model with the quality as a covariate should be enough to model the differences between states, without having separate parameters. The M_{t,α_n} model would also allow information to be shared between all states for a better estimate than if all α_s had to be estimated independently.

The M_{t,α_n} model could easily be extended to account for more than one covariate. An example still lies in the use of quality measure: since it is most likely computed per locus, each value could be used as a different covariate with its own parameter. If there is reason to believe that some loci inform more or less than others on an individual's identity, differentiating them in the probit may lead to better results.

We have used the data quality information to improve the estimate of the number of error-prone histories and their selection, bringing the genotyping data a step closer to the model. A major step now remains to bring it even closer by potentially using information about proximity, i.e. how close genotypes are to each other. By modelling the identification probability as a function of relatedness, the ghosts would be matched to those likely to have produced them, rather than at random. Thus, a single capture history with no close proximity to any history would have

a low likelihood as a ghost, and conversely one with close proximity to another history would have a high likelihood as a ghost.
However, one of the challenges is to define the relation between relatedness and the probability of a good identification. To model the identification probability as a function of relatedness, we need to calculate how likely a misidentification is depending on all other samples. The relatedness to the other samples from the history should increase the probability of identifying it from that individual. And the relatedness to samples from other histories should support the fact that they are not from the same individual. If the relatedness to other samples is left out, the identification probability for individuals with unique capture could not be calculated, or would have to be 1. In addition, a lot of information would be needed and stored (the relatedness of all samplers to each other). Although the moves could be sampled in a more efficient way and convergence could be reached with less iteration than what we have reported, building an MCMC for a model that include relatedness seems to remain a challenge.

Chapter three

Repeated observations on an occasion

3

3.1 Aim of the chapter

One of the main hypotheses of the LMM published by Link et al. (2010) [74] is that individuals can only be seen once per occasion. As many studies use eDNA extracted from faeces, it is common for multiple samples to be collected from the same individual on the same occasion. These individuals are 'captured' multiple times. Although these studies could benefit from modelling misidentification, the LMM is not suitable for these experiments because the single-capture assumption is violated. In this chapter I extend the LMM to account for repeated observations of individuals on the same occasion. To do this, I use a Poisson process to model the observation process. In addition, I propose an MCMC algorithm that generates a latent set of histories accounting for repeated observations. Finally, I apply the new model to a real data set from a study of Eurasian otters.

3.2 A Poisson extension of the model

3.2.1 The Poisson model

In the case where several captures of the same individual can occur at the same occasion, the observable history ω_i of an individual is thus composed of counts representing the number of time the individual was observed and identified. For example we might observe the following individual history: $(0, 2, 0, 3, 1)$ where 2 is the number of times the individual is detected at occasion 2. In an experiment in which individuals can be "captured" several times on one occasion, one of the captures might result in a misidentification and the creation of a ghost history, while the other captures are correct identifications. In contrast to the model $M_{t,\alpha}$, an individual producing a ghost at a given occasion can be seen at the same occasion if another capture of it results in a good identification.

Due to the new structure of the data, we must modify the notations. A history may contain several misidentifications at the same occasion. To represent latent error histories ν_j, we note the total number of observations of an individual on each occasion with the number of these observations that resulted in misidentification as a superscript. For example, the observed history $(0, 2, 0, 3, 1)$ might have been generated by the latent error history $(1^{(1)}, 2, 0, 3, 3^{(2)})$. In this example, the observation at the first occasion and two observations at the fifth occasion were misidentified, resulting in zero observations at occasion 1 and one observation at occasion 5. The latent capture history ξ_k is the same as the latent error history without the superscripts. In our example, the latent capture history is $(1, 2, 0, 3, 3)$.

In this section, we make the same assumptions about the misidentifications as in the model $M_{t,\alpha}$. Identifications are independent, misidentifications always result in ghosts (i.e. false individuals), and ghosts cannot be seen again (i.e. ghosts have exactly one observation in their history).

The observed histories of an experiment can be summarised in the frequency vector $\mathbf{y}$. The different latent error histories possibly responsible for the observed histories can be summarised in the frequency vector $\mathbf{x}$. As in the model $M_{t,\alpha}$, the vector $\mathbf{y}$ is the linear transformation of the vector $\mathbf{x}$: $\mathbf{y} = \mathbf{A}\mathbf{x}$, where $\mathbf{A}$ is a known matrix.

The Table 3.1 shows an example of $\mathbf{y}$, $\mathbf{x}$ and $\mathbf{A}$ in the case where the three histories $(1,2)$, $(0,3)$ and $(1,0)$ were observed, $(1,2)$ three times, $(0,3)$ twice, and $(1,0)$ only once. Any observed history with at least 2 observations corresponds necessarily to a real individual because a central assumption of our model is that no two misidentifications may produce the same error (note that this means that the genotyping process must be sufficiently discriminant). Under this assumption, the set of observed histories in Table 3.1 implies that there are at least 5 individuals in our population: 3 with observed history $(1,2)$, and 2 with observed history $(0,3)$. As for the observed history $(1,0)$, it may correspond to a sixth individual that was correctly identified or not–this last point we cannot know–or be a ghost generated by the misidentification at the first occasion of one of the 5 individuals already mentioned. We have actually 4 different possible sets of latent error histories:

1. $\{(1,2),(1,2),(1,2),(0,3),(0,3),(1,0)\}$
2. $\{(1,2),(1,2),(2^{(1)},2),(0,3),(0,3)\}$
3. $\{(1,2),(1,2),(1,2),(0,3),(1^{(1)},3)\}$
4. $\{(1,2),(1,2),(1,2),(0,3),(0,3),(1^{(1)},0)\}$

The Table 3.1 presents the second possibility.

	v_j	$(1,2)$	$(0,3)$	$(1,0)$	$(2^{(1)},2)$	$(1^{(1)},3)$	$(1^{(1)},0)$
ω_i	$\mathbf{y}/\mathbf{x}$	2	2	0	1	0	0
$(1,2)$	3	1	0	0	1	0	0
$(0,3)$	2	0	1	0	0	1	0
$(1,0)$	1	0	0	1	1	1	1

Table 3.1: The first and second columns are, respectively, the observed histories and the observed history frequencies. The first and second rows are, respectively, the possible latent error histories and one possible set of frequencies of the latent error histories. The rest of the table is the matrix $\mathbf{A}$.

Similarly, the latent capture histories can be summarised in the frequency vector $\mathbf{z}$. Also, $\mathbf{z}$ is a linear transformation of $\mathbf{x}$: $\mathbf{z} = \mathbf{B}\mathbf{x}$. The Table 3.2 continues the example from above, showing possible $\mathbf{z}$ and $\mathbf{B}$.

	v_j	$(1,2)$	$(0,3)$	$(1,0)$	$(2^{(1)},2)$	$(1^{(1)},3)$	$(1^{(1)},0)$
ξ_k	$\mathbf{z}/\mathbf{x}$	2	2	0	1	0	0
$(1,2)$	2	1	0	0	0	0	0
$(0,3)$	2	0	1	0	0	0	0
$(1,0)$	0	0	0	1	0	0	1
$(2,2)$	1	0	0	0	1	0	0
$(1,3)$	0	0	0	0	0	1	0

Table 3.2: The first and second columns are, respectively, the latent capture histories and the latent capture history frequencies. The first and second rows are, respectively, the latent error histories and the same set of latent error history frequencies as in Table 3.1. The rest of the table is the matrix $\mathbf{B}$.

Let $\boldsymbol{\lambda} = (\lambda_1, ..., \lambda_T)$ be the set of parameters involved in the capture process, modelled by a Poisson process (each λ_t being a parameter of the Poisson process at occasion t). The likelihood of the model, conditional on $\mathbf{x}$ and $\mathbf{z}$ is given by the Equation 3.1.

$$[\mathbf{y}, \mathbf{x}, \mathbf{z}|N, \boldsymbol{\lambda}, \alpha] = I(\mathbf{y} = \mathbf{A}\mathbf{x})\,[\mathbf{x}|\mathbf{z}, \alpha]\,[\mathbf{z}|N, \boldsymbol{\lambda}] \tag{3.1}$$

We now describe the two elements that change compared to model $M_{t,\alpha}$, i.e. $[\mathbf{z}|N, \boldsymbol{\lambda}]$ and $[\mathbf{x}|\mathbf{z}, \alpha]$.

- $[\mathbf{z}|N, \boldsymbol{\lambda}]$

We model the true number of observations of an individual at an occasion with a Poisson distribution. The probability π_k that an individual has a given latent capture history ξ_k is given by:

$$\pi_k = \prod_{t=1}^{T} \frac{\boldsymbol{\lambda}_t^{\xi_{k,t}}}{\xi_{k,t}!} e^{-\lambda_t} \tag{3.2}$$

The capture likelihood has a multinomial form:

$$[\mathbf{z}|N, \boldsymbol{\lambda}] = \frac{N!}{\prod_k z_k!} \prod_k \pi_k^{z_k} \tag{3.3}$$

- $[\mathbf{x}|\mathbf{z}, \alpha]$

All realised captures are potentially subject to misidentifications. Let $o_{j,t}$ be the number of good identifications for individuals with history v_j at occasion t. Then, knowing the true number of captures, the probability that it was correctly identified $o_{j,t}$ times is Binomial. Thus:

$$[\mathbf{x}|\mathbf{z}, \alpha] = \prod_j \prod_{t=1}^{T} \binom{v_{j,t}}{o_{j,t}} o_{j,t}^{\alpha} (v_{j,t} - o_{j,t})^{1-\alpha} \tag{3.4}$$

3.2.2 Estimating the parameters

To the algorithm used for model $M_{t,\alpha}$, we need to add how we sample $\boldsymbol{\lambda}$, which is a new parameter, but also the way we propose a new $\mathbf{x}$. The way we sample α does not really change.

The MCMC is constructed this way:

1. Let $\lambda_t \sim \Gamma(\alpha_0^{(t)}, \beta_0^{(t)})$ be the Gamma prior over $\boldsymbol{\lambda}$, and $\beta(a_0^{\alpha}, b_0^{\alpha})$ denote the beta prior on α.

2. Initialize all parameters as well as a set of latent histories satisfying $\mathbf{y}$=$\mathbf{A}\mathbf{x}$. Such a set can be obtained by assuming that no mistakes were made. The latent frequencies of the histories containing 2's are 0 and all the other match the observed frequencies one-to-one. In order to run several chains with different initialisations, one can take the previous initialisation of $\mathbf{x}$ and follow the later steps of (5) by only adding misidentification to the set and always accepting the proposed ones without going though the Metropolis-Hasting

acceptance. In the initial latent set, fix the number of unseen individual to a random realistic number.

3. Sample the capture parameter $\boldsymbol{\lambda}$ with Gibbs sampling. The likelihood being a product of Poisson, it follows that the Gamma priors lead to full conditional Gamma posterior distribution:

$$\lambda_t|\mathbf{z} \sim \Gamma(\alpha_0^{(t)} + a^{(t)}, \beta_0^{(t)} + b^{(t)})$$

, where $a^{(t)}$ [1] is the total number of observations on occasion t and $b^{(t)}$ is the number of individuals that were available at occasion t.

1: $a^{(t)} = \sum_k \xi_{k,t}$

4. Sample the identification rate with Gibbs sampling. Similarly to the capture rate, it has a full conditional beta posterior distribution.

$$\alpha|\mathbf{x} \sim \beta(a_0^{\alpha} + a^{\alpha}, b_0^{\alpha} + b^{\alpha})$$

where a^{α} [2] is the total number of correct identifications and b^{α} [3] is the total number of misidentifications.

2: $a^{\alpha} = \sum_j \sum_t x_j I(v_{j,t} = 1)$

3: $b^{\alpha} = \sum_j \sum_t x_j I(v_{j,t} = 2)$

5. Sample jointly N and $\mathbf{x}$. To update the frequencies of the latent histories $\mathbf{x}$ and $\mathbf{z}$ and the number of individuals we use Metropolis-Hastings. Since an individual can be correctly identified on one of its observations and misidentified on another on the same occasion, to propose an $\mathbf{x}'$ we need to be able to add an error to any individual. Except for this point, the concept of the algorithm remains the same. Randomly add or remove a misidentification. To add a misidentification, sample an occasion t from those that could have generated one (i.e., the occasions for which there is at least one individual with a unique capture, at that occasion). Then sample a history v_0 from all the available ones in the current set of histories. And "merge" a history with a unique capture at occasion t into the sampled history (i.e., add a capture and a misidentification at occasion t to an individual with history v_0 and remove one history v_{1t}). To remove a misidentification, sample an occasion t from those where at least one misidentification has occurred. Then sample a history v_2 that contains a misidentification at that occasion t, remove a capture and an error at occasion t to one individual with history v_2, and add an individual with a unique capture at occasion t.
   More formally, follow the steps:
   a) Define:
      - $\boldsymbol{\nu}^{(1t)}$ the history with a unique capture at time t (potential ghost),
      - $\chi_{2,t}(\mathbf{x}) = \{v_j|v_{j,t} - o_{j,t} > 0, x_{v_j} > 0\}$ the set of histories *containing* at least one ghost at occasion t, for the given $\mathbf{x}$.
   b) With probability 0.5, go to (i), otherwise go to (ii).
      i. Add a misidentification (i.e. a ghost) to the latent set.
         - Sample $t \in \{t|x_{\nu^{(1t)}} > 0\}$.
         - Sample $\boldsymbol{\nu}^{(0)}$ uniformly from the set of histories for which $\mathbf{x}_j > 0$.
         - Set $\boldsymbol{\nu}^{(2)} = \boldsymbol{\nu}^{(0)} + \boldsymbol{\nu}^{(1t)}$ and add one error at occasion t. For example, if $\boldsymbol{\nu}^{(1t)} = (0, 1^{(0)})$, and $\boldsymbol{\nu}^{(0)} = (1^{(0)}, 3^{(1)})$, then $\boldsymbol{\nu}^{(2)} = (1^{(0)}, 4^{(2)})$.

- Define the move $b_{\boldsymbol{\nu}^{(0)},\boldsymbol{\nu}^{(1t)},\boldsymbol{\nu}^{(2)}} = (-1,-1,+1)$, and $b_\nu = 0$ for all other latent histories.

ii. Remove a misidentification from the latent set.

- Sample t in the set of occasions where at least one misidentification occurred.
- Sample $\boldsymbol{\nu}^{(2)} \in \chi_{2,t}(x)$, the history containing a misidentification at occasion t.
- Define $\boldsymbol{\nu}^{(0)} = \boldsymbol{\nu}^{(2)} - \boldsymbol{\nu}^{(1t)}$ and remove one error at occasion t. For example, if $\boldsymbol{\nu}^{(1t)} = (0, 1^{(0)})$, and $\boldsymbol{\nu}^{(2)} = (1^{(0)}, 4^{(2)})$ then $\boldsymbol{\nu}^{(0)} = (1^{(0)}, 3^{(1)})$.
- Define the move $b_{\boldsymbol{\nu}^{(0)},\boldsymbol{\nu}^{(1t)},\boldsymbol{\nu}^{(2)}} = (+1,+1,-1)$, and $b_\nu = 0$ for all other latent histories.

c) Define $\mathbf{x}' = \mathbf{x}^{(k-1)} + b$.

d) Calculate $\mathbf{z}' = \mathbf{B}\mathbf{x}'$ and $N' = \sum \mathbf{x}'$.

e) With probability r_1, set $\mathbf{x}^k = \mathbf{x}'$, $\mathbf{z}^k = \mathbf{z}'$ and $N^(k) = N'$. Otherwise set $\mathbf{x}^k = \mathbf{x}^{k-1}$, $\mathbf{z}^k = \mathbf{z}^{k-1}$ and $N^(k) = N^{(k-1)}$.

$$r_1 = min\left(1, \frac{[\mathbf{y}, \mathbf{x}', \mathbf{z}'|N', \boldsymbol{\lambda}, \alpha]}{[\mathbf{y}, \mathbf{x}^{(k-1)}, \mathbf{z}^{(k-1)}|N, \boldsymbol{\lambda}, \alpha]} \frac{q(\mathbf{x}^{(k-1)}|\mathbf{x}')}{q(\mathbf{x}'|\mathbf{x}^{(k-1)})}\right). \tag{3.7}$$

6. Sample the number of unseen individuals x_1:

a) set $\mathbf{x}' = \mathbf{x}$, $\mathbf{z}' = \mathbf{z}$ and x_0 the number of unseen individual in $\mathbf{z}$ (and $\mathbf{x}$),

b) sample a move $c \in [-D, D]$ where D is fixed integer,

c) define $x'_0 = x_0 + c$,

d) set the number of unseen individuals in $\mathbf{x}'$ and $\mathbf{z}'$ to x'_0,

e) accept it with probability r_2:

$$r_2 = min\left(1, \frac{[\mathbf{z}'|N', \boldsymbol{\lambda}]}{[\mathbf{z}|N, \boldsymbol{\lambda}]}\right). \tag{3.8}$$

7. repeat steps 3 to 6 as much as needed.

The proof of convergence of the algorithm stays mainly the same as in Section 1.2.5. The first condition becomes that step 3 and 6 produce chains which converge to $\pi(N, \boldsymbol{\lambda}, |\mathbf{z})$ for any $\mathbf{z}$ such that $\mathbf{z} = \mathbf{B}\mathbf{x}$ for some $\mathbf{x} \in \mathcal{F}_y$. The validity of the condition is trivial again. The second and third conditions do not change, and their proof remain the same.

Proposal density

The proposal densities are calculated by multiplying the probabilities of each sampling step used for defining the move. They are successively: the probability of adding (or removing) an error, the probability of choosing the ν_0 (or ν_2) and the probability of choosing the t knowing the sampled ν. When adding an error, the proposal density q is:

$$q(x'|x^{k-1}) = \frac{0.5}{\#\chi_0.\#\{t|\nu_{0,t} = 0, x_{\nu_{1t}} > 0\}} \tag{3.5}$$

and when removing an error, is:

$$q(x'|x^{k-1}) = \frac{0.5}{\#\chi_2.\#\{t|\nu_{2,t} = 2\}} \tag{3.6}$$

where $\#S$ denotes the cardinality of S.

3.3 Simulation study design

3.3.1 Comparison of models for repeated observations

We conducted a simulation study in which we evaluated the $M_{\lambda,\alpha}$ model against the model proposed by Yoshizaki et al. (2011) [67], in which all histories with a single capture are excluded. We also tested the standard capture-recapture M_t model to demonstrate the importance of disregarding misidentifications.

[67]: Yoshizaki et al. (2011), 'Modeling misidentification errors that result from use of genetic tags in capture–recapture studies'

We simulated datasets according to the $M_{\lambda,\alpha}$ model for a population of $N = 500$ individuals. We aimed for capture rates (with the meaning "seen at least once" or "not seen at all" on an occasion) of 0.1, 0.2, 0.3, and 0.4, so we set $\lambda_t = 0.11, 0.23, 0.36, 0.51$. We simulated identification rates

as in the first study, $\alpha = 0.8, 0.9, 0.95$. We simulated 10 datasets for these 12 scenarios.

3.3.2 Practical implementation

As we did previously, we used NIMBLE ([80]) to implement the model. The advantage of NIMBLE is that it allows writing all the samplers of the MCMC (mandatory here) and new distributions. Thus, we wrote the likelihood of the model and the sampler of $\mathbf{x}$. We were also able to rewrite all the Gibbs samplers of the MCMC for maximum computational efficiency.

[80]: Valpine et al. (2017), 'Programming With Models'

the $M_{\lambda,\alpha}$ model, was run over 1E6 iterations after a burn-in period of 100,000 iterations and the chains were thinned by a factor of 1/100 in order to limit memory usage. The M_t model was run over 100,000 iteration after a burn-in period of 10,000 iterations, and the chains were thinned by a factor of 1/10. Yoshizaki's model was run over 10,000 iterations after a burn-in period of 1000 iterations. These numbers of iterations were to ensure that, when it can be achieved, all chains converge, but less iteration are needed in most cases.

We ran two chains for each simulation with two different starting points. For the first one, $\mathbf{X}$ was initialised as the set of observed histories, as if there was no error. In the second one, we arbitrarily added 40 errors randomly. Previous tests indicated that two chains were sufficient to see convergence.

3.4 Results

3.4.1 Simulation study results

Running two chains of 1,100,000 iterations for the $M_{\lambda,\alpha}$ model on a 3.0GHz Intel processor took less than 10 minutes. We checked the convergence graphically by looking at the N chains, as N is the slowest moving parameter. We also looked at the $\hat{R}$ and the resulting effective sample size. The M_t model always converged perfectly and there was no evidence of any problems. Yoshizaki's model converged for all simulations, but some chains for $p = 0.1$ seemed to be constrained by a lower bound. For the $M_{\lambda,\alpha}$, the MCMC converged in all scenarios. The convergence was always faster for higher identification rates.

The population size estimates for both models are shown in Figure 3.1. The M_t model always overestimated the population size, up to a factor of two. Lower identification rates and higher capture rates both led to greater overestimation. In these scenarios, Yoshizaki's model and model $M_{t,\alpha}$ had the same trend of estimates. For very low capture rates, the chains seemed to minimize the population size, resulting in an unlikely very low identification probability estimate. For a capture rate of 0.2, there was only a small bias with seven or fewer occasions, and no bias with nine occasions. There was no bias for higher capture rates. The confidence interval estimated by the $M_{\lambda,\alpha}$ model was smaller than that estimated by Yoshizaki's model for most scenarios with seven or fewer occasions. They were slightly larger with nine occasions, but the true

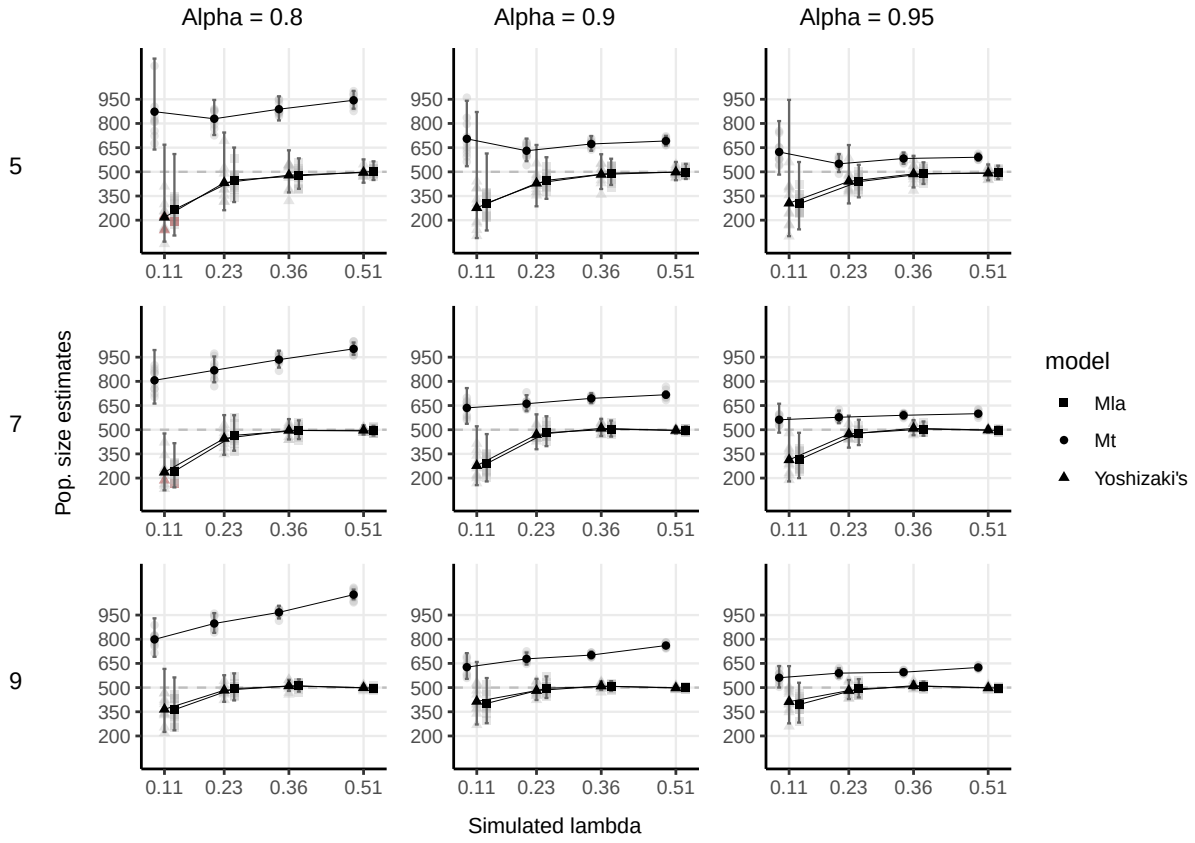

Figure 3.1: Multi-capture CMR population size estimations (y axis) depending on capture probability (x axis), identification probability (columns), number of capture occasions (lines) and model used (dot shape). The $M_{\lambda,\alpha}$ model is noted Mla. Horizontal dashed lines indicate true population size. Grey and red symbols show simulation-specific estimates of the population-size posterior mean. Red ones indicates that the N chains had $\hat{R} > 1.1$ or an effective sample size under 50. Black symbols connected by lines show the mean estimates averaged across simulations. Error bars show the estimate of the 97.5% and 2.5% quantiles averaged across simulations.

population size was more often in the confidence interval estimated by the $M_{\lambda,\alpha}$ model.

3.4.2 Application to Otter dataset

We applied our model to data from a study of the Eurasian otter (*Lutra lutra*), in Upper Lusatia, Saxony, Germany [93]. Otters are nocturnal and elusive and pose challenges for (live-)trapping [94]. Otter faeces (spraints) are particularly suitable for studying the species, as they are used for intraspecific communication. Otters produce up to 30 spraints a day, tending to defecate on frequently visited visible terrestrial sites at specific locations throughout their home range. Data collection involved collecting spraints over five consecutive days ($T = 5$) from 2006 to 2012 (excluding 2009). Sampling was conducted in March (2006, 2010, 2011, 2012), April (2007), and May (2008). The authors considered it unlikely that repeated PCR could completely eliminate all genotyping errors due to the relatively high genotyping error rates and low success rates. They used the error-incorporating misidentification model proposed by Lukacs & Burnham (2005) [71] (hereafter model L&B), implemented in the MARK program [95]. However this model relies on several assumptions that are unlikely to be met in practice and does not adequately address ghost

Figure 3.2: European otter (*Lutra lutra*) at the British wildlife centre (Surrey), by karen Bullock
`https://www.flickr.com/photos/karen_cb/`

[93]: Lampa et al. (2015), 'Non-Invasive Genetic Mark-Recapture as a Means to Study Population Sizes and Marking Behaviour of the Elusive Eurasian Otter (Lutra lutra)'

[94]: Kruuk (2006), *Otters: Ecology, behaviour and conservation*

[71]: Lukacs et al. (2005), 'Estimating population size from DNA-based closed capture-recapture data incorporating genotyping error'

Table 3.3: Closed population model estimates of the otter population in Upper Lusatia (Saxony, Germany). N is the population size, and here $N\alpha$ indicates the number of misidentifications. The values are the mean estimate ± the standard error.

	Model	2006	2007	2008	2010	2011	2012
N	Yoshizaki's	15.1 ± 0.8	20.4 ± 1.4	14.7 ± 0.5	15.2 ± 0.2	23.6 ± 0.9	18.4 ± 0.3
	$M_{\lambda,\alpha}$	16.1 ± 0.4	22.7 ± 0.9	14.3 ± 0.5	16.1 ± 0.3	23.6 ± 0.8	18.1 ± 0.3
	M_t	21.7 ± 0.9	30.5 ± 1.4	20.4 ± 0.7	18.1 ± 0.2	24.2 ± 0.4	21.1 ± 0.3
	L&B	19.0 ± 2.6	24.0 ± 3.5	19.4 ± 2.2	15.4 ± 2.1	25.1 ± 1.8	21.5 ± 2.2
$N\alpha$	$M_{\lambda,\alpha}$	5.8	7.5	7.8	4.9	2.5	5.9
	L&B	8.3	12.9	4	17.6	3.9	13

capture histories resulting from misidentification [67, 74]. After model selection, the authors of the previous study kept the model M_0 without individual or time heterogeneity. For simplicity and because we expect no substantial difference, we applied our $M_{\lambda,\alpha}$ model with time-dependent capture. We also applied Yoshizaki's model (discarding single capture histories) and the standard model M_t, in which misidentifications are ignored. The identification rates from the L&B model and the $M_{\lambda,\alpha}$ cannot be compared directly because the L&B model consider only one capture of an individual per occasion. We will compare the number of estimated misidentifications. When estimating the population size, we took Yoshizaki's model estimates as reference, as it estimates one parameter less and does not rely on any strong assumption.

The estimates are shown in Table 3.3. For 2011, all models estimate a similar population size. For all other years, the M_t model overestimates the population size compared to the other models. For most years, the $M_{\lambda,\alpha}$ model and Yohizaki's model have similar estimates. The minimum population size (numbers of observed individuals that cannot be ghosts considering the hypothesis of the LMM) are 16, 22, 14, 16, 23, 18 for each year respectively. For 2006, 2007 and 2010, the population size estimated by Yoshizaki's model is under this minimum.
Considering the hypothesis of the LMM, the maximum number of misidentifications were 6, 8, 8, 5, 3 and 6 for each year respectively. For all years, the model $M_{\lambda,\alpha}$ estimated that all the histories with a single capture were probably ghosts. In contrast, for most years, the mean estimate identification rates from model L&B indicated more misidentifications than our assumptions allowed. Other histories than those with only one capture would need to contain misidentifications. Compared to the simulations we conducted, the otter dataset was on the favourable side for good estimates. The capture rates were higher than what were tested in the simulations, and the identification rate was around the highest tested (95%). However, the population size was very small compared to the simulations.

[95]: White et al. (1999), 'Program MARK'

[67]: Yoshizaki et al. (2011), 'Modeling misidentification errors that result from use of genetic tags in capture–recapture studies'

[74]: Link et al. (2010), 'Uncovering a Latent Multinomial'

3.5 Discussion

In this chapter, we developed and implemented an extension of the LMM to account for repeated observations of individuals on the same occasion. We used a Poisson process to model the counts, replacing the Bernoulli

detection process. This new model demonstrates the flexibility of the LMM framework. With this extension of the model, we have removed the constraining assumption that individuals cannot be captured more than once on the same occasion. This model, which deals with misidentification in the presence of repeated observations, could be used in the many studies of mammals that use faecal DNA. Using it will help confirm whether or not the data contain misidentifications, if the estimated α is 1 or less than 1. And if misidentifications are present, then it can estimate the population size without bias.
We also compare our model to Yoshizaki's model. To use Yoshizaki's model, we remove the multiple observations from the data, keeping a maximum of one per individual and occasion. Yoshizaki's model also estimates the population size correctly, but its estimated confidence interval is higher than that of the $M_{\lambda,\alpha}$. This is probably due to the fact that Yoshizaki's model does not use repeated observations but uses only the first observation on each occasion, and also discard all the potential ghost.

We also applied the $M_{\lambda,\alpha}$ model to a Eurasian Otter data-set. The model estimates the population size as Yoshizaki's model for most of the years, and it confirms the thoughts of Lampa et al. (2015) [93] that misidentifications were present in the datasets. However the population size is quite small, and the variance of the observed counts in each year's dataset is in most cases larger than the mean, up to four times. Some simulations with large population sizes but similar overdispersion to that observed suggest that overdispersion is not in itself a problem. However, for other simulations with a small population size (and no overdispersion) the population size was often underestimated and the identification rate was underestimated.

[93]: Lampa et al. (2015), 'Non-Invasive Genetic Mark-Recapture as a Means to Study Population Sizes and Marking Behaviour of the Elusive Eurasian Otter (Lutra lutra)'

The $M_{\lambda,\alpha}$ model performs poorly in a similar way to the model $M_{t,\alpha}$ with very low capture probabilities (i.e. ≤ 0.1). To mitigate this problem, as in the previous chapter, we can use a probit model to estimate the misidentifications through an identification quality covariate. As the probit model was shown to improve the results compared to the original $M_{t,\alpha}$ model, it would most likely also improve the estimates of the $M_{\lambda,\alpha}$ model. Since more identifications are made when several observation can be made of an individual at the same occasion, the probit approach is probably even more interesting with this model.

Chapter four

Open population modelling of misidentification

4

4.1 Aim of chapter

So far, we have presented the LMM and some extensions for dealing with misidentification in closed population. However, a large proportion of capture-recapture experiments focus on studying the survival rates in open populations, where births and deaths (and eventually immigration and emigration) can occur. The LMM has already been extended to the open population [79]. However, this has been done for a different type of misidentification than the one I consider in this thesis. Bonner et al. (2015) [79] consider the case where two individuals are mistaken for each other: the one actually seen is noted as unseen while the one actually unseen is noted as seen. This is not relevant to our case where we assume that misidentifications only result in the creation of ghosts. Therefore, in this chapter I present several extensions of the model to open populations. I consider extensions for single-state, multi-state, and single-state with a probit model for an identification covariate. I use simulations to show the importance of modelling the misidentifications in open population models, and to evaluate the performance of the single-state model. The other models were developed as necessary steps for the next chapter and so were not tested with repeated simulations.

[79]: Bonner et al. (2015), 'Extending the latent multinomial model with complex error processes and dynamic markov bases'

4.2 Single-state open population models

4.2.1 Notations

Parameters

- p_t: Probability that an individual is captured at time t,
- α: Probability that a captured individual correctly identified
- ϕ: probability that an individual survives from one occasion to the next one.
- $\psi_{r,s}$: in multistate, probability that an individual transition from state r to state s between two consecutive occasions.

Data and Latent variables

- ω_i: Observed history i
- ν_j: Latent error history j (in which misidentification are noted down)
- ξ_k: Latent capture history k (real capture history)

Statistics

- N: number of individuals seen at least once,
- $(n_t) = (n_1, ..., n_T)$: number of individuals first seen at each occasion,
- y_i: number of observed history i
- x_j: number of latent error history j

- z_k: number of latent capture history k
- $\mathbf{y} = (y_1, ..., y_{2^T-1})$): vector of counts of observed histories
- $\mathbf{x} = (x_1, ..., x_{3^T})$): vector of counts of latent error histories
- $\mathbf{z} = (z_1, ..., z_{2^T})$): vector of counts of latent capture histories

4.2.2 Single state open population: Cormack-Jolly-Seber

To study the probability of survival of individuals in a population over a period of time (say daily), individuals can be captured and released at regular intervals of a duration of the period of interest (one day) T times. Individuals are captured with probability p_t at occasion t. When captured for the first time an individual is marked to be recognised and a capture history is created for it. When recaptured at a later occasion a 1 is registered in the history. Otherwise if not recaptured for an occasion, a 0 is registered in its history. I will consider losses on capture, i.e. the possibility of not releasing a captured animal, because it was killed when manipulated. When this happens at occasion t, a -1 is registered in the history at occasion $t + 1$, indicating the end of the history. Assuming we cannot recover a dead individual, this leads to N capture histories composed of 0s, 1s and -1s.

With the Cormack-Jolly-Seber model (CJS) ([28], [29]), the likelihood of an history is conditional on the first capture. Then, each of the histories can be seen as the combination of two independent partial histories. The first corresponds to the consecutive sightings of an individual (for which we know it is alive). It is of the form $1, ..., 1$. The second one corresponds to the last observation and beyond, and is of the form $1, 0, ..., 0$. These two parts being independent of each other, the likelihood of an history can be separated in two.

[28]: Jolly (1965), 'Explicit estimates from capture-recapture data with both death and immigration-stochastic model'

[29]: Seber (1965), 'A note on the multiple-recapture census'

The likelihood of history ω_i from the first capture at occasion C to the last capture at occasion D is

$$L_{i,1} = \phi_C p_D \prod_{t=C+1}^{D-1} \left[\phi_t P_{i,t}\right] \tag{4.1}$$

where $P_{i,t} = p_t^{I(\omega_{i,t} \in \{-1,1\})}(1 - p_t)^{I(\omega_{i,t}=0)}$ and $I(test)$ equals 1 if $test$ is true and 0 otherwise.

The likelihood of history ω_i, from the last capture at $t = D$ to the end of study, conditionally on the fact the individual was released at D, is given as a recursion by Catchpole et al. (1998) [96]. If χ_t is the probability that an individual, alive at t, is not seen again after, then:

[96]: Catchpole et al. (1998), 'Integrated recovery/recapture data analysis'

$$\chi_t = \begin{cases} (1 - \phi_t) + \phi_t(1 - p_{t+1})\chi_{t+1} & \text{if} \quad D \leq t \leq T - 1 \\ 1 & \text{if} \quad t = T \end{cases} . \tag{4.2}$$

Then, the complete likelihood of history ω_i is $\pi_i = L_{i,1} \cdot \chi_D$ and the vector of counts of the observed histories, $\mathbf{y}$ is a product multinomial random variable with density

$$[\mathbf{y}|\phi, \mathbf{p}] = \frac{\prod_{t=1}^{T} n_t!}{\prod_i y_i!} \prod_i \pi_i^{y_i}, \tag{4.3}$$

where n_t is the number of individuals first seen at occasion t.

4.2.3 Open population LMM

LMM likelihood reminder

Capture process:

$$[\mathbf{z}|N, \mathbf{p}] = \frac{N!}{\prod_k z_k!} \prod_k \pi_k^{z_k}$$

With single state histories probabilities:

$$\pi_k = \prod_1^T p_t^{I(\xi_{k,t}=1)} (1-p_t)^{I(\xi_{k,t}=0)}$$

Identification process:
If $I(\mathbf{z} = \mathbf{Bx})$, then

$$[\mathbf{x}|\mathbf{z}, \alpha] = \frac{\prod_k z_k!}{\prod_j x_j!} \prod_j \left[\prod_{t=1}^T A_{j,t} \right]^{x_j}$$

$$A_{j,t} = \alpha^{I(v_{k,t}=1)} (1-\alpha)^{I(v_{k,t}=2)}$$

In Section 1.2.3 we broke the likelihood of the LMM into two components: the capture one and the identification one. For the model we will call CJS_α, the likelihood will take the same form, with one component for the capture process and one for the identification process. Similarly to Section 1.3.3 and Section 2.2.2 we can now change these components to create a CJS_α model for an open population experiment. As we already presented the model $M_{t,\alpha}$ in Section 1.2.3 and the CJS in Section 4.2.2, we already described the needed processes. The likelihood is

$$[\mathbf{y}, \mathbf{x}, \mathbf{z}|(n_t), \phi, \mathbf{p}, \alpha] = I(\mathbf{y} = \mathbf{Ax})\,[\mathbf{x}|\mathbf{z}, \alpha]\,[\mathbf{z}|(n_t), \phi, \mathbf{p}] \tag{4.4}$$

The capture process for a CJS_α is simply the capture process of the CJS so the likelihood is the same. Thus, the likelihood is computed with Equation 4.3. The probabilities π_i are calculated by replacing observed histories ω by the latent capture histories ξ. The values of y_i are also replaced by those of the z_k and the n_t are also the ones corresponding to the latent capture histories.

$$[\mathbf{z}|\phi, \mathbf{p}] = \frac{\prod_{t=1}^T n_t!}{\prod_i z_i!} \prod_k \pi_k^{z_k} \tag{4.5}$$

The identification process is exactly the same for the CJS_α model as for the $M_{t,\alpha}$. As such, the first capture can very well be a misidentification. Hence, the number of individuals first captured at an occasion will change from one iteration to another in the MCMC.

4.2.4 Estimating the parameters of the CJS_α

The MCMC is constructed this way:

1. Let $\beta(a_0^t, b_0^t)$ denote the beta prior on p_t, $\beta(a_0^\alpha, b_0^\alpha)$ denote the beta prior on α and $\beta(a_0^\phi, b_0^\phi)$ the beta prior on ϕ.

2. Initialize all parameters as well as a set of latent histories satisfying **y**=**Ax**. Such a set can be obtained by assuming that no mistakes were made. The latent frequencies of the histories containing 2's are 0 and all the other match the observed frequencies one-to-one. In order to run several chains with different initialisations, one can take the previous initialisation of **x** and follow the later steps of (6) by only adding misidentification to the set and always accepting the proposed ones without going though the Metropolis-Hasting acceptance.

3. Sample the capture rate with Gibbs sampling. The likelihood being multinomial, it follows that the beta priors lead to full conditional beta posterior distribution:

$$p_t|\mathbf{z}, \phi \sim \beta(a_0^t + a^t, b_0^t + b^t)$$

where a^t is the number of individuals seen at least once before t that were captured at t and b^t is the number of individuals seen at least once before t that were alive but unseen at t [1].

1: To compute b^t, the latent state 'dead or alive' of each individual must be sampled. The probability that an individual last released before t and never seen again is still alive at t can be computed with the forward-backward algorithm.

4. Sample the survival rate with Gibbs sampling. Again, it has a full conditional beta posterior distribution.

$$\phi|\mathbf{z}, \mathbf{p} \sim \beta(a_0^\phi + a^\phi, b_0^\phi + b^\phi)$$

where a^ϕ is the total number of of times an individual survived from one occasion to the next after its first sight and b^ϕ is the total number of individual seen once or more that have died during before the last occasion. After the last release of an individual, there is no way of knowing it survived or died, but the transition toward one state or another can be sampled [2].

2: This again is done using the forward and backward algorithms.

5. Sample the identification rate using Gibbs sampling. Similar to the capture rate, it has a full conditional beta posterior distribution:

$$\alpha|\mathbf{x} \sim \beta(a_0^\alpha + a^\alpha, b_0^\alpha + b^\alpha)$$

where a^α [3] is the total number of correct identifications and b^α [4] is the total number of misidentifications.

3: $a^\alpha = \sum_j \sum_t x_j I(v_{j,t} = 1)$

4: $b^\alpha = \sum_j \sum_t x_j I(v_{j,t} = 2)$

6. Sample $\mathbf{x}$. Sampling $\mathbf{x}$ is done the exact same way as it is with model $M_{t,\alpha}$. It requires to be able to sample from $\mathcal{F}_\mathbf{y}$. To do so, sample moves from a dynamic Markov basis [83] which is the set of moves $M(x)$ that connect each $\mathbf{x}$ to some neighbours. Randomly add or remove an error from the set of latent histories. To add an error, sample a history that may have generated a ghost (i.e. a history containing a 0), and "merge" it with a potential ghost (i.e. replace the 0 by a 2 and remove the ghost history). To remove an error, sample a history containing a 2, replace it by a 0 and add a history with a unique capture (coded 1) at that time.
   More formally, follow the steps:

   a) Define:
      - ν_{1t} the history with a unique capture at time t (potential ghost),
      - $\chi_{0,t}(\mathbf{x}) = \{\nu | \nu_t = 0, x_\nu > 0, x_{\nu_{1t}} > 0\}$ the set of histories having *potentially* generated a ghost at time t, for the given $\mathbf{x}$,
      - $\chi_{2,t}(\mathbf{x}) = \{\nu | \nu_t = 2, x_\nu > 0\}$ the set of histories *containing* a ghost at time t, for the given $\mathbf{x}$.

   b) With probability 0.5, go to (i), otherwise go to (ii).

      i. Add a misidentification (i.e. a ghost) to the latent set.
         - Sample $\boldsymbol{\nu}^{(0)} \in \chi_{0.}(x) = \bigcup_t \chi_{0,t}(x)$.
         - Sample $t \in \{t | \nu_t^{(0)} = 0, x_{\nu_{1t}} > 0\}$.
         - Define $\boldsymbol{\nu}^{(2)} = \boldsymbol{\nu}^{(0)} + 2\boldsymbol{\nu}^{(1t)}$.
         - Define the move $b_{\boldsymbol{\nu}^{(0)}, \boldsymbol{\nu}^{(1t)}, \boldsymbol{\nu}^{(2)}} = (-1, -1, +1)$, and $b_\nu = 0$ for all other latent histories.

      ii. Remove a misidentification from the latent set.
         - Sample $\boldsymbol{\nu}^{(2)} \in \chi_{2.}(x) = \bigcup_t \chi_{2,t}(x)$.
         - Sample $t \in \{t | \boldsymbol{\nu}_t^{(2)} = 2\}$.
         - Define $\boldsymbol{\nu}^{(0)} = \boldsymbol{\nu}^{(2)} - 2\boldsymbol{\nu}^{(1t)}$.

[83]: Dobra (2012), 'Dynamic markov bases'

- Define the move $b_{\nu^{(0)},\nu^{(1t)},\nu^{(2)}} = (+1, +1, -1)$, and $b_\nu = 0$ for all other latent histories.

c) Define $\mathbf{x}' = \mathbf{x}^{(k-1)} + b$.
d) Calculate $\mathbf{z}' = \mathbf{B}\mathbf{x}'$.
e) Calculate the proposal numbers of first capture per occasion $(n_t)'$.
f) Set $\mathbf{x}^{(k)} = \mathbf{x}'$ with probability r. Otherwise set $\mathbf{x}^{(k)} = \mathbf{x}^{(k-1)}$.

$$r = min\left(1, \frac{[\mathbf{y}, \mathbf{x}', \mathbf{z}'|(n_t)', \phi', p, \alpha]}{[\mathbf{y}, \mathbf{x}^{(k-1)}, \mathbf{z}^{(k-1)}|(n_t), \phi, p, \alpha]} \frac{q(\mathbf{x}^{(k-1)}|\mathbf{x}')}{q(\mathbf{x}'|\mathbf{x}^{(k-1)})}\right) \quad (4.8)$$

7. repeat steps 3 to 6 as much as needed.

The proof of convergence is once again the same as in Section 1.2.5. The first condition becomes that Step 3 and 4 produces chains which converge to $\pi(\phi, \mathbf{p}, |\mathbf{z})$ for any $\mathbf{z}$ such that $\mathbf{z} = \mathbf{B}\mathbf{x}$ for some $\mathbf{x} \in \mathcal{F}_y$. The validity of the condition is as trivial as it was in the original proof. The second and third conditions do not change, and their proof remain the same.

Proposal density

The proposal densities are calculated by multiplying the probabilities of each sampling step used for defining the move. They are successively: the probability of adding (or removing) an error, the probability of choosing the ν_0 (or ν_2) and the probability of choosing the t knowing the sampled ν. When adding an error, the proposal density q is:

$$q(x'|x^{k-1}) = \frac{0.5}{\#\chi_0.\#\{t|\nu_{0,t} = 0, x_{\nu_{1t}} > 0\}} \quad (4.6)$$

and when removing an error, is:

$$q(x'|x^{k-1}) = \frac{0.5}{\#\chi_2.\#\{t|\nu_{2,t} = 2\}} \quad (4.7)$$

where $\#S$ denotes the cardinality of S.

4.2.5 Covariate of identification in open-population

We can also extend the CJS_α to include a covariate of identification as we did in Chapter 2. We develop this model (CJS_{α_n}) in this section. Starting with the model described in Chapter 2, we would only have to modify the capture likelihood to specify a CJS_{α_n}. To keep the chapters as independant as possible, we re-describe the identification process but note that it is the same as in Chapter 2. Because we cannot summarise the histories due to the covariate, we use the complete history sets $\mathbf{Y}$, $\mathbf{X}$, and $\mathbf{Z}$ instead of the frequency vectors $\mathbf{y}$, $\mathbf{x}$ and $\mathbf{z}$. (Note that the notations $\mathbf{x}$ and $\mathbf{z}$ are still used to account for the number of matrices identical, with the exception of individual labelling.) Let $\alpha_{n,t}$ be the probability of correct identification for individual n at occasion t. If n was not captured at t, we take $\alpha_{n,t} = 1$.

The likelihood, conditional to a given $\mathbf{X}$ is:

$$[\mathbf{Y}, \mathbf{X}, \mathbf{Z}|(n_t), \phi, \mathbf{p}, \boldsymbol{\alpha}, \boldsymbol{\theta}_\alpha] = I(\mathbf{Y}|\mathbf{X})[\mathbf{X}|\mathbf{Z}, \boldsymbol{\alpha}][\mathbf{Z}|(n_t), \phi, \mathbf{p}][\boldsymbol{\alpha}|\boldsymbol{\theta}_\alpha], \quad (4.9)$$

Notations reminder

- ϕ: survival probability,
- $\mathbf{p} = (p_1, ..., p_T)$: capture probabilities,
- $\boldsymbol{\alpha} = (\alpha_{n,t})$: identification probabilities,
- (n_t): numbers of individuals first capture for each occasion,
- $\mathbf{Y}$: matrix of observed histories,
- $\mathbf{X}$: matrix of latent error histories,
- $\mathbf{Z}$: matrix of latent capture histories.

- $I(\mathbf{Y}|\mathbf{X})$

Let's define a function f such that, for a latent error history ν_j, it results in the corresponding set of observed histories (ω_i): $f(\nu_j) = (\omega_i)$. For example, $f((1,1,2)) = \{(1,1,0),(0,0,1)\}$. If we apply f to all the latent histories in $\mathbf{X}$, the resulting set of histories (ω_i) must be equal to $\mathbf{Y}$, except for index inversions. An example is given on Figure 2.1 in Section 2.2.2. Thus, $I(\mathbf{Y}|\mathbf{X})$ is 1 if $f(\mathbf{X}) = \cup_j f(\nu_j) = \mathbf{Y}$ and 0 otherwise. We write this as $I(\mathbf{Y} = f(\mathbf{X}))$.

- $[\mathbf{X}|\mathbf{Z}, \boldsymbol{\alpha}]$

First, we rewrite the part $I(\mathbf{z} = \mathbf{B}\mathbf{x})$ of the Equation 1.5. (Remember that the identification process was the same for the models $M_{t,\alpha}$ and CJS_α.) As for $I(\mathbf{Y}|\mathbf{X})$, let's define a function g that, for a latent error history ν_j, results in the corresponding latent capture history ξ_j: $g(\nu_j) = \xi_j$. For

example $g((1,1,2)) = (1,1,1)$. If we apply g to all histories in $\mathbf{X}$, the resulting set of histories (ξ_j) must be equal to $\mathbf{Z}$. An example is given in Figure 2.2 in Section 2.2.2. Thus, $I(\mathbf{X}|\mathbf{Z})$ is 1 if $g(\mathbf{X}) = \cup_j g(v_j) = \mathbf{Z}$ and 0 otherwise. We write this as $I(\mathbf{Z} = g(\mathbf{X}))$.

The identification likelihood is a product of Bernoulli trials for all captured individuals at all times they were captured. The product of factorial denotes the various ways of ordering the individuals.

$$[\mathbf{X}|\mathbf{Z}, \boldsymbol{\alpha}] = I(\mathbf{Z} = g(\mathbf{X})) \frac{\prod_k z_k!}{\prod_j x_j!} \prod_{n=1}^{N} \prod_{t=1}^{T} \alpha_{n,t}^{I(v_{n,t}=1)} (1 - \alpha_{n,t})^{I(v_{n,t}=2)} \ . \quad (4.10)$$

- $[\mathbf{Z}|\mathbf{p}, \phi, (n_t)]$

This part is very similar to the capture process of the CJS_α:

$$[\mathbf{Z}|(n_t), \phi, \mathbf{p}] = \frac{\prod_{t=1}^{T} n_t!}{\prod_k z_k!} \prod_{n=1}^{N} \pi_n, \quad (4.11)$$

where π_n is computed as in Section 4.2.2. As the capture probabilities are constant across individuals, the capture likelihood simplifies as Equation 4.5

- $[\boldsymbol{\alpha} \mid \boldsymbol{\theta}_\alpha]$

For this part, similar to Mc clintock et. al (2014) [75], we chose to develop a probit model. Other links could be used, especially since there are no missing covariates. The probit model gives us

[75]: McClintock et al. (2014), 'Probit models for capture-recapture data subject to imperfect detection, individual heterogeneity and misidentification'

$$\alpha_{n,t} = \boldsymbol{\phi}(a \cdot \tau_{n,t} + b)$$

where $\boldsymbol{\phi}$ is the standard normal cumulative distribution function. Thus, $\boldsymbol{\theta}_\alpha = (a, b)$. We propose a model where $b \neq 0$. To understand why, let's consider what would happen if we kept a sample for which $\tau = 0$ (i.e. having observed no loci at all for that sample). In that case, we could only randomly assign the sample in an already existing history or in a new one. The probability of putting it in the right history would be very low. On the other hand, if τ is large enough (i.e. having observed most loci with good confidence), the probability of misidentifying the sample would be very small. Thus we want $\alpha = \boldsymbol{\phi}(a \cdot 0 + b) \approx 0$, so $b < 0$.

To fully specify the probit model, we define $u_{n,t}$ as a binary indicator of the success of the identification of the capture of individual n at occasion t. That is, $u_{n,t} = 1$ if the sample n, t resulted in a correct individual identification, and 0 otherwise. We also define $\tilde{u}_{n,t}$, a continuous latent process of $u_{n,t}$. We set $\tilde{u}_{n,t} \sim \mathcal{N}(a\tau_{n,t} + b, 1)$ and if $\tilde{u}_{n,t} < 0$ then $u_{n,t} = 0$, or else if $\tilde{u}_{n,t} > 0$ then $u_{n,t} = 1$.

We note $\mathbf{u} = (u_{n,t})_{n\in[1,N],t\in[1,T]}$ and $\tilde{\mathbf{u}} = (\tilde{u}_{n,t})_{n\in[1,N],t\in[1,T]}$ Since all covariates are known, conditional on $\mathbf{X}$, all the $u_{n,t}$ are known. So the definition of $\tilde{u}_{n,t}$ is not really needed, but it does allow for Gibbs sampling of a and b (see Section 2.2.3).

We only have left to specify priors for a and b:

$$a \sim \mathcal{N}(\mu_a, \sigma_a^2),$$

$$b \sim \mathcal{N}(\mu_b, \sigma_b^2).$$

4.2.6 Estimating the parameters of the CJS_{α_n}

To construct the MCMC, the parameters ϕ and $\mathbf{p}$ are sampled as described in Section 4.2.3 while $\boldsymbol{\theta}_\alpha$ is sampled as in Section 2.2.3. The matrix of latent histories $\mathbf{X}$ is sampled as in Section 2.2.3 but updating the numbers of individuals first captured (n_t) instead of the total number of individuals. The MH ratio given by Equation 2.7 does not changes but the history probabilities are computed same as for the likelihood (see Section 4.2.2).

The MCMC is constructed this way:

1. Let $\beta(a_0^t, b_0^t)$ denote the beta prior on p_t, $\mathcal{N}(\mu_a, \sigma_a^2)$ the normal prior on a and $\mathcal{N}(\mu_b, \sigma_b^2)$ the normal prior on b.

2. Initialize all parameters as well as a set of latent histories satisfying $I(\mathbf{Y} = f(\mathbf{X}))$. Such a set can be obtained by assuming that no mistakes were made, i.e. the exact set of observed histories. In order to run several chains with different initialisations, one can take the previous initialisation of $\mathbf{X}$ and follow the later steps of (5) by only adding misidentifications to the set and always accepting the proposed ones without going though the Metropolis-Hasting acceptance. In the initial latent set, fix a random realistic number of all-zero histories to be part of the population.

3. Sample the capture rate with Gibbs sampling. The likelihood being multinomial, it follows that the beta priors lead to full conditional beta posterior distribution:

$$p_t|\mathbf{Z}, \phi \sim \beta(a_0^t + a^t, b_0^t + b^t)$$

   where a^t is the number of individuals seen at least once before t that were captured at t and b^t is the number of individuals seen at least once before t that were alive but unseen at t [5].

4. Sample the survival rate with Gibbs sampling. Again, it has a full conditional beta posterior distribution.

$$\phi|\mathbf{Z}, \mathbf{p} \sim \beta(a_0^\phi + a^\phi, b_0^\phi + b^\phi)$$

   where a^ϕ is the total number of of times an individual survived from one occasion to the next after its first sight and b^ϕ is the total number of individual seen once or more that have died during before the last occasion. After the last release of an individual, there is no way of knowing it survived or died, but the transition toward one state or another can be sampled [6].

5. Sample the identification rate probit parameters with Gibbs sampling. With the probit model, the parameters a and b can be sampled in their full conditional posterior. First, the $\tilde{\mathbf{u}}$ also need to be updated by sampling the $\mathbf{u}_{n,t}$ in their full conditional posterior

5: To compute b^t, the latent state 'dead or alive' of each individual must be sampled. The probability that an individual last released before t and never seen again is still alive at t can be computed with the forward-backward algorithm.

6: This again is done using the forward and backward algorithms.

using the values of $\boldsymbol{\tau}$. For simplification of the notation, we introduce notation L denoting the total number of capture realised and $l = 1, \ldots, L$ the indexes of each capture. Then:

$$\begin{aligned} \tilde{u}_l|\cdot &\sim \begin{cases} \mathcal{TN}_{(0,+\infty)}(a\tau_l + b, 1) & \text{if} \quad \nu_l = 1, \\ \mathcal{TN}_{(-\infty,0)}(a\tau_l + b, 1) & \text{if} \quad \nu_l = 2, \end{cases} \\ a|\cdot &\sim \mathcal{N}(\mu'_a, \sigma'^2_a), \\ b|\cdot &\sim \mathcal{N}(\mu'_b, \sigma'^2_b), \end{aligned} \tag{4.12}$$

where $\mathcal{TN}$ is the truncated normal distribution and

$$\begin{cases} \sigma'^2_a = \left(\frac{1}{\sigma_a^2} + \sum_{l=1}^{L} \tau_l^2\right)^{-1}, \\ \mu'_a = \sigma'^2_a \left(\frac{\mu_a}{\sigma_a^2} + \sum_{l=1}^{L} \tau_l(\tilde{u}_l - b)\right), \end{cases} \tag{4.13}$$

and

$$\begin{cases} \sigma'^2_b &= \frac{\sigma_b^2}{L\sigma_b^2+1}, \\ \mu'_b &= \sigma'^2_b \left(\frac{\mu_b}{\sigma_b^2} + \sum_{l=1}^{L}(\tilde{u}_l - a\tau_l)\right). \end{cases} \tag{4.14}$$

6. Sample $\mathbf{X}$ with Metroplis-Hastings. In order to propose an $\mathbf{X}'$, the definitions of sets that where sampled ($\chi_{0,t}(x)$ and $\chi_{2,t}(x)$) need to be changed. Let $\chi_{0,t}(\mathbf{X}) = \{i|\nu_{i,t} = 0\}$ be the set of individual which were unseen at time t in their latent error history if at least one individual is seen only at occasion t. Otherwise $\chi_{0,t}(\mathbf{X}) = \emptyset$. Let $\chi_{2,t}(\mathbf{X}) = \{i|\nu_{i,t} = 2\}$ be the set of individual which were misidentified at time t in their latent error history. Finally, let $\chi_{1,t}(\mathbf{X}) = \{i|\nu_{i,t} = 1, (\nu_{i,s})_{s\neq t} = 0\}$ be the set of individual which were only seen once, at time t, in their latent error history. Then, to generate $\mathbf{X}^{(k)}$, use the following steps:
   a) Set $\mathbf{X}' = \mathbf{X}^{(k-1)}$.
   b) With probability 0.5 go to step c, otherwise go to step d.
   c) Add a ghost to the proposal set of latent histories $\mathbf{X}'$.
      i. Sample uniformly $t \in \{t|\chi_{0,t}(\mathbf{X}) \neq \emptyset\}$, the set of occasions for which at least one individual is unseen and one individual is only seen at that occasion.
      ii. Sample uniformly $i_0 \in \chi_{0,t}(\mathbf{X})$ the set of unseen individuals at occasion t.
      iii. Sample $i_1 \in \chi_{1,t}(\mathbf{X})$ proportionally to $1 - \alpha_{i_1,t}$.
      iv. Set $\nu'_{i_0,t} = 2$, with the covariate of identification associated to $\nu_{i_1,t}$.
      v. Remove ν_{i_1} from the set of individuals.
      vi. Go to step e.
   d) Remove a ghost from the proposal set of latent histories $\mathbf{X}'$.
      i. Sample uniformly $t \in \{t|\chi_{2,t}(\mathbf{X}) \neq \emptyset\}$, the set of occasions where at least one misidentification is present.
      ii. Sample $i_2 \in \chi_{2,t}(\mathbf{X})$ proportionally to $\alpha_{i_2,t}$.
      iii. Add an individual with a single capture at time t with the covariate of identification that is associated to $\nu_{i_2,t}$.
      iv. Set $\nu'_{i_2,t} = 0$.
   e) Compute $\mathbf{Z}' = g(X')$. Set n' as the number of individuals in $\mathbf{X}'$. Set (n'_t) the numbers of individuals first caught per occasion.
   f) With probability r, set $\mathbf{X}^{(k)} = \mathbf{X}'$, $\mathbf{Z}^{(k)} = \mathbf{Z}'$ and $(n_t^{(k)}) = (n'_t)$.

Otherwise set $\mathbf{X}^{(k)} = \mathbf{X}^{(k-1)}$, $\mathbf{Z}^{(k)} = \mathbf{Z}^{(k-1)}$ and $(n_t^{(k)}) = (n_t^{(k-1)})$.

$$r = min\left(1, \frac{\prod_t n_t'! \prod_{i=1}^{N'} \pi_i}{\prod_t n_t! \prod_{i=1}^{N} \pi_i} \frac{q(\mathbf{X}^{(k-1)}|\mathbf{X}', \boldsymbol{\alpha})}{q(\mathbf{X}'|\mathbf{X}^{(k-1)}, \boldsymbol{\alpha})}\right), \tag{4.15}$$

where $\pi_i = L_{i,1} \cdot \chi_D \prod_t \alpha_{i,t}^{I(v_{i,t}=1)}(1-\alpha_{i,t})^{I(v_{i,t}=2)}$ (see Equation 4.1 and Equation 4.2 for $L_{i,1}$ and χ_D) and $[\mathbf{X}'|\mathbf{X}, \boldsymbol{\alpha}]$ is the proposal density for $\mathbf{X}'$. When adding a ghost:

$$[\mathbf{X}'|\mathbf{X}, \boldsymbol{\alpha}] = \frac{0.5(1-\alpha_{v_1,t})}{\sum_{i\in\chi_{1,t}(\mathbf{X})}(1-\alpha_{i,t})\#\{t|\chi_{0,t}(\mathbf{X}) \neq \emptyset\}\#\chi_{0,t}} \tag{4.16}$$

and when removing a ghost:

$$[\mathbf{X}'|\mathbf{X}, \boldsymbol{\alpha}] = \frac{0.5\alpha_{v_2,t}}{\sum_{i\in\chi_{2,t}(\mathbf{X})}(\alpha_{i,t})\#\{t|\chi_{2,t}(\mathbf{X}) \neq \emptyset\}} \tag{4.17}$$

where $\#S$ denotes the cardinal of ensemble S.

7. repeat steps 3 to 6 as much as needed.

The proof of convergence of this algorithm is the same as in Section 1.2.4. Some notations change, due to the moves not being written as vectors. In the proof, we replace the definition of moves as vectors by functions of a set of latent histories: $b(\mathbf{X}) \in \mathcal{M}(\mathbf{X})$. The reverse move of $b(\mathbf{X})$ is $b^{-1}(b(\mathbf{X}))$. The full proof can be re-written with these new notation. It will stay the same so the proof is still valid.

4.3 Multi-state open population models

4.3.1 Multistate open population: Arnason-Schwarz model

If the individuals can be observed in different states (corresponding to biological states, locations...), then the state they are in when seen can be registered. We assume the state is always identified without error. Assuming S states ($s \in \mathcal{S}$) can be observed, the capture histories are composed of S+1 different numbers (0 if not captured and $1, \ldots, S$ for the S states). Additionally, on a loss on capture, a -1 is recorded at $t+1$. King et al. (2003) [97] give a closed form of the commonly named Arnason-Schwarz model (first developed by Arnason & Schwarz[30–32, 98]).

Let $O_{(c,d)}(r,s)$ denote the probability that an animal in state $r \in \mathcal{S}$ at time c remains unobserved until it is subsequently resighted in state $s \in \mathcal{S}$ at time $d+1$. Then, for $c \leq d$,

$$O_{(c,d)}(r,s) = p_{d+1,s} Q_{(c,d)}(r,s), \tag{4.18}$$

where $Q_{(c,d)}(r,s)$ denotes the probability that an animal changes from state r at time c to state s at time $d+1$, and is unobserved between these

[97]: King et al. (2003), 'Closed-form likelihoods for Arnason-Schwarz models'

[30]: Neil Arnason (1972), 'Parameter estimates from mark-recapture experiments on two populations subject to migration and death'

[31]: Neil Arnason (1973), 'The estimation of population size, migration rates and survival in a stratified population'

[32]: Schwarz et al. (1993), 'Estimating migration rates using tag-recovery data'

[98]: Schwarz et al. (1996), 'A general methodology for the analysis of capture-recapture experiments in open populations'

times, and is given by

$$Q_{(c,d)}(r,s) = \begin{cases} \phi_{r,c}\psi_{r,s} & \text{if} \quad c = d \\ \phi_{r,c} \sum_{l \in \mathcal{S}} \left[(1 - p_{c+1,l}) \psi_{r,l} Q_{(c+1,d)}(l,s) \right] & \text{if} \quad c < d \end{cases} . \tag{4.19}$$

Let $(c,d) \in \mathcal{CD}$ denote the pairs of consecutive times the individual was captured. Then the likelihood $L_{i,1}$ of the part of history ω_i that precedes the last sighting at D is

$$L_{i,1} = \prod_{(c,d) \in \mathcal{CD}} O_{(c,d)}(r,s) \quad . \tag{4.20}$$

The second part of the likelihood of history ω_i, if the individual was released at D is $\chi_{r,D}$. It is the probability that an individual, alive at D in state r, is not seen again afterwards. It is given by

$$\chi_{r,t} = \begin{cases} (1 - \phi_{r,t}) + \phi_{r,t} \sum_{s \in \mathcal{S}} \left[\psi_{r,s} (1 - p_{s,t+1}) \chi_{s,t+1} \right] & \text{if} \quad D \leq t \leq T-1 \\ 1 & \text{if} \quad t = T \end{cases} . \tag{4.21}$$

Then, the complete likelihood of history ω_i is $\pi_i = L_{i,1} \cdot \chi_{r,D}$. And the vector of counts of the observed histories, $\mathbf{y}$ is a product multinomial random variable with density

$$[\mathbf{y} | \phi, \boldsymbol{\psi}, \mathbf{p}] = \frac{\prod_{s \in \mathcal{S}} \prod_{t=1}^{T-1} n_{s,t}!}{\prod_i y_i!} \prod_i \pi_i^{y_i} \quad , \tag{4.22}$$

where $n_{s,t}$ is the number of individuals first captured at t in state s.

4.3.2 Arnason-Schwarz LMM

Now that we've introduced the AS model, we can present a last general extension of the LMM, in open-population with multistate observations: the AS_α model. It should now be clear as to how this is done. The likelihood conditional to the set of latent histories is:

$$[\mathbf{y}, \mathbf{x}, \mathbf{z} | (n_{s,t}), \phi, \boldsymbol{\psi}, \mathbf{p}, \alpha] = I(\mathbf{y} = \mathbf{A}\mathbf{x}) [\mathbf{x} | \mathbf{z}, \alpha] [\mathbf{z} | (n_{s,t}), \phi, \boldsymbol{\psi}, \mathbf{p}] \tag{4.23}$$

The capture likelihood $[\mathbf{z} | (n_{s,t}), \phi, \boldsymbol{\psi}, \mathbf{p}]$ is the same as for the Arnason-Schwarz. This part is exactly the same as in Section 1.3.3. If we consider that the probability of correctly identifying an individual is constant between states, the identification part $[\mathbf{x} | \mathbf{z}, \alpha]$ does not change much compared to model $M_{t,\alpha}$. The latent error histories ν are composed of $2S+1$ different values: one per state when correctly identified, one per state when misidentified and the 0. The likelihood of the identification process is computed with Equation 1.5, rewriting $A_{j,t} = \alpha^{I(\nu_{j,t} \in [1,S])} (1-\alpha)^{I(\nu_{j,t} > S)}$. For example, if three states are considered, the identification likelihood of latent history $(1,4,0,2,6)$ is

$$A_{(1,4,0,2,6)} = \alpha \times (1-\alpha) \times 1 \times \alpha \times (1-\alpha) = \alpha^2 (1-\alpha)^2$$

State heterogeneity can be considered for the identification process by setting $\boldsymbol{\alpha} = (\alpha_1, ..., \alpha_S)$. In that case the previous example likelihood

would simply be

$$A_{(1,4,0,2,6)} = \alpha_1 \times (1 - \alpha_1) \times \alpha_2 \times (1 - \alpha_3)$$

4.3.3 Estimating the parameters of the AS_α

The construction of the MCMC is done the same way as for the CJS_α except that the transitions are also to be sampled.

The MCMC is constructed this way:

1. Let $\beta(a_0^t, b_0^t)$ denote the beta prior on p_t, $\beta(a_0^\alpha, b_0^\alpha)$ denote the beta prior on α, $\beta(a_0^\phi, b_0^\phi)$ the beta prior on ϕ and $\psi_{s,\cdot} \sim Dir(\alpha_0^{\psi_s})$ be the Dirichlet prior on $\psi_{s,\cdot}$.

2. Initialize all parameters as well as a set of latent histories satisfying **y**=**Ax**. Such a set can be obtained by assuming that no mistakes were made. The latent frequencies of the histories containing 2's are 0 and all the other match the observed frequencies one-to-one. In order to run several chains with different initialisations, one can take the previous initialisation of **x** and follow the later steps of (6) by only adding misidentification to the set and always accepting the proposed ones without going though the Metropolis-Hasting acceptance.

3. Sample the capture rate with Gibbs sampling. The likelihood being multinomial, it follows that the beta priors lead to full conditional beta posterior distribution:

$$p_t|\mathbf{z}, \phi, \boldsymbol{\psi} \sim \beta(a_0^t + a^t, b_0^t + b^t)$$

   where a^t is the number of individuals seen at least once before t that were captured at t and b^t is the number of individuals seen at least once before t that were alive but unseen at t [7].

4. Sample the survival rate with Gibbs sampling. Again, it has a full conditional beta posterior distribution.

$$\phi|\mathbf{z}, \mathbf{p}, \boldsymbol{\psi} \sim \beta(a_0^\phi + a^\phi, b_0^\phi + b^\phi)$$

   where a^ϕ is the total number of of times an individual survived from one occasion to the next after its first sight and b^ϕ is the total number of individual seen once or more that have died during before the last occasion. After the last release of an individual, there is no way of knowing it survived or died, but the transition toward one state or another can be sampled [8].

5. Sample the transition rates with Gibbs sampling. They have a full conditional beta posterior distribution.

$$\psi_{s,\cdot}|\mathbf{z}, \mathbf{p}, \psi \sim Dir(\mathbf{a_0}^{\psi_{s,\cdot}} + \mathbf{a}^{\psi_{s,\cdot}})$$

   where $\mathbf{a}^{\psi_{s,\cdot}}$ is the number of of times an individual transitioned from state s to the others. Just like for the initial states, we can't know what transition occurred for an unseen individual but it can also be sampled [9].

7: To compute b^t, the latent state 'dead or alive' of each individual must be sampled. The probability that an individual last released before t and never seen again is still alive at t can be computed with the forward-backward algorithm.

8: This again is done using the forward and backward algorithms.

9: This again is done using the forward and backward algorithms.

6. Sample the identification rate with Gibbs sampling. Similarly to the capture rate, it has a full conditional beta posterior distribution.

$$\alpha \sim \beta(a_0^\alpha + a^\alpha, b_0^\alpha + b^\alpha)$$

where a^α [10] is the total number of correct identifications and b^α [11] is the total number of misidentifications.

10: $a^\alpha = \sum_j \sum_t x_j I(\nu_{j,t} \in [1, S])$

11: $b^\alpha = \sum_j \sum_t x_j I(\nu_{j,t} = \in [S+1, 2S])$

7. Sample $\mathbf{x}$. Sampling $\mathbf{x}$ is done the exact same way as it is with the multistate model $M_{t,\alpha}$. It requires to be able to sample from $\mathcal{F}_\mathbf{y}$. To do so, sample moves from a dynamic Markov basis [83] which is the set of moves $M(x)$ that connect each $\mathbf{x}$ to some neighbours. Randomly add or remove an error from the set of latent histories. To add an error, sample a history that may have generated a ghost (i.e. a history containing a 0), and "merge" it with a potential ghost (i.e. replace the 0 by a 2 and remove the ghost history). To remove an error, sample a history containing a 2, replace it by a 0 and add a history with a unique capture (coded 1) at that time.
More formally, follow the steps:

[83]: Dobra (2012), 'Dynamic markov bases'

   a) Define:
      - $\boldsymbol{\nu}^{(1st)}$ the history with a unique capture at occasion t in state e_s (potential ghost),
      - $\chi_{0,s,t}(x) = \{\boldsymbol{\nu} | \nu_t = 0, x_\nu > 0, x_{\nu^{(1st)}} > 0\}$ the set of histories having *potentially* generated a ghost in state e_s at occasion t, for the given $\mathbf{x}$,
      - $\chi_{2,s,t}(x) = \{\boldsymbol{\nu} | \nu_t = s + S, x_\nu > 0\}$ the set of histories *containing* a ghost in state e_s at occasion t, for the given $\mathbf{x}$.
   b) Sample a state s uniformly from $1, ..., S$.
   c) With probability 0.5, go to (i), otherwise go to (ii).
      i. Add a misidentification (i.e. a ghost) to the latent set.
         - Sample $\boldsymbol{\nu}^{(0)} \in \chi_{0,s,.}(x) = \bigcup_t \chi_{0,s,t}(x)$.
         - Sample $t \in \{t | \boldsymbol{\nu}_t^{(0)} = 0, x_{\nu^{(1st)}} > 0\}$.
         - Set $\boldsymbol{\nu}^{(2)} = \boldsymbol{\nu}^{(0)}$ and then $\nu_t^{(2)} = s + S$.
         - Define the move $b_{\boldsymbol{\nu}^{(0)}, \boldsymbol{\nu}^{(1st)}, \boldsymbol{\nu}^{(2)}} = (-1, -1, +1)$ and $b_\nu = 0$ for all other latent histories.
      ii. Remove a misidentification from the latent set.
         - Sample $\boldsymbol{\nu}^{(2)} \in \chi_{2,s,.}(x) = \bigcup_t \chi_{2,s,t}(x)$.
         - Sample $t \in \{t | \boldsymbol{\nu}_{s,t}^{(2)} = S + s\}$.
         - Define $\boldsymbol{\nu}^{(0)} = \boldsymbol{\nu}^{(2)}$ and then $\nu_t^{(0)} = 0$.
         - Define the move $b_{\boldsymbol{\nu}^{(0)}, \boldsymbol{\nu}^{(1st)}, \boldsymbol{\nu}^{(2)}} = (+1, +1, -1)$ and $b_\nu = 0$ for all other latent histories.
   d) Define $\mathbf{x}' = \mathbf{x}^{(k-1)} + b$.
   e) Calculate $\mathbf{z}' = \mathbf{B}\mathbf{x}'$.
   f) Calculate the proposal numbers of first capture per occasion $(n_{s,t})'$.
   g) Set $\mathbf{x}^{(k)} = \mathbf{x}'$ with probability r. Otherwise set $\mathbf{x}^{(k)} = \mathbf{x}^{(k-1)}$.

$$r = min\left(1, \frac{[\mathbf{y}, \mathbf{x}', \mathbf{z}' | (n_{s,t})', \boldsymbol{\phi}, \boldsymbol{\psi}, \mathbf{p}, \alpha]}{[\mathbf{y}, \mathbf{x}, \mathbf{z} | (n_{s,t}), \boldsymbol{\phi}, \boldsymbol{\psi}, \mathbf{p}, \alpha]} \frac{q(\mathbf{x}^{(k-1)} | \mathbf{x}')}{q(\mathbf{x}' | \mathbf{x}^{(k-1)})}\right) \quad (4.26)$$

8. repeat steps 3 to 7 as much as needed.

The proof of convergence is once again the same as in Section 1.2.5. The first condition becomes that Step 3 and 4 produces chains which converge

Proposal density

The proposal densities are calculated by multiplying the probabilities of each sampling step used for defining the move. They are successively: the probability of adding (or removing) an error, the probability of choosing the ν_0 (or ν_2) and the probability of choosing the t knowing the sampled ν. When adding an error, the proposal density q is:

$$q(x' | x^{k-1}) = \frac{0.5}{\#\chi_{0.} \#\{t | \nu_{0,t} = 0, x_{\nu_{1t}} > 0\}} \quad (4.24)$$

and when removing an error, is:

$$q(x' | x^{k-1}) = \frac{0.5}{\#\chi_{2.} \#\{t | \nu_{2,t} = 2\}} \quad (4.25)$$

where $\#S$ denotes the cardinality of S.

to $\pi(\phi, \mathbf{p}, |\mathbf{z})$ for any $\mathbf{z}$ such that $\mathbf{z} = \mathbf{B}\mathbf{x}$ for some $\mathbf{x} \in \mathcal{F}_y$. The validity of the condition is as trivial as it was in the original proof. The second and third conditions do not change, and their proof remain the same.

4.4 Simulation study

4.4.1 Scenarios and implementation

From the models above, we tested only tested the CJS_α. We simulated ten datasets for each of the following scenario: survival $\phi = 0.6, 0.7, 0.8, 0.9$, capture probability $p_t = 0.3, 0.4, 0.5, 0.6, 0.7$, identification probability $\alpha = 0.8, 0.9, 0.95$. For every scenario, the initial population size is 500 and it was maintained approximately constant. We simulated constant capture rate but they were treated as time dependant by the models. These simulations have two objectives. The first is showing the impact of misidentification over the estimation of survival while the second is to show how the LMM perform in the various scenarios of survival, capture and identification rates. hence, for all simulations we ran both the CJS and the CJS_α.

The CJS_α was implemented with NIMBLE ([80]) and the CJS was ran using the same scripts but without the samplers for the set of latent histories $\mathbf{x}$ and the identification probability α were removed. The MCMC for the CJS_α was run over 1,000,000 to 2.000.000 iterations with an additional burn-in phase of 100,000 to 400,000 iterations. We ran two chains for each simulation with two different starting points. For the first one, $\mathbf{x}$ was initialised as the set of observed histories, as if there was no error. In the second one, we arbitrarily added 50 errors randomly. Obviously for running the CJS to simulate the effect of misidentifications, there is no latent set $\mathbf{x}$ but just the observed data $\mathbf{x}$. Knowing that the auto-correlation is very high due to slow movements through the fiber $\mathcal{F}_\mathbf{y}$, we ran the samplers for α and $\mathbf{x}$ (steps 5 and 6 in the algorithm of Section 4.2.3) five times each alternating them in an iteration of the MCMC. This almost amounts to having five times as many iteration but for a much lower computing time than if we had to update every parameters at each of the iteration.

[80]: Valpine et al. (2017), 'Programming With Models'

4.4.2 Results

Running two chains of 1,400,000 iteration on a 3.0GHz Intel processor took a bit more than two hours. We checked the convergence graphically by looking at the ϕ chains and the number of misidentifications chains. We also looked at the $\hat{R}$ and the resulting effective sample size. The CJS always converges perfectly and there are no indication of any problems. The CJS_α also always converge perfectly if we look at the ϕ chains but the chains of misidentifications number show a poor convergence for $\phi \leq 0.7$.

The relative bias on the estimates with the CJS is shown on the Figure 4.1, and the relative bias on the estimates with the CJS_α is shown on the Figure 4.2. The Figure 4.1 shows that the CJS always underestimates the survival. The relative bias does not seem to depend on the true survival

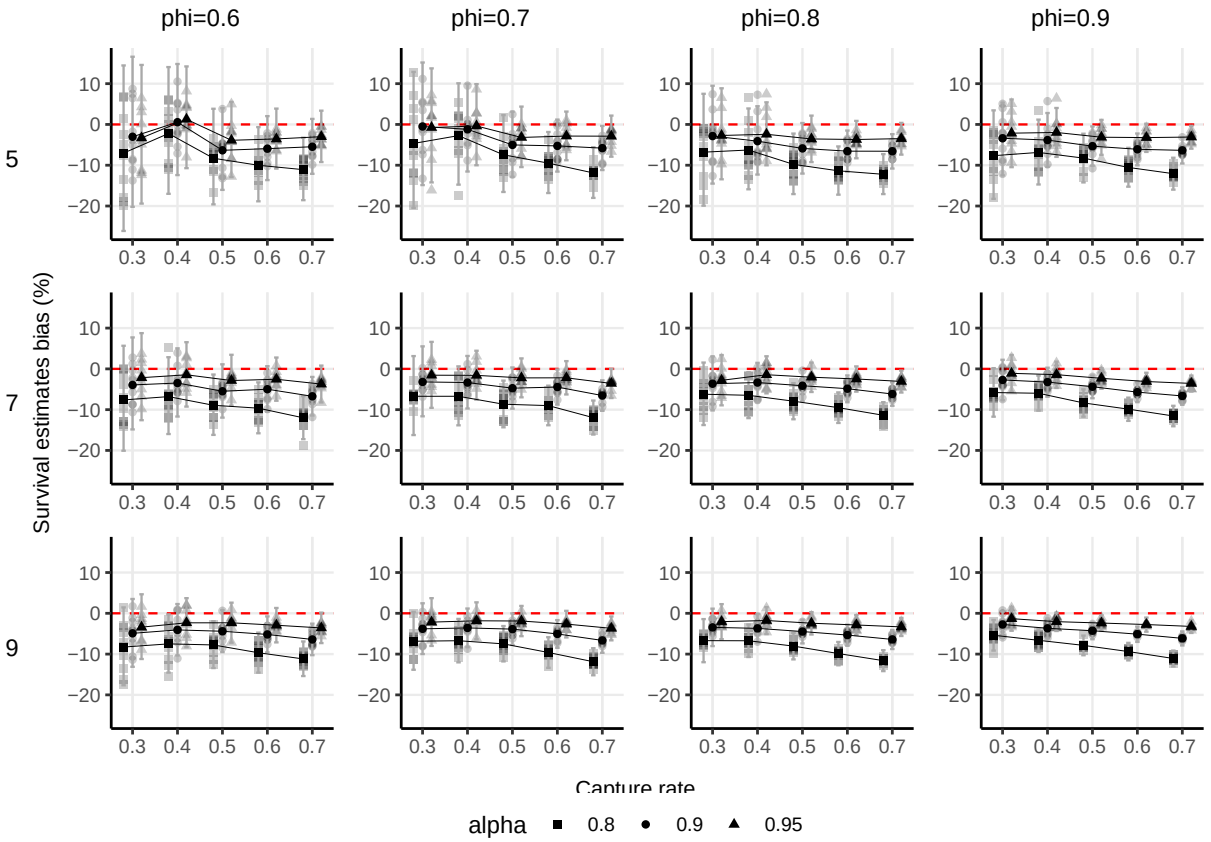

Figure 4.1: CJS relative survival estimate bias (y axis) depending on capture probability (x axis), identification probability (dot shapes), number of capture occasions (rows) and simulated survival (columns). Grey and red symbols show simulation-specific estimate bias of the survival posterior mean, red ones indicating that the N chains had $\hat{R} > 1.1$ or an effective sample size under 50. Black symbols connected by lines and error bars show, respectively, the estimate bias of the mean and the 95% credible intervals of the posterior distribution of survival averaged across simulations.

but rather on the capture rate and the identification rate. For $\alpha = 0.95$, the relative bias is low, a few percents, but when $\alpha = 0.8$, the relative bias reaches 10%. Also, the higher the capture rate is, the higher the bias is. The Figure 4.2 shows that the CJS_α is not biased given that either there are seven or more occasions, either the true survival is of 0.8 or more, either the capture rate is at 0.5 or more. The identification rate does not cause bias in the estimates, but as it get smaller,it increases the size of the 95% confidence interval. When $\phi \leq 0.7$ and $T = 5$ and $p \leq 0.4$, the average bias is still low but the uncertainty is quite large (between 0.5 and 0.75 for $\phi = 0.6$).

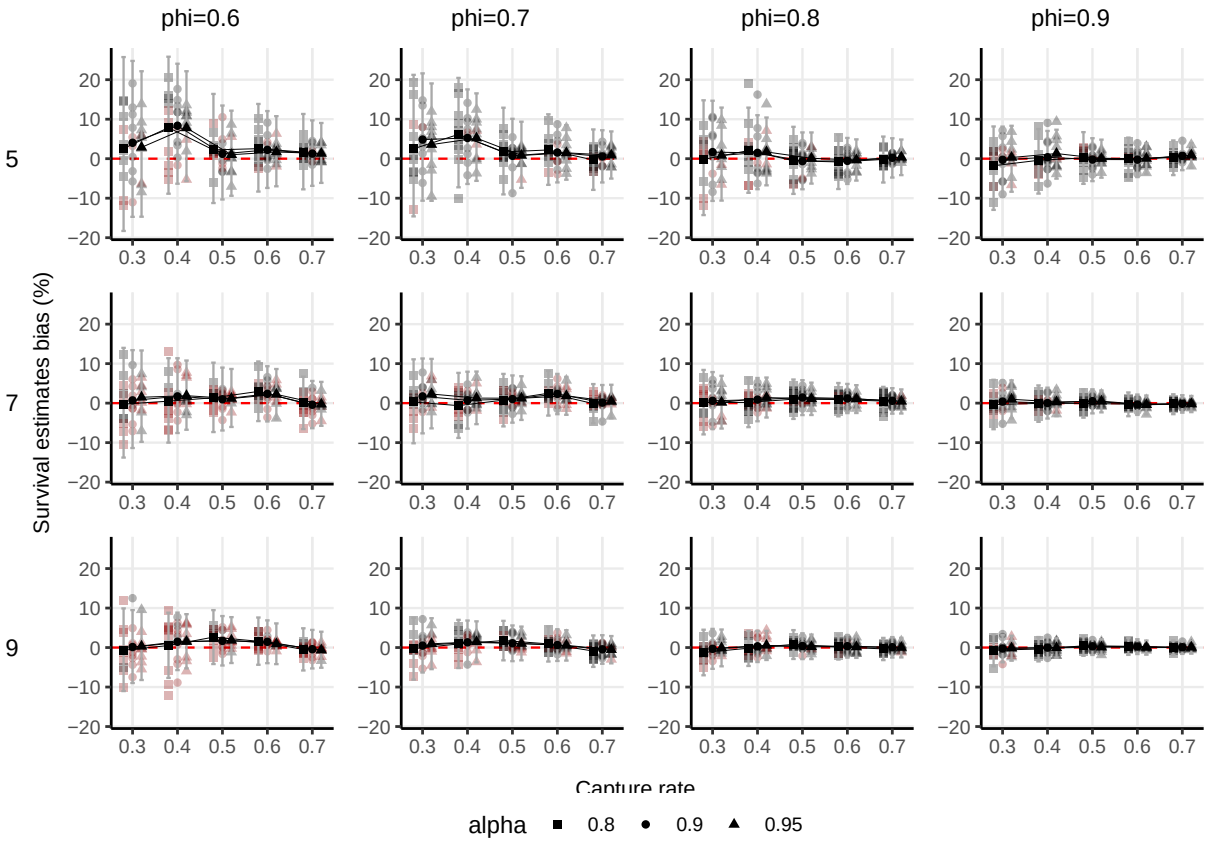

Figure 4.2: CJS_α relative survival estimate bias (y axis) depending on capture probability (x axis), identification probability (dot shapes), number of capture occasions (rows) and simulated survival (columns). Grey and red symbols show simulation-specific estimate bias of the survival posterior mean, red ones indicating that the N chains had $\hat{R} > 1.1$ or an effective sample size under 50. Black symbols connected by lines and error bars show, respectively, the estimate bias of the mean and the 95% credible intervals of the posterior distribution of survival averaged across simulations.

4.5 Discussion

In this chapter I have extended the latent multinomial model to open population capture recapture data. I have developed the LMM in the Cormack-Jolly-Seber framework. As in Chapter 2, I then extended it using a probit approach in order to consider an identification quality covariate. In the same way, I also developed the LMM in the Arnason-Schwarz framework. The development of the model AS_α is of fundamental importance because the aim of many studies is migration or transition rates in the context of an open population, and we have shown in Chapter 1 that transition estimates were biased if the capture or identification rates depend on the states and misidentifications are ignored. I conducted a simulation study of the CJS_α, and showed that the model estimates the survival without bias, even with an identification rate as low as 0.8.

In order to access the full fiber $\mathcal{F}_x = \{\mathbf{x}|\mathbf{y} = \mathbf{A}\mathbf{x}\}$, and to be able to explore the full parameter space with the MCMC, there should be a move that leads to histories containing only misidentifications (*e.g.*: history 020). Such a history is only theoretical and cannot actually happen, because if there is only one capture in a history, then it cannot really be misidentified since identification is more about matching samples together than matching them to individuals. However, in the LMM, such

histories have a positive probability, so the algorithm should be able to propose such histories in the MCMC. In closed population, there are individuals with unseen histories, which the algorithm can use to propose histories with only misidentifications (if a misidentification is added to an unseen history). In the CJS framework, the likelihood is conditioned on the first capture so the unseen histories are not considered and the possibility of moving toward an all-misidentification history must be allowed as a move of the MCMC.
I have implemented the MCMC with this in mind, but some tests comparing estimates with and without such a possibility of move seem to indicate that it may not be absolutely necessary.

Parameters estimation for open population models is much slower than for closed population models, and the MCMC requires many more iterations for low capture probabilities. In addition, with low survival, the computation time increased substantially due to the increased autocorrelation of the MCMC and the need for more iterations.

Both the AS_α model and the CJS_{α_n} model have not been tested with repeated simulations on a diverse panel of scenarios.
The AS_α model was only developed as a necessary step for the next chapter. Therefore we did not take the time to run such simulations. However, it would be interesting to see how the survival and the transitions are estimated by the model in presence of misidentifications.
The CJS_{α_n} model was developed to be applied to a real data set. However, this has not been possible. Therefore, testing the model and comparing it with the CJS_α remains to be done. As in Chapter 2, we expect that the CJS_{α_n} would lead to more precise results than the CJS_α, but the evaluation of this extension remains to be done.

I developed some models in this chapter but other potentially useful models still need to be developed. The Table 4.1 summarises which models have been developed and which models have been omitted .

	Single observation		Multiple observation	
	no covariate	covariate	no covariate	covariate
Single state	Yes	Yes	No	No
Multistate	Yes	No	No	No

Table 4.1: Current state of development for open population models. "Yes" indicates that the model is developed and implemented and "No" indicates that the model is not developed.

The interest of a multi-state open population model using an identification quality covariate lies in improving the estimates of the model. This model is built by replacing the capture process of the CJS_{α_n} with the capture process of the AS model, and adapting the algorithm to sample the state in which we add or remove an error.
All the models for studies with multiple observations on an occasion (see Chapter 3) are an extension that would allow many studies to use the LMM, namely, most studies that collect faeces to study individual survival.

For the closed population studies, and for any study where misidentifications could have occurred, the data should be analysed with a model such as the CJS_α to confirm whether or not misidentifications occurred. If the estimated α is around 1, this validates that no misidentifications have occurred and the data can be analysed using a classical model. If the estimated α is less than 1 and the estimates of the classical model and

the one modelling misidentifications are significantly different then the CJS_α (or the CJS_{α_n}) is the model to use.

Identifiability issues can arise especially when the capture or the survival or the identification rate is low. Testing the sensitivity to the prior of α can help to detect lack of identifiability.

Since, in our simulations, the CSJ_α estimates the survival without bias even when it is as low as 0.6, with a low capture rate of 0.3 and a low identification probability of 0.8, many more samples could be kept by allowing misidentifications. This could open the door to new study designs where a large percentage of the samples would not pass the quality threshold for eliminating misidentifications. In the next chapter I present an example of a study where it would be expected that deleting low quality samples would lead to keep almost no sample. The chapter shows how the LMM could be used to estimate the survival of mosquito larvae with CMR.

Chapter five

Study of mosquito larvae survival rate using capture-recapture

5

5.1 Introduction

5.1.1 Project aim

This PhD project was part of a larger research project that advocates the study of the adaptation of mosquito vectors to environmental modifications, in particular, how global change might impact their fitness and life-history traits and influence their vectorial capacity. This will allow more accurate prediction of the epidemiological consequences of niche expansion and the spread of mosquito-borne pathogens. Such predictions are essential in order to adapt disease control programmes and avoid the emergence of vector-borne diseases.

This chapter presents a simulation study of mosquito larvae survival rate using capture-recapture in the context of misidentification. Although the study is based on simulations, these have been made as realistic as possible using external biological information on mosquitoes and knowledge from the field. Thus, this chapter may be viewed as a good illustration of the opportunities offered by the latent multinomial model (LMM).

The chapter starts by introducing the context of the research project as well as the current state of knowledge about the demographic parameters of mosquitoes. It then describes two possible protocols for studying mosquito larvae survival. In these protocols, individual capture histories are obtained by 'marking' mosquito larvae with genetic fingerprints using eDNA. The specific extensions of the AS_α model (see Chapter 4) used to analyse the data in the two protocols are defined. Based on the literature, realistic simulations were created corresponding to the two protocols and these were then compared using the models developed.

5.1.2 Mosquitoes, vectors of disease

Mosquitoes transmit some of the most acute infectious diseases impacting humans. Despite recent progress, malaria, Bancroftian filariasis, and viruses such as dengue, chikungunya, Zika and yellow fever continue to pose major challenges to public health. These challenges are further compounded by global changes in both the environment and society that are promoting the emergence and resurgence of these diseases worldwide. Among these global environmental changes, the demographic and spatial growth of densely populated urban areas is causing a shift in the way mosquito-borne diseases spread. Previously, these diseases mainly affected rural areas with lower host populations. However, with the expansion of urban centres and the resulting increase in population density, new risks for the transmission of these pathogens are arising. In fact, the least developed countries are those experiencing the highest urbanisation rates, often in the range of 2–6% per year. This trend is

especially evident in tropical Africa, where major disease vectors such as the Anopheles gambiae complex, responsible for transmitting malaria and filariasis, are adapting to the haphazard growth of urban centres. These mosquitoes are now thriving in anthropogenically polluted water, which were previously unsuitable as larval habitats [99].

Increased resilience to environmental stressors, such as pollutants, appears to be a key factor driving this phenomenon [100]. Another illustration of recent niche expansion involves certain freshwater species such as *Aedes albopictus*, *Ae. aegypti*, and *An. coluzzii*, which are adapting to brackish waters within certain regions of their distribution [101, 102]. This adaptation carries the potential of heightened transmission risk, particularly as rising global temperatures contribute to sea-level rise, resulting in the expansion of saline and brackish water along coastlines. Similarly, the globalisation of transport and trade has facilitated the widespread invasion of new ecological niches by *Ae. albopictus*. This expansion has given rise to the development of traits such as diapause and increased inter-specific competitive ability. Yet there is limited understanding of whether these adaptations linked to niche expansion will amplify or diminish the intensity of disease transmission. Moreover, there is uncertainty whether these adaptations will drive the dissemination of mosquito-borne pathogens into even more distant areas. For instance, the exposure of mosquitoes to foreign substances and pollutants in urban hubs might encourage the emergence of resistance to insecticides, potentially undermining our ability to manage these vectors. On the other hand, costs associated with the evolution of adaptive traits may affect life-history traits and fitness [103]. These aspects, in turn, may influence the vector's capacity to transmit disease and the force of transmission. Studying how current global transformations influence life-history traits connected to fitness will allow us to more accurately anticipate the epidemiological consequences of anthropogenic mosquito vector niche expansion and the proliferation of mosquito-borne diseases.

[99]: Kamdem et al. (2012), 'Anthropogenic habitat disturbance and ecological divergence between incipient species of the malaria mosquito Anopheles gambiae'

[100]: Tene Fossog et al. (2013), 'Physiological correlates of ecological divergence along an urbanization gradient'

[101]: Ramasamy et al. (2012), 'Global climate change and its potential impact on disease transmission by salinity-tolerant mosquito vectors in coastal zones'

[102]: Tene Fossog et al. (2015), 'Habitat segregation and ecological character displacement in cryptic African malaria mosquitoes'

[103]: Ramasamy et al. (2014), 'Biological differences between brackish and fresh water-derived Aedes aegypti from two locations in the Jaffna peninsula of Sri Lanka and the implications for arboviral disease transmission'

5.1.3 Life cycle and environmental conditions

The life cycle of mosquitoes is shown in Figure 5.1. It is characterised by four distinct stages: egg, larva, pupa and adult. Mosquitoes undergo a process known as complete metamorphosis, which involves a series of morphological and physiological changes. Each stage plays a vital role in mosquito development and survival, with the immature stages being particularly susceptible to environmental factors and control interventions.

The life cycle of a mosquito begins when a female mosquito lays her eggs in or near a water source. These habitats can vary from natural breeding sites such as stagnant ponds, rice fields and river edges to artificial containers such as rainwater-filled tyres or water storage containers. The female *An. gambiae* typically lays her eggs in clusters of highly variable size, depending on her size [104], parity [105], access to blood or oviposition site [105], plasmodium infection [106], parasites [107] or sugar sources [108]. The number of eggs ranges between a few dozen to more than 200. The time required for mosquito eggs to hatch varies based on environmental conditions, primarily temperature and humidity. In optimal conditions, eggs can hatch within 24 to 48 hours.

[104]: Lyimo et al. (1993), 'Effects of adult body size on fecundity and the pre-gravid rate of Anopheles gambiae females in Tanzania'

[105]: Dieter et al. (2012), 'The effects of oviposition-site deprivation on Anopheles gambiae reproduction'

[105]: Dieter et al. (2012), 'The effects of oviposition-site deprivation on Anopheles gambiae reproduction'

[106]: Hogg et al. (1997), 'The effects of natural Plasmodium falciparum infection on the fecundity and mortality of Anopheles gambiae s. l. in north east Tanzania'

[107]: Nnakumusana (1986), 'The effect of Coelomomyces indicus on the fecundity and longevity of Anopheles Gambiae, Culex fatigans and Aedes aegypti exposed to infection at each larval instar'

[108]: Manda et al. (2007), 'Effect of discriminative plant-sugar feeding on the survival and fecundity of Anopheles gambiae'

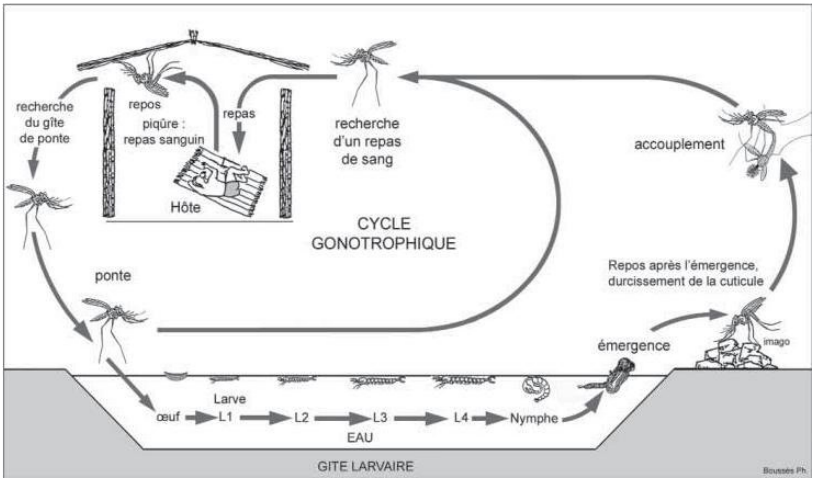

Figure 5.1: Biological cycle of the Anophele (Boussès Ph.).

Upon hatching, the mosquito larvae emerge from the eggs and enter the larval stage. The larval phase of *An. gambiae* is a multi-stage developmental period marked by a series of moults, resulting in four distinct larval instars. *An. gambiae* larvae progress through four larval instars, with each stage characterised by specific morphological and physiological changes. First-instar larvae emerge upon hatching from the eggs, displaying a relatively simple structure. As they progress through successive instars, their bodies become more segmented and elongated. Notably, the mouthparts become more developed, allowing them to feed more efficiently on microorganisms and organic matter in their aquatic habitat. The duration of each instar varies, with the fourth instar being the longest, lasting up to several days. Larvae primarily feed on microorganisms, detritus and organic matter present in the water, contributing to nutrient cycling in aquatic ecosystems. They undergo multiple moults, shedding their exoskeletons as they grow.

As the larval stage nears completion, larvae transition into pupae. Pupae are more mobile than larvae, but do not engage in feeding activity. Instead, their primary focus is the completion of their transformation into adult mosquitoes. The pupal stage is a critical period for the development of adult structures, including wings, legs and reproductive organs. The pupal stage typically lasts about 2 days, after which the adult *An. gambiae* emerges onto the water's surface. At this stage, the mosquito is fragile, and it takes some time for its exoskeleton to harden and its wings to fully expand before it becomes a functional, flight-capable adult.

Once the female flies away, she mates and then requires a blood meal to complete the reproductive cycle. Once the eggs are developed, the female looks for a larval site and lays the eggs. Then, without needing to be fecundated again, the female starts another cycle, looking for a blood meal. The female's lifespan is around 3 to 4 weeks in sub-Saharan Africa.

5.1.4 Development time of the immature stages

Table 5.1 provides several estimates of the average time spent in each larval stage for *An. gambiae*. The overall duration from hatching to adult emergence varies by up to a factor of two. A large part of these variations

is likely due to variations in environmental temperature, which is often not measured or reported, though it is known to have a significant effect [109]. While the estimates differ slightly, they are generally consistent in approximating the duration of the complete immature cycle from hatching to adulthood. The complete cycle takes between one and two weeks. Service (1971, 1973 [110, 111]) observed a longer development time for the fourth instar. This is consistent with the fact that a fourth instar larva must prepare for the pupa stage and does not feed during this time. However, Bayoh et al. (2003) [109] observed this phenomenon only for temperatures below 25°C. At temperatures around 25°C, a similar development time was observed for all instars, and for temperatures above 25°C they reported a shorter development time for the fourth instar than for the other instars.

[109]: Bayoh et al. (2003), 'Effect of temperature on the development of the aquatic stages of Anopheles gambiae sensu stricto (Diptera'

[110]: Service (1971), 'Studies on sampling larval populations of the Anopheles gambiae complex'

[111]: Service (1973), 'Mortalities of the larvae of the Anopheles gambiae Giles complex and detection of predators by the precipitin test'

	°C	Instar 1	Instar 2	Instar 3	Instar 4	Pupa	Total
[112]	24-26	-	-	-	5-8	1	6-9
[113]	?	-	-	-	9.2	1.2	10.4
[110]	?	1.5	3	2	4	2	12.5
[111]	?	1.42	2.88	1.93	3.75	1.79	11.77
[109]	30	1.90	2.09	2.27	1.42	1.11	8.79

Table 5.1: Instar development time per state (in days). If only state 4 is given, it is for the complete larval cycle.

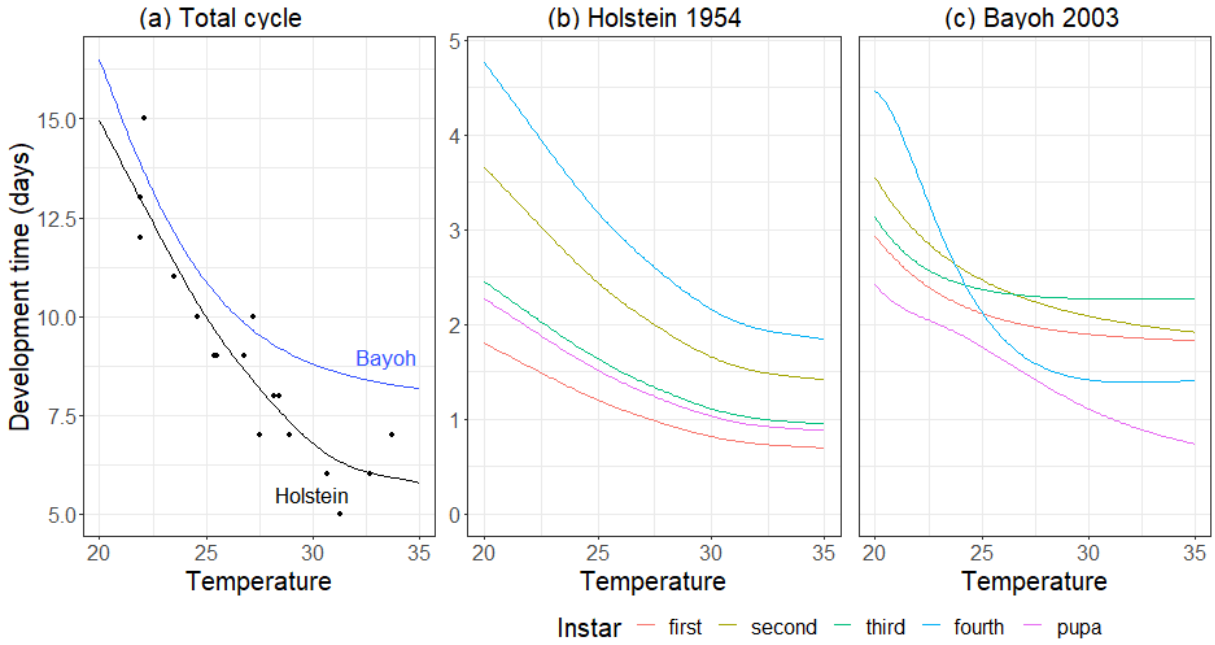

Figure 5.2: Average development time per instar depending on temperature.

In addition, Holstein (1954) [113] measured the cycle time from egg to adult as a function of temperature. By subtracting from these total times one day for the egg duration status, we obtain the development time of the immature post-hatching stages as a function of temperature. In Figure 5.2 (a), the total immature cycle time depending on the temperature as measured by Holstein (1954) [113] and by Bayoh (2003) [109] is shown. If we consider the development time reported by Service (1973) [111] to be representative of the ratio of time spent in each stage, these ratios can then be used to calculate the average time per stage and per temperature, as estimated by Holstein (954) [113]. Figure 5.2 (a) and (b) show the average development time for each instar given the temperature based

[113]: Holstein (1954), 'Biology of Anopheles gambiae. Research in French West Africa.'

[109]: Bayoh et al. (2003), 'Effect of temperature on the development of the aquatic stages of Anopheles gambiae sensu stricto (Diptera'

on Holstein's or Bayoh's data.

5.1.5 Survival of the immature stages

The survival of mosquito larvae has long been a topic of study. Bates (1941) [114] used the duration of the larval instar to estimate the theoretical proportion of each larval instar in a population if there was no mortality. He compared it with observations in several sites. The method has since been improved and life tables have been constructed. First, larvae from the different stages are collected. The numbers of each instar are then divided by the corresponding instar duration. Then these values are plotted against the age (in days) of the larvae and pupa. A survivorship curve (see Figure 5.3) is then fitted. From this curve, one can obtain the number of larvae surviving to each age in days. These numbers of surviving individuals are then used to create life tables, leading to survival probability for each age.

[114]: Bates (1941), 'Field Studies of the Anopheline Mosquitoes of Albania.'

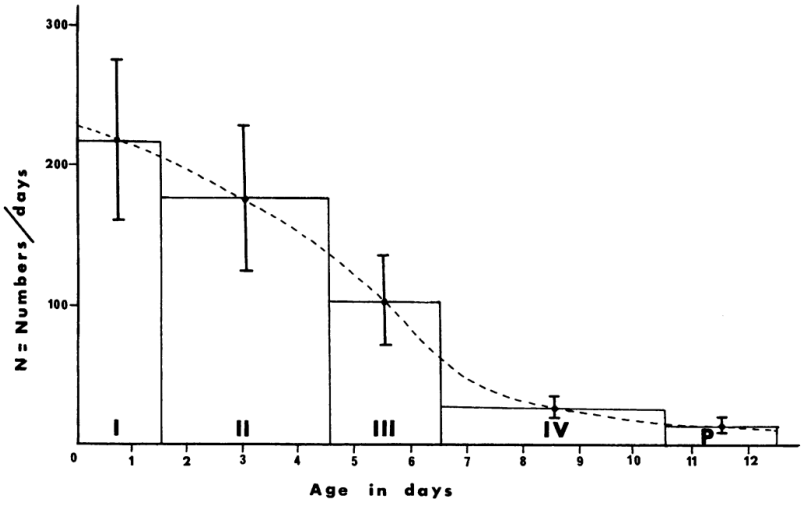

Figure 5.3: Example of survivorship curve (from Service 1971 [110])

This method has been used in several studies to estimate the survival of immature stages of *An. gambiae* in different habitats, including borrow pits and marshes (in Kenya [110]), Rabour and Chiga (in Kenya [111]), sprayed rice fields, unsprayed rice fields, and pools and ponds (in Kenya [115]). The survival probability estimates from the previously cited studies are given in Table 5.2. The survival probability is estimated daily as a function of larval age. Service (1971) [110] estimates survival over a longer lifetime, so an additional survival probability is given. The overall mortality over the complete immature cycle is between 0.9 and 0.96 days. I aggregated the values of Table 5.2 in Section 5.3.1 to simulate survival.

[110]: Service (1971), 'Studies on sampling larval populations of the Anopheles gambiae complex'

[111]: Service (1973), 'Mortalities of the larvae of the Anopheles gambiae Giles complex and detection of predators by the precipitin test'

[115]: Service (1977), 'Mortalities of the immature stages of species B of the Anopheles gambiae complex in Kenya'

Table 5.2: Survival probability to age $t + 1$.

Age (days)	0	1	2	3	4	5	6	7	8	9	10	11
Marshes	0.93	0.92	0.90	0.85	0.80	0.66	0.59	0.62	0.72	0.80	0.77	0.84
Borrow pits	0.94	0.92	0.90	0.87	0.83	0.77	0.69	0.63	0.60	0.61	0.61	0.65
Rabour	0.96	0.92	0.89	0.85	0.77	0.71	0.65	0.55	0.58	0.61	0.63	
Chiga	0.94	0.88	0.84	0.81	0.78	0.74	0.70	0.59	0.63	0.63	0.65	
Rice field	0.81	0.79	0.83	0.84	0.83	0.81	0.80	0.76	0.75	0.68	0.59	
Pools & ponds	0.95	0.94	0.90	0.87	0.81	0.80	0.76	0.75	0.75	0.68	0.52	
Sprayed rice field	0.95	0.94	0.91	0.90	0.88	0.73	0.80	0.78	0.77	0.80	0.76	

5.2 Capture-recapture on mosquito larvae

5.2.1 Experimental design

This section presents two capture-recapture (CR) protocols for mosquito larvae. An idea for a protocol was proposed by the team working on the research project (at the MIVEGEC research institute in France). From this, I defined two protocols, which are detailed below, indicating their points in common and how they differ.

On a capture occasion t, larvae are captured in a larval site and kept for 24 hours in contaminant-DNA-free water. They are released in their larval site immediately after the next capture occasion. By releasing a larva at $t + 1$, it cannot be captured twice consecutively. The impacts of the manipulation on the larvae are thus limited. Thus, we can assume that captures do not affect the survival or transition rates between states of released individuals. An additional benefit of releasing an individual after the next capture is that the state of the individual is known for two consecutive occasions.

DNA is extracted from the water in which a larva has been kept and is genotyped. The genotypes are matched to each other in order to construct capture histories. However, obtaining DNA from a small sample of water that contained a larva for one day is challenging. Typically, little DNA is available, so the genotype is very likely to be incomplete. Hence, misidentification can occur. In this case, the probability of identification [48] is kept low enough to avoid confusing two individuals. Any sample that could result in confusion because the identification probability is too high is discarded. So the only possible source of error is the creation of a new individual (a ghost). I also considered that there is no confusion between states, as each instar is sufficiently different.

[48]: Taberlet et al. (1999), 'Non-invasive genetic sampling and individual identification'

The protocol involves manipulating fragile individuals. Larvae are sampled from the larval site and rinsed in DNA-free water to minimise contamination before being put in individual tubes. This leads to the likelihood that some individuals will be killed accidentally. This mortality is added to natural mortality occurring in the tubes for different reasons.

The first and second instars are smaller and more delicate than the other instars. Consequently, there is a higher risk of accidentally causing harm or killing them when trying to capture and handle them. Moreover, due to their smaller size, they are likely to release less DNA. Abstaining from capturing the first and second instars reduces the population of interest

(to third and fourth instars and pupae). Thus, for equivalent capture effort (in the number of captured individuals), the capture rate will be higher (because less individuals are available). Considering all these points, I designed two different protocols.

For the first protocol, the pupae caught are killed and identified. The reason is that pupae take only around two days to develop to adulthood. Thus, if one is released, the probability of recapturing it is very low. Furthermore, when it becomes an imago, the probability of seeing it again is zero. Additionally, using the individuals DNA to identify pupae ensures sufficient DNA to avoid misidentification at this stage. The drawback of this is that there will be no information in the data-set about the survival probability of pupae. Fourth instar developing into pupae during the 24h of a capture are released. This is for reasons of simplicity as it makes encoding simpler. (Captures were always over two occasions, even for pupa, as the second occasion indicates the end of the history.)

To solve the problem of pupa survival, I designed a second protocol. In this protocol, larval sites are covered with a net to allow the capture of all emerging adults. Compared to the first protocol, captured pupae are released and potentially recaptured as pupae, although this is unlikely. However, they are recaptured as adults if they successfully develop. The net also protects the larval site from large predators of the larvae, which helps to compare the survival in different larval sites in relation to conditions of interest such as salinity and water pollution. However, the net also prevents new eggs from being laid in the site.

To code individual histories, unseen individuals are denoted as is usual with 0s. Then we denote the capture of individuals at different stages starting from 1 for third instar, to 4 for adults. In the latent error history, individual misidentification is noted with numbers above the one used for the last available stage. Thus, in the first protocol, as no adults are captured, misidentification is noted 4 and 5 for the third and fourth instar respectively. In the second protocol, misidentification is noted 5, 6 and 7 for the third instar, fourth instar and pupae respectively. Right censoring occurs either because an individual is killed accidentally when handling it or voluntarily because it is a pupa (in the first protocol) or an adult (in the second protocol). The right censoring of a history is denoted with a 6 in the first protocol and with an 8 in the second protocol.

Capture notations

Protocol 1

- 1, 2, 3: Third instar, fourth instar, pupa
- 4, 5: Misidentification of third and fourth instar
- 6: History end (right censoring)

Protocol 2

- 1, 2, 3, 4: Third instar, fourth instar, pupa, adult
- 5, 6, 7: Misidentification of third instar, fourth instar and pupa
- 8: History end (right censoring)

For the first protocol, a history is shown below for an individual first captured in the third instar, released in the third instar after the next occasion, then later recaptured in the fourth instar and released as a pupa, then recaptured as a pupa and killed:

$$1 \quad 1 \quad 0 \quad 2 \quad 3 \quad 3 \quad 6$$

For the second protocol, the same history is considered for the five first occasions. At occasion 6, the pupa is not recaptured, but it is recaptured as an adult at occasion 7:

$$1 \quad 1 \quad 0 \quad 2 \quad 3 \quad 0 \quad 4 \quad 8$$

.

5.2.2 Modelling the capture data of the larvae

To model the data from these experiments, I consider a special case of the AS_α model from Section 4.3.2. The models for both protocols are very similar. However, the $\boldsymbol{\psi}$ parameter matrix takes into account transitions up to pupa in the first protocol and up to adult in the second. In the second protocol, the additional capture probability for the adult state is set to 1.

The structure of the transition matrix for both protocols is:

$$\boldsymbol{\phi}^{(P1)} = \begin{pmatrix} \psi_{1,1} & \psi_{1,2} & 0 \\ 0 & \psi_{2,2} & \psi_{2,3} \\ 0 & 0 & 1 \end{pmatrix} \quad \boldsymbol{\phi}^{(P2)} = \begin{pmatrix} \psi_{1,1} & \psi_{1,2} & 0 & 0 \\ 0 & \psi_{2,2} & \psi_{2,3} & 0 \\ 0 & 0 & \psi_{3,3} & \psi_{3,4} \end{pmatrix}$$

In a small larval site, the population can vary significantly in a day. Thus, it is considered that the capture probability is time dependent. Second, as it is possible to capture larvae independently of their state, it is considered that the capture probability is not state dependent. Finally, the identification probability is considered constant.

The formulation of the likelihood is the same as in Section 4.3.2:

$$[\mathbf{y}, \mathbf{x}, \mathbf{z}|(N_{s,t}), \boldsymbol{\phi}, \boldsymbol{\psi}, \mathbf{p}, \alpha] = I(\mathbf{y} = \mathbf{A}\mathbf{x})\,[\mathbf{x}|\mathbf{z}, \alpha]\,[\mathbf{z}|(N_{s,t}), \phi, \boldsymbol{\psi}, \mathbf{p}] \quad (5.1)$$

Notation reminder

- $(N_{s,t})$: number of individuals first captured at t in state s
- $\mathbf{y}$: frequencies of observed histories
- $\mathbf{x}$: frequencies of latent error histories
- $\mathbf{z}$: frequencies of latent capture histories
- $\boldsymbol{\phi}$: survival probabilities,
- $\boldsymbol{\psi}$: transition probabilities,
- $\mathbf{p}$: capture probabilities,
- α: identification probability

- $[\mathbf{z}|\boldsymbol{\phi}, \boldsymbol{\psi}, \mathbf{p}]$

To model the capture process, we need to take into account that a captured individual is kept for 24h and is released just after the next capture occasion, having a capture probability of 1 at $t+1$ conditional on being captured at t. To implement this constraint, we first modify Equation 4.18. Let $\mathcal{S}$ be the set of states a larvae may be observed in. Let $O_{(c,d)}(r_1, r_2, s_1, s_2)$ be the probability that an animal in state $r_1 \in \mathcal{S}$ and $r_2 \in \mathcal{S}$ at occasions c and $c+1$ remains unobserved until it is subsequently resighted in states $s_1 \in \mathcal{S}$ and $s_2 \in \mathcal{S}$ at times $d+1$ and $d+2$, for $C \leq c \leq d \leq D-2$

$$O_{(c,d)}(r_1, r_2, s_1, s_2) = \psi_{r_1,r_2}^{I(c=C)} p_{d+1,s_1} \psi_{s_1,s_2} Q_{(c+1,d)}(r_2, s_1), \quad (5.2)$$

where $Q_{(c,d)}(r_2, s_1)$ denotes the probability that an animal changes from state r_2 at time c to state s_1 at time $d+1$, and is unobserved between these times. It is the same as in Equation 4.19. The likelihood of the sighting in history ξ_k, $L_{k,1}$ takes the same form as Equation 4.20:

$$L_{k,1} = \prod_{(c,d)\in\mathcal{C}\mathcal{D}} O_{(c,d)}(r, s) \quad . \quad (5.3)$$

The probability that an individual last released at D in state r is not resighted again, $\chi_{r,D}$, is the same as in Equation 4.21. It need not be changed since it only concerns when larvae were unobserved.

The likelihood of history ξ_k is $\pi_k = L_{k,1} \cdot \chi_{r,D}$ and the likelihood for the capture process is

$$[\mathbf{z}|\phi, \mathbf{p}] = \frac{\prod_{s\in\mathcal{S}} \prod_{t=1}^{T} N_{s,t}!}{\prod_k z_k!} \prod_k \pi_k^{z_k} \quad . \tag{5.4}$$

- $[\boldsymbol{\alpha}|\boldsymbol{\theta}]$

Next we change the likelihood of identification. Since a capture always consists of two observations, an identification always concerns two capture occasions. The likelihood of the identification process is computed with Equation 1.5, rewriting $A_j = \alpha^a(1-\alpha)^b$, with a the number of times an individual with latent history v_j is correctly identified and b the number of times they are misidentified. The identification likelihood is:

$$[\mathbf{x}|\mathbf{z}, \alpha] = I(\mathbf{z} = \mathbf{B}\mathbf{x}) \frac{\prod_k z_k!}{\prod_j x_j!} \prod_j \left[\prod_{t=1}^{T} \alpha^{J_t \cdot I(v_{j,t}\in\mathcal{S})} (1-\alpha)^{J_t \cdot I(v_{j,t}>S)} \right]^{x_j} \tag{5.5}$$

where J_t is the indicator that the individual was captured at occasion t and will be released after occasion $t+1$. $\mathcal{S}$ is the set of states that can be misidentified. For the first protocol, $\mathcal{S} = \{1, 2\}$, as only the third and fourth instar can be misidentified. For the second protocol, $\mathcal{S} = \{1, 2, 3\}$, as only the adults cannot be misidentified.

5.2.3 Estimating parameters of the larvae models

The algorithm to estimate the parameters of the model is almost the same as in Section 4.3.3. However, the way of indicating a different vector of latent counts $\mathbf{x}'$ is adapted to the specific capture process.

The MCMC is constructed this way:

1. Let $\beta(a_0^t, b_0^t)$ denote the beta prior on p_t, $\beta(a_0^\alpha, b_0^\alpha)$ denote the beta prior on α and $\beta(a_0^\phi, b_0^\phi)$ the beta prior on ϕ.

2. Initialise all parameters as well as a set of latent histories satisfying **y**=**Ax**. Such a set can be obtained by assuming that no mistakes were made. The latent frequency of the histories containing 2s are 0, and all the others match the observed frequency one-to-one. In order to run several chains with different initialisations, one can take the previous initialisation of **x** and follow the later steps of (6) by only adding misidentification to the set and always accepting the proposed values without going though the Metropolis-Hasting acceptance.

3. Sample the capture rate with Gibbs sampling. The likelihood being multinomial, it follows that the beta priors lead to full conditional beta posterior distribution:

$$p_t|\mathbf{z}, \boldsymbol{\phi}, \boldsymbol{\psi} \sim \beta(a_0^t + a^t, b_0^t + b^t)$$

where a^t is the number of individuals seen at least once before t that were available and captured at t, and b^t is the number of individuals seen at least once before t that were alive but unseen at t [1]. The individuals that were captured at $t-1$ are not available for capture at t, so they do not contribute to the posterior of p_t.

1: To compute b^t, the latent state 'dead or alive' of each individual must be sampled. The probability that an individual last released before t and never seen again is still alive at t can be computed with the forward-backward algorithm.

4. Sample the survival rate with Gibbs sampling. Again, it has a full conditional beta posterior distribution.

$$\phi|\mathbf{z}, \mathbf{p}, \boldsymbol{\psi} \sim \beta(a_0^\phi + a^\phi, b_0^\phi + b^\phi)$$

where a^ϕ is the total number of times an individual survived from one occasion to the next after its first sighting, and b^ϕ is the total number of individuals seen once or more that have died before the last occasion. After the last release of an individual, there is no way of knowing if it survived or died, but the transition towards one state or another can be sampled.[2] The observation of survival before the release of captured individuals does not contribute to this survival probability.

2: This is also done using the forward and backward algorithms.

5. Sample the transition rates with Gibbs sampling. They have a full conditional beta posterior distribution.

$$\psi_{s,\cdot}|\mathbf{z}, \mathbf{p}, \phi \sim Dir(\mathbf{a_0}^{\psi_{s,\cdot}} + \mathbf{a}^{\psi_{s,\cdot}})$$

where $\mathbf{a}^{\psi_{s,\cdot}}$ is the number of of times an individual transitioned from state s to the others. Just like for the initial states, we can't know what transition occurred for an unseen individual but it can also be sampled. [3].

3: This is also done using the forward and backward algorithms.

6. Sample the identification rate with Gibbs sampling. Similar to the capture rate, it has a full conditional beta posterior distribution.

$$\alpha|\mathbf{x} \sim \beta(a_0^\alpha + a^\alpha, b_0^\alpha + b^\alpha)$$

where a^α is the total number of correct identifications and b^α the total number of misidentifications.

7. Sample $\mathbf{x}$. Sampling $\mathbf{x}$ is done the same way as in previous models. Randomly add or remove an error from the set of latent histories. The difference is that an individual is always seen twice in a row, so misidentification can only come from individuals where at least two consecutive 0s are observed. To add an error, sample a history that may have generated a ghost (i.e. a history containing two consecutive 0s), and 'merge' it with a potential ghost (i.e. replace the first 0 by $s+S$, [4] the second 0 by r and remove the ghost history). To remove an error, sample a history containing a value used for misidentification, replace it by a 0, as well as the following number in its history, and add a history with only two consecutive captures (coded s and r) at that time. More formally, follow the steps:

4: s is the state the ghost was observed when captured, and r the state it was observed before released.

   a) Define:
      - $\boldsymbol{\nu}^{(1,srt)}$ the history with only two captures at time t and $t+1$ in states s and r (potential ghost)
      - $\chi_{0,srt}(\mathbf{x}) = \{\boldsymbol{\nu}|\nu_t = 0, \nu_{t+1} = 0, x_\nu > 0\}$ the set of histories having *potentially* generated a ghost in state s at time t

and r at time $t+1$, for the given $\mathbf{x}$

- $\chi_{2,srt}(\mathbf{x}) = \{\boldsymbol{\nu} | \nu_t = s + S, \nu_{t+1} = r, x_\nu > 0\}$ the set of histories *containing* a ghost at time t in consecutive states s and r, for the given $\mathbf{x}$.

b) With a probability of 0.5, go to (i), otherwise go to (ii).

i. Add a misidentification (i.e. a ghost) to the latent set.
   - Sample a state s and a state r that can follow the state s.
   - Sample $t \in \{t | \exists \boldsymbol{\nu} : \nu_{(t,t+1)} = 0, P(\nu_{(t,t+1)} = s, r) > 0, x_{\nu^{(1,srt)}} > 0\}$.
   - Sample $\boldsymbol{\nu}^{(0)} \in \chi_{0,srt}(x)$.
   - Set $\boldsymbol{\nu}^{(2)}$ as $\boldsymbol{\nu}^{(0)}$, $\boldsymbol{\nu}_t^{(2)} = s$, $\nu_{t+1}^{(2)} = r$.
   - Define the move $b_{\nu^{(0)},\nu^{(1,srt)},\nu^{(2)}} = (-1, -1, +1)$, and $b_\nu = 0$ for all other latent histories.

ii. Remove a misidentification from the latent set.
   - Sample a state s and a state r that can follow the state s.
   - Sample $t \in \{t | \exists \boldsymbol{\nu} : \nu_t = s + S, nu_{t+1} = r, x_{\nu^{(1,srt)}} > 0\}$.
   - Sample $\boldsymbol{\nu}^{(2)} \in \chi_{2,srt}(x)$.
   - Set $\boldsymbol{\nu}^{(0)}$ as $\boldsymbol{\nu}^{(2)}$, $\boldsymbol{\nu}_t^{(0)} = 0$, $\boldsymbol{\nu}_{t+1}^{(0)} = 0$.
   - Define the move $b_{\boldsymbol{\nu}^{(0)},\boldsymbol{\nu}^{(1,srt)},\boldsymbol{\nu}^{(2)}} = (+1, +1, -1)$, and $b_\nu = 0$ for all other latent histories.

c) Define $\mathbf{x}' = \mathbf{x}^{(k-1)} + b$.

d) Calculate $\mathbf{z}' = \mathbf{B}\mathbf{x}'$.

e) Calculate the proposal numbers of first capture per occasion and state $(N_{st})'$.

f) Set $\mathbf{x}^{(k)} = \mathbf{x}'$ with probability $min(1, r)$. Otherwise set $\mathbf{x}^{(k)} = \mathbf{x}^{(k-1)}$.

$$r = min\left(1, \frac{[\mathbf{y}, \mathbf{x}', \mathbf{z}' | (N_{st})', \boldsymbol{\psi}\boldsymbol{\phi}', p, \alpha]}{[\mathbf{y}, \mathbf{x}^{(k-1)}, \mathbf{z}^{(k-1)} | (N_{st}), \boldsymbol{\psi}, \boldsymbol{\phi}, p, \alpha]} \frac{q(\mathbf{x}^{(k-1)} | \mathbf{x}')}{q(\mathbf{x}' | \mathbf{x}^{(k-1)})}\right) \quad (5.6)$$

Proposal density

The proposal density is calculated by multiplying the probability of each sampling step used for defining the move. They are successive: the probability of adding (or removing) an error, the probability of choosing the states s and r, the probability of choosing the t, and the probability of choosing the ν_0 (or ν_2).

8. Repeat steps 3 to 6 as much as needed.

5.3 Simulation study

5.3.1 Simulations

- **Experimental design**

The study considered simulations with nine occasions. Several reasons led to that choice. First, a larva would not be available for more than nine days between the third instar and the pupa. Second, nine occasions provide very good estimates with the LMM. And lastly, nine consecutive experimental days seems reasonable. Hence, a capture history is ten occasions long. The last occasion is different from 0 only if the individual was captured on occasion 9.

- **Population size and initial states**

The focus of the study was on urbanised areas, so small larval sites were simulated, such as those provided by abandoned tyres. The population size was initialised randomly between 500 and 1000 individuals, and all

instars were taken into account. The resulting number of individuals available for capture for the third instar, the fourth instar and the pupa state was between 150 and 300. For the first protocol, the population was maintained through hatching by randomly sampling additional individuals (around one-fifth of the initial population size) at each occasion. For the second protocol, the net that trapped the individuals prevented new eggs from being laid. Thus, no hatching occurred. Although eggs already present at the outset could hatch around the first and second occasion, these were ignored.

At the beginning of the experiment, there is no way to know if a larval site is at equilibrium. So the initial state probability was set to differ from the equilibrium. The initial state probability was set to:

Instar 1	Instar 2	Instar 3	Instar 4	Pupa
0.4	0.3	0.15	0.1	0.05

- **Capture effort**

A fixed number of captures was simulated at each occasion, which is easy to do experimentally. In addition, it helps in genotyping, as we know approximately how many samples need to be processed at each occasion. It also helps to plan the budget. I chose to sample 100 individuals per occasion. I assumed that all individuals had the same probability of being sampled on an occasion, but the probability was different between occasions. I took into account that if the number of individuals available is close to the targeted number, the target may be difficult to reach. Hence, if 110 or less individuals were available for a target of 100, between 90% and 100% of them were randomly captured .

For the first protocol, where the size of the population remains constant, the resulting capture rate is very high for populations of around 500 individuals. After the first occasion, there are less than 100 individuals available (in third instar or later states), while around 100 individuals are waiting to be released from the previous occasion. So the capture rate is above 90%. For populations with 1000 individuals, the capture rates are still between 0.35 and 0.45. For the second protocol, no eggs can be laid in the site, so the size of the population decreases. Thus, the capture probability increases to the limit of 0.9–1, even with an initial population size of 1000 individuals.

- **Development time**

In order to simulate transitions between instars as realistically as possible, the time spent by a larva in an instar is drawn randomly from an asymmetric normal distribution, as suggested by Birley (1979) [116]. I chose for the means the values from Figure 5.2 (b) [instead of Figure 5.2 (c)] because the development time is longer for the fourth instar. The standard deviations for the asymmetric normal distribution are those estimated for *Ae. aegypti* by Birley (1979) [116], although this is not the same species.

[116]: Birley (1979), 'The estimation and simulation of variable developmental period, with application to the mosquito Aedes aegypti (L.)'

Figure 5.4 gives an overview of the density of the time needed to go from the third instar to the fourth instar. The time spent in a state is defined according to a medium temperature that remains constant throughout

the experiment (assuming no extreme temperature has occurred). For simplicity, all the datasets were simulated at the same temperature, 25°C. The corresponding average transition probability (from the third instar to the adult stage) is:

$$\psi = \begin{pmatrix} 0.41 & 0.59 & 0 & 0 \\ 0 & 0.75 & 0.25 & 0 \\ 0 & 0 & 0.40 & 0.60 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

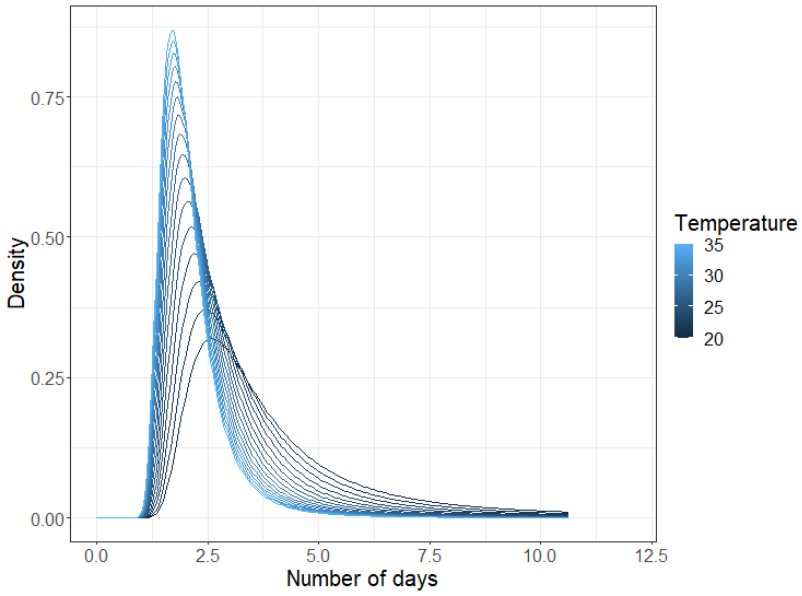

Figure 5.4: Probability density of time spent by larvae in state 3 depending on temperature.

- **Survival**

Taking the survival estimates from Table 5.2 and averaging the values over the days corresponding to each instar, we can obtain an average daily survival probability per instar. Additionally, I averaged them over the various sites and studies. These values are shown in Table 5.3 and in Figure 5.5.

	Instar 1	Instar 2	Instar 3	Instar 4	Pupa
Marsh	0.93	0.85	0.62	0.73	0.84
Borrow pit	0.93	0.87	0.73	0.61	0.65
Rabour pond	0.94	0.84	0.68	0.58	0.63
Chiga pond	0.91	0.81	0.72	0.62	0.65
Rice field	0.80	0.83	0.80	0.73	0.59
Pools	0.94	0.86	0.78	0.73	0.52
Sprayed rice	0.95	0.90	0.76	0.78	0.76
Average	0.91	0.85	0.73	0.68	0.66

Table 5.3: Average daily survival probability for each instar.

- **Death on capture**

It is very likely that the survival probability will differ if an individual is

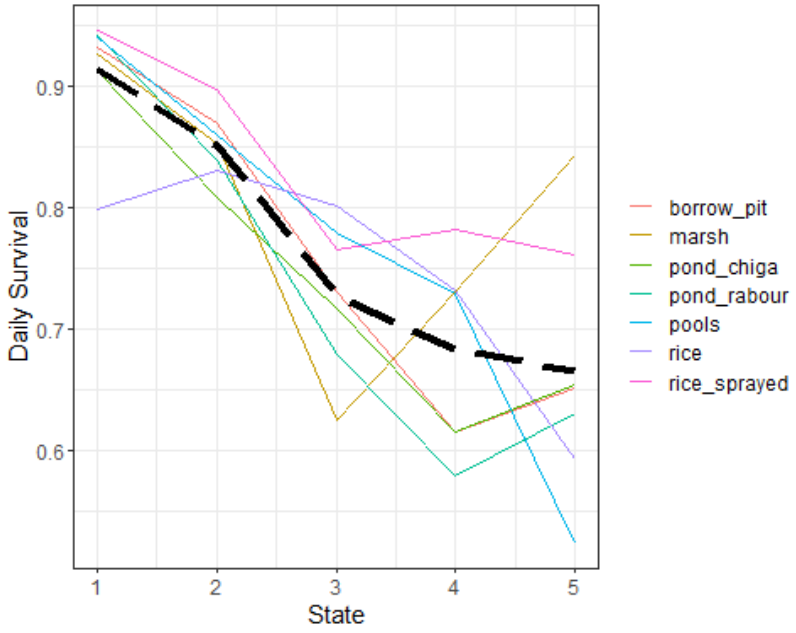

Figure 5.5: Average daily probability of survival by state. Coloured lines are the estimates and the black dashed line the average.

kept captive in a DNA-free water tube or if it is in its natural environment. However, we do not know if the 'survival during capture' will be higher or lower than the actual survival in the larval site. Since we can assume that accidental death may occur on capture, I chose to simulate a lower 'survival during capture'. First the population was simulated with its deaths, and then the captures were simulated. There is an additional 0.025 probability that the individual will die in the tube.

- **Number of simulations**

For each protocol, 100 datasets were simulated. For each simulation, survival and transition probability of each instar was estimated with the LMM.

5.3.2 Results

Figure 5.6 shows estimates of larvae survival when using the first protocol (pupae killed and identified without error). The results show that the survival is estimated without bias for the third instar and the fourth instar. However, as expected, the survival of the pupa is highly underestimated. This is explained by the fact that the two events – (1) that a pupa dies and (2) that a pupa becomes an adult – cannot be distinguished. Thus, we can only estimate 'apparent survival'.

Figure 5.7 shows estimates of larvae survival when using the second protocol (adults trapped, caught and identified without error). The survival probability is estimated without bias, independently of the true survival for all studied stages. But the uncertainty for pupae survival is quite high, especially for the lower values of survival.

When comparing the survival estimates of the two protocols, both estimates of third and fourth instar survival are equivalent. But only the second protocol allows pupae survival to be estimated. For both

protocols, confidence intervals are smaller for the fourth instar than for the other instars. The first protocol confidence interval increases when survival increases, while for the second, the confidence interval decreases when survival increases. The second protocol is more precise than the first when true survival is around 0.7 or more: the 95% confidence intervals are almost three-quarters smaller for the highest survival. For the lowest survival, the difference between the confidence interval for the two protocols is low.

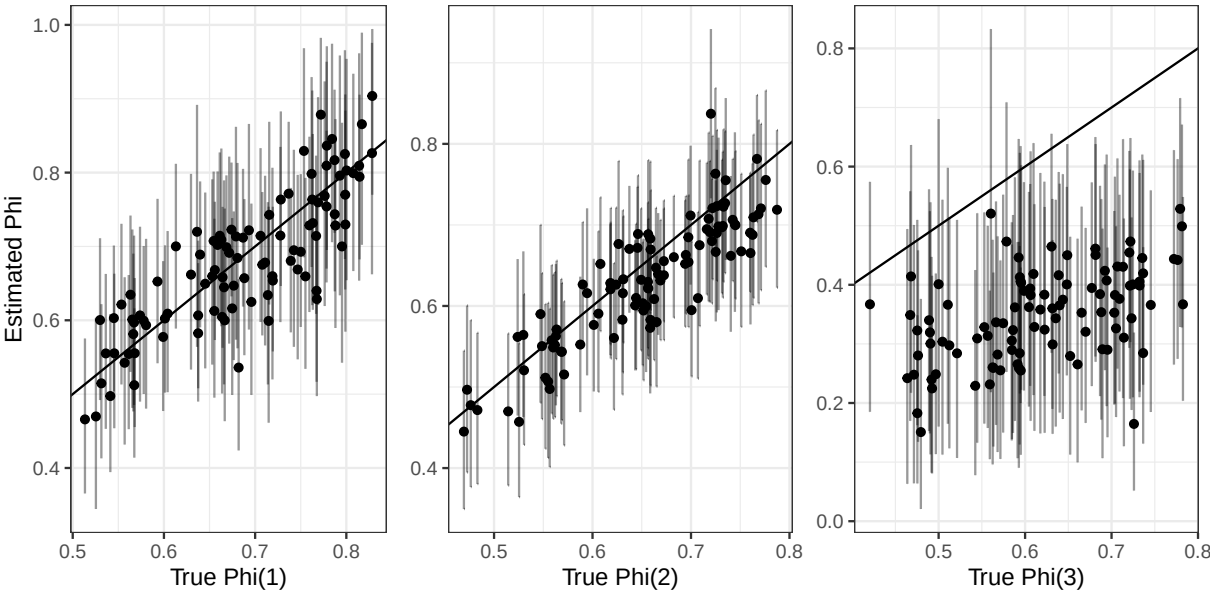

Figure 5.6: Estimates of survival for third (1) and fourth (2) instar and pupa (3), using the first protocol where pupae are killed. The equation of the black line is $y = x$.

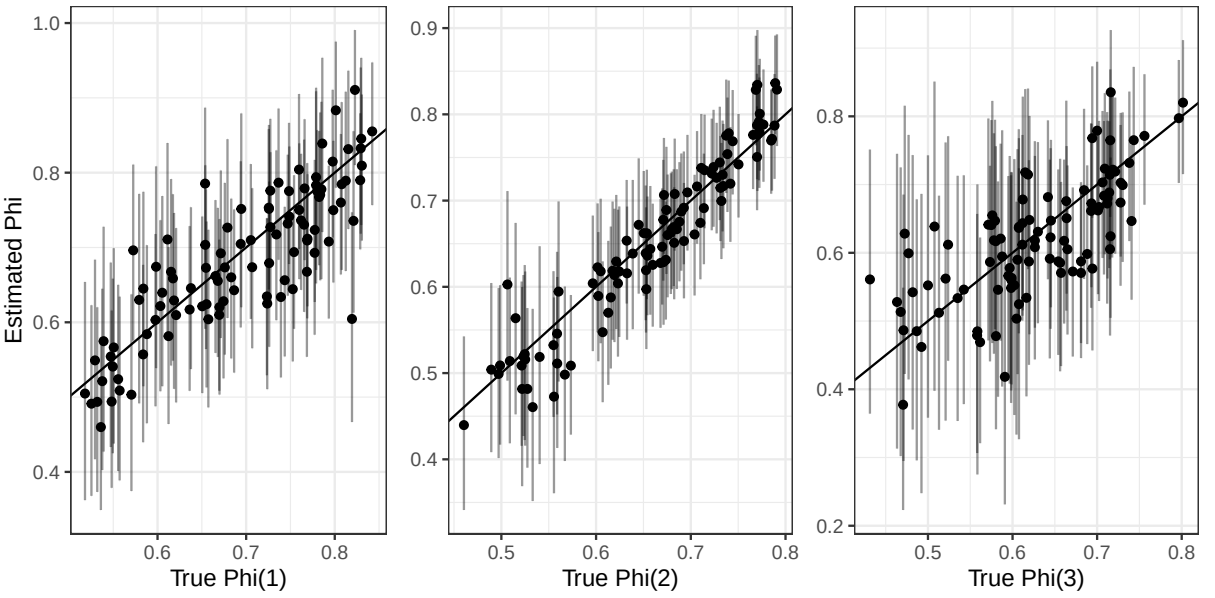

Figure 5.7: Estimates of survival for third (1) and fourth (2) instar and pupa (3), using the second protocol where adults are trapped and captured. The equation of the black line is $y = x$.

Figure 5.8 shows the estimates of the transition rates using both protocols.

The first protocol leads to biased estimates of all transitions. In addition, there are no estimates for transitions from the pupa stage. In contrast, the second protocol leads to unbiased estimates of all transitions, up to the adult stage.

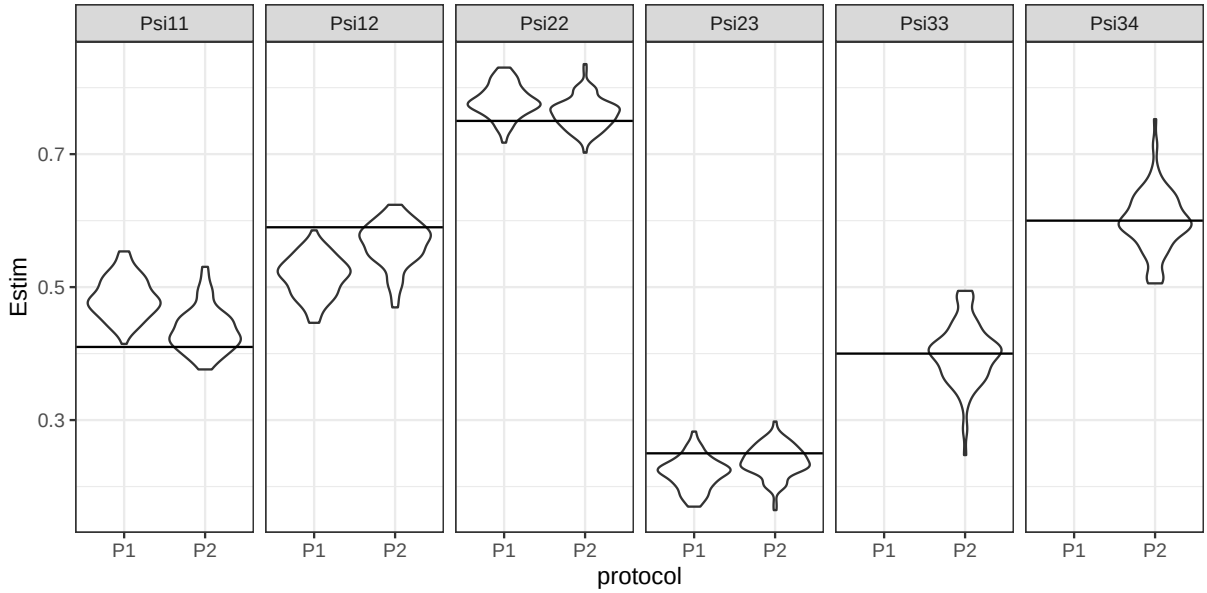

Figure 5.8: Estimation of transition rates for third (1) and fourth (2) instar and pupa (3), using both protocols. P1 is the first protocol (pupae killed) and P2 the second (adult captured). The third instar is number 3, the fourth instar number 4 and pupae number 5, so Psi33 is the probability that a third instar stays in third instar until the next occasion. The horizontal lines indicate the simulated transition rate.

5.4 Discussion

This study investigated an example of a capture-recapture experiment taking into account misidentification of individuals. This study design could be used to study mosquito larvae survival and transitions between stages with capture-recapture. We proposed two different field protocols. The first protocol consists of killing the captured pupae, while the second uses a net allowing the recapture of the adult stage of the released pupae. While the first protocol is easier to implement, the second allows the estimation of pupae survival. Simulations were also conducted to evaluate both protocols and check the accuracy of the survival estimations, even for the lowest simulated survival.

The findings show that the two evaluated protocols are bias-free for estimating the survival of third and fourth instars. But only the second protocol estimates pupae survival. Pupae survival is estimated with high uncertainty in cases of low survival ($0.5 < \phi < 0.6$). But the simulated survival in these cases is lower than the lowest survival reported in the literature ($\phi \approx 0.6$). This means that actual survival is likely to be higher, and lower uncertainty can be expected. The lower uncertainty on fourth instar survival may be due to the low transition rate from fourth instar to pupa, which makes the fourth instar stage longer, and as a consequence, more represented in the population.

Only the second protocol estimates the transition rate without bias.

Transitions are slightly biased in the first protocol. This may be due to the non-observation of pupae to adulthood. This lack of information is reflected in the transition estimates.

Overall, the second protocol performed better in the simulations. The price is that a net must be installed and controlled so that adult mosquitoes cannot escape from the study area, and that effort must be made to catch all adults in the net before a capture. These additional efforts make the second protocol less convenient to use. Furthermore, it is not known what consequences the net may have on the larvae environment. It protects the larvae from outside predators and prevents new eggs from being laid. This is likely to increase the survival of the larvae. If the aim is to study survival in the absolute, then testing the impact of the net is a prerequisite to using the second protocol. However, if the objective is to compare survival in different sites and perhaps link the difference in survival to environmental conditions (such as pollution), then it may not be a problem to alter survival. In fact, the effect of predators could be controlled by the net and help to compare survival based on environmental conditions.

This study is the first to put forward and test a CR protocol on mosquito larvae, and could pave the way for new experiments. As it is an original study, it was based on several hypotheses. As I explain below, some of these are more significant and likely to affect the results, and others are more minor.

The model I developed aimed to evaluate the effect of an environmental covariate on survival. This requires the experiment to be deployed on a reasonably large number of larval sites of equivalent size with different values for the environmental covariate. In such cases, the model could easily be extended to model the datasets from all sites together. Such a model would allow an estimation of the effect of the covariate of interest. If only a few sites are studied, or the effect of a covariate is too small to be detected, the model would still give estimates of several parameters of interest, namely the survival rate and the transition rate. Such estimates are very important for using models such as agent-based population models, which aim to realistically characterise the impacts of mosquito control.

The main hypothesis, conditioning the validity of the study design, is that the capture process does not affect true survival. I tried to meet this assumption by letting at least 24h pass before recapturing an individual in order to minimise the potential impact of the capture process. However, this hypothesis was not tested. The fact that a larva has been captured previously could lower its survival probability even after being released. In the same way, the study design relies on the hypothesis that the capture process does not change the speed of the immature life cycle. This was not tested either. In addition, with the second protocol, the larvae population size will decrease because no hatching occurs. This may affect the survival of the remaining larvae: for example, because competition will decrease.

In the model it was also assumed that the transition probability was the same in the tube during a capture and in the larval site. This may not hold because individuals in tubes have nothing to feed on, which may delay their growth, for example. This could be easily resolved by

modelling different transition probabilities, at the likely cost of increased uncertainty of the estimates.

For the second protocol, it was assumed that all adults are caught with a probability of 1 on the first occasion after emergence. If the actual capture probability of adults is lower, then this would need to be modelled. Estimating the capture probability of adults would require estimating how many of them escape the net. This appears complicated and would likely increase the uncertainty concerning pupae survival.

It was also considered that all stages had the same capture probability. This could be the case using a strict protocol, but pupae behave differently than larvae and are highly mobile to avoid predators: as such they are harder to catch. If the pupae capture probability must be estimated apart from the other stages, the uncertainty might be high, especially given that there are few available pupae. This increased uncertainty of capture probability would likely increase the uncertainty of pupae survival.

The model was also based on state-independent identification probability. If more data was available, identification probability could be considered state-dependent if there is evidence that some states lead to better data for identification, or the model could use a covariate of quality as shown in Section 2.2.2 and Section 4.2.5.

For the sake of simplicity, second instar larvae were not considered. They could easily be captured, so the model would be able to estimate the transition rate from second to third instar and second instar survival. This could help improve to some extent the uncertainty of third instar survival. But it would increase the population size of interest in a larval site and so diminish the capture rate if the number of captures is kept constant as in my simulation.

The transition time for each state of a larva was simulated independently. Thus, a larva could take a very short time to get from an instar to the next stage (compared to other larvae) and then a very long time to get to the following stage. This may not be very realistic biologically, but should not affect the validity of the simulations.

It would be valuable to test these protocols in the field, which would improve the quality and usefulness of the LMM and allow it to be developed. This study remains conceptual with no proof on the practical feasibility of the experiment. However, I hope that it will help lead to new experiments in the future for mosquitoes or for other species.

Discussion

Discussion

The original objective of this thesis was to develop a capture-recapture model to estimate the effect of environmental covariates on the survival of mosquito larvae. As individual misidentification of mosquito larvae was expected, the goal was to integrate this in the model. The protocol to obtain the data about the larvae was developed in parallel by another team. After two years, it became clear that the protocol would not be ready to use during the thesis as initially planned, and that no data about larvae would be available. For that reason, the focus was shifted to the sole question of modelling misidentification in CMR, and I proceeded with the mosquito case to explore the potential offered by the model.

I will first discuss the extensions of the LMM that were developed (*The latent multinomial model*) and some limitations of this.

I will then discuss the application of the models (*Applications*).

I will then discuss the implementation of the models (*Code*). I will summarise what has been achieved and what is yet to be done.

Finally, the section *What comes next* examines the perspectives of the results for current and future methods of planning and analysing CR data from low-quality DNA, and more broadly in the presence of potential misidentification, as well as possible future extensions of this model.

The latent multinomial model

The model and its extensions

The latent multinomial model (LMM) stands as a robust tool within the field of CMR models, specifically designed to handle the complexities of misidentification. This model takes into account the unavoidable errors that occur when identifying individuals from low-quality eDNA. One of the distinct advantages of employing the LMM is its ability to estimate misidentification rates from traditional capture-recapture data (unlike models that estimate genotyping errors such as Wright's model). The LMM uses all available data from CMR studies, ensuring that no potentially valuable information is left unused. This is better than Yoshizaki's model, which discards histories with single captures. In addition, the LMM is a highly adaptable framework, making it a good choice for a wide range of applications beyond a specific study. Its flexibility sets it apart from other models designed to address misidentification issues.

In the first step, I covered all the main models of a closed CMR, which aims to estimate population size. I started by extending the LMM to multiple states. Such a model could potentially be used for studies like the one on great crested newts by Worthington et al. (2019) [89], where the states correspond to ponds. In that study, the newts were identified manually based on visual patterns. It was assumed that no misidentification occurred. However, misidentification could have occurred. And in similar

[89]: Worthington et al. (2019), 'Estimation of population size when capture probability depends on individual states'

studies, the matches of photographs could be made by an algorithm, potentially generating misidentification. This covers the closed-population cases of CMR, especially since the closed population model was extended to account for individual heterogeneity by McClintock et al. (2014) [75].

[75]: McClintock et al. (2014), 'Probit models for capture-recapture data subject to imperfect detection, individual heterogeneity and misidentification'

I then extended the LMM to open populations, first for a single state with the CJS framework (CJS_α model), and then for multiple states with the Arnason-Schwarz framework (AS_α model). These extensions allow the estimation both of survival and transitions between states. Ignoring misidentification in standard models can lead to biased survival estimates if misidentification occurs. While the CJS model estimates survival, not recruitment, I gave guidelines to write an open population LMM that would estimate recruitment.

I gave particular attention to the case of low recapture rates (between 0.1 and 0.2). In that situation, Yoshizaki et al. (2011) [67] and Vale et al. [77] highlighted that the parameters of their models were unidentifiable. In the LMM, I proposed the use of identification quality data, readily available with genetic tags, to improve the estimates. I developed an extension of the LMM that incorporates a measurement of identification quality as a covariate of identification probability, using a probit model. Provided that there are enough occasions (nine or more) and a high identification probability ($\alpha \geq 0.9$), the model can estimate the population size without bias, even in cases of very low capture rates ($p = 0.1$). And in any case, this model performed better than the model without a covariate. This extension can be used for any kind of data, as long as a continuous measure of quality can be obtained about the identification process. For example, with photographic tags, a continuous measure could be developed that would take into account the angle of the photo and its clarity or blurriness.

Finally, one of the key assumptions of the original LMM is that individuals can only be captured once per occasion. This assumption was constraining. It is unrealistic for many studies such as the ones using DNA from faeces. To address this, I proposed allowing individuals to be spotted several times on one occasion in the LMM.

General limitations

In addition to the specific limitations discussed in each chapter, three general points can be added that concern this study.

First, while the simulations conducted provided valuable insights, it is worth noting that due to the multitude of scenarios explored, the simulations were limited to ten in many cases. There is little reason why significantly different results would emerge with additional simulations. But the open population scenario results showed some inconsistencies. Bias was observed for five occasions, $\phi = 0.6$ and $p = 0.4$, but not with $p = 0.3$. For these scenarios, more simulations might improve the results.

Second, the two extensions developed in Chapter 2 and Chapter 3 were not developed for all cases. The probit model was developed and implemented for a single state only (closed or open population). In an open population, the probit model was tested on some simulations, but

not with repetitions on a complete set of scenarios. The Poisson model was only developed in a single state for a closed population. However, the formulation of the LMM as two independent processes, capture and identification, would easily allow the model to be extended. Extending the probit and Poisson models to multi-state and an open population would greatly extend the LMM's range of applicability.

Applications

The third limitation concerns model validation. The use of real-world data, particularly to validate the probit model and the larval model, would have been highly beneficial. I tried twice to obtain data to validate the probit model, but unfortunately the data would not have been suitable for the model. A great deal of time was spent trying to obtain this data and analysing it to find it would not suit the needs of the project.

However, I was able to obtain a real dataset on otters to illustrate the Poisson extension of the LMM. I also developed and compared two protocols for collecting mosquito larvae.

Application to otters

In one application, I used the model $M_{\lambda,\alpha}$ on a dataset on the Eurasian otter. I compared its estimates with those of the M_t model and Yoshizaki's model. The results showed that when there was evidence of misidentification, the $M_{\lambda,\alpha}$ model provided estimates similar to Yoshizaki's model, but with enhanced precision.

In this thesis study, running the LMM in parallel to Yoshizaki's model and the model M_t helped identify the years where misidentification occurred. The consensus between the $M_{\lambda,\alpha}$ model and Yoshizaki's model for most years indicates that the model can be used on real data.

Application to mosquito larvae

The previous chapter describes a model developed specifically to analyse data from a study on mosquito larvae. Although the data acquisition protocol is still in development, I conducted a simulation study to compare the two potential protocols and determine the advantages and disadvantages of each. The previous chapter explains that provided certain assumptions are valid, one of the protocols performs better, and the survival and transition probabilities between the various instars can be estimated without bias. This represents the first such protocol for mosquitoes, and the simulation tests show the potential of the LMM for studying specific situations in the wild.

The final objective of the project on mosquito larvae was to study the effect of environmental covariates on survival. Although I developed a model to estimate survival, I did not develop a model to analyse the effect of environmental covariates as originally planned. Based on the larvae extension of the LMM, I do not think this will be a difficult problem. But as a consequence, I did not perform a sensitivity analysis on an estimate

of the covariate effect, but simply estimated survival and observed the uncertainty arising on simple estimates. It is unknown if the precision of the estimates will allow the detection of a potential environmental effect on the survival rate. These further analyses remain to be done in future studies, when data is available.

Thus, the final step of the application of the model will be carried out when data is collected. For simplicity, I did not consider capturing second instars, but the final protocol may allow their capture and identification. In that case, the model will need to be modified accordingly. The code modifications for that should be minor and simple to implement.

Code

A major downside of the LMM is the near absence of the practical implementation of the model. To date, the only available implementation (Vale et al. 2014 [77]) was for a single-state closed population and was written with the Automatic Differentiation Model Builder (ADMB) [76]. Consequently, the many studies focusing on open populations or transitions between sites or states cannot make use of it, and modifying this implementation is out of reach of most researchers. As a result, researchers interested in using the LMM find themselves with the challenge of writing their own implementation, which is a barrier to its widespread use.

[77]: Vale et al. (2014), 'Maximum likelihood estimation for model $M_{t,\alpha}$ for capture–recapture data with misidentification'

[76]: Fournier et al. (2012), 'AD Model Builder'

At the outset of my thesis project, I began by discovering the LMM published by Link [74] and the improvements made to the sampling algorithm by Schofield & Bonner [78, 79]. I then implemented the model using the R language and the package NIMBLE for all the extensions described in this thesis. This type of implementation has great potential for specific use cases, as the R language is widely used by researchers for CMR, and they could modify the code to fit the specific use they have. This is an improvement over the sole implementation previously available from Vale et al. (2014) [77].

[74]: Link et al. (2010), 'Uncovering a Latent Multinomial'

[78]: Schofield et al. (2015), 'Connecting the latent multinomial'

[79]: Bonner et al. (2015), 'Extending the latent multinomial model with complex error processes and dynamic markov bases'

In *The Lord of the Rings*, in order to destroy the ring, Frodo first has to get to Mordor: and this step is the longest. In this thesis project, the first step of understanding and implementing the LMM as published by Link with the improved algorithm from Bonner & Schofield was also the longest. Yet despite the intricacies and complexities of the model, NIMBLE proved a flexible enough tool that even as somewhat of a novice I could get the model to work.

Now that the model has been generalised to the basic cases, the main limitation to its use is the absence of accessible implementation. During this project, I developed the R/Nimble codes for running all the models detailed in this thesis. These codes constitute a solid basis for implementing a package that will allow anyone without specific knowledge in NIMBLE code or a detailed understanding of the model to run it. In a next step, I plan to develop such a package.

What comes next

A new way to think about capture-recapture

This study could lead to major changes in the way we approach and use genetic data in ecological studies. Currently, researchers often face the problem of dealing with risky data, which leads to the exclusion of valuable information. By extending the latent multinomial model (LMM) to all general cases of CMR, this work paves the way for collecting and using more data, rather than focusing only on high-quality samples.

As the cost of sequencing continues to fall, and new methods such as genotyping-in-thousands [117] emerge, there is a growing opportunity to explore the potential of low-quality DNA samples. This is leading to the current transition from the use of microsatellites to SNPs in DNA studies. The latter have many advantages, such as being in short sequences and being bi-allelic. Although they are less informative for identification, using larger panels compensates for this problem. In contrast to microsatellites, the larger panel makes it much less likely to lead to the recapture of a 'ghost' (i.e. ghost histories with several captures). The required hypothesis that ghosts are unique is therefore much more robust with SNPs.

[117]: Campbell et al. (2015), 'Genotyping-in-Thousands by sequencing (GT-seq)'

Ongoing studies could be improved by reconsidering any poor quality data that has already been collected. Studies that are being planned can go even further. By considering samples not yet taken into account because of their low quality, this could broaden the range of data that can be used. For example, studies using faecal samples usually collect only fresh faeces. The model would allow considering the collection of older faeces with more degraded DNA. This development may lead to new ideas and methods for DNA collection, particularly for species that have been challenging to study using capture-mark-recapture methods. As this thesis has shown that a study with a low capture rate (around 0.1) is still possible using a quality covariate, studies could be developed with an expected low recapture rate and potential misidentification. The implications of these advances are far reaching and offer exciting prospects for the field of population dynamics.

Future developments of the model

In his 2014 paper, McClintock identified several extensions needed for the LMM: "to allow misidentifications to match legitimate individuals" and "to allow for ghost histories consisting of multiple encounters". The second implies that a ghost can be resighted: that is, several misidentifications could lead to the same ghost. Although this thesis did not tackle these challenges, I discuss some ideas about solving them below.

"Allowing identification errors to match legitimate individuals"

Chapter 2 mentions matching misidentifications to legitimate individuals and discusses how to model the identification rate as a function of the relatedness. Going a step further, I suggest the identification probability is modelled as a function of relatedness using two components, one intra

and one inter individual component. The intra component would use the relatedness of a sample to the other samples in the same history. This component would inform on the probability of matching the sample with the individual it has been matched to. The inter component would use the relatedness of the sample to all other samples of other individuals. This component would inform on the probability of matching the sample as another individual. A drawback is that the computations might be time-consuming.

To make the MCMC efficient and limit computation time, the relatedness could be used to propose $\mathbf{x}'$ in an efficient way. Say we have a vector of continuous data r indicating the relatedness of a potential ghost to all individuals. This vector could be used in a very simple way in the algorithm by proposing a new latent set. When adding an error to the set and after having sampled a potential ghost (ν_1), instead of uniformly sampling the individual in which the ghost will be placed (the ν_0), it can be sampled proportionally to its relatedness r to the ghost. The proposal density $[\boldsymbol{X}'|\mathbf{X}]$ is adapted to take it into account.

Another 'naive' idea is for simple situations when the population is small, and the number of possible misidentifications is low. In these cases, it might be possible to identify all the likely sets of latent histories. With this prior work, one would only need to move between the different identified sets. The algorithm would be much simpler than it is currently. We would need to sample a latent set from the identified sets. This could be done either uniformly or proportionally to the difference in the number of misidentifications between the current and the proposed set, to avoid large jumps that would likely be rejected. Of course, the initial definition of likely latent sets would be very important in this design. If some possible latent sets are forgotten, problems may arise. If the number of sets is not too large, maximum likelihood might be possible.

"Allowing for ghost histories to consist of multiple encounters"

In this study, the hypothesis was retained that ghosts can never be resighted, so their histories only have one capture. Here I suggest some ideas about how a model could be constructed that would allow for resighting ghosts. We will suppose that only ghosts from the same individual can be resighted. For example, an individual A with the true capture history 11111, misidentified on occasions 2 and 3, may generate the ghost 01100. However, two different individuals A and B, respectively misidentified on occasions 2 and 3, cannot generate together the ghost 01100.

Say a latent history generated several ghosts identified as the same individual. For example, the true history 101111 can be observed as histories 10010, 00101 and 000001. Then in the latent history, the two misidentifications are linked to indicate that they were identified together. We can rewrite the latent histories with a superscript for each linked misidentification. In the example, the first and second misidentifications were linked, so we write the latent history $102^{(1)}12^{(1)}2^{(2)}$. The superscript "(1)" indicate that the capture were identified as the same ghost, and the "(2)" was identified as another ghost. Then we can reconstruct a matrix $\mathbf{A}$ that gives $\mathbf{y} = \mathbf{A}\mathbf{x}$. If we consider the identification rate to be the same for

any number of linked misidentifications, the likelihood would actually be the same as for the $M_{t,\alpha}$ model. The algorithm proposing a new set of latent histories would change to sample any history as a potential ghost. However, the uniform sampling of the ghost histories would probably be very inefficient.

Evolving marks

Yoshizaki et al. [64] highlighted another challenge that the LMM could be faced with: evolving natural tags. This is particularly important for studies using visual marks such as scars. When scars are used as a natural identification marking (because no two individuals have the exact same scars), issues can arise if new scars appear between two captures, hiding the older scar. If the pattern changes to the point that the individual is not recognisable, the result is that one individual history can be divided into several histories. For example, the history 1111 could be observed as three histories: 1000, 0110 and 0001. Yoshizaki developed a model 'EV' (EvolVing marks) for such cases, but did not make the latent states explicit. Similar to how Link et al. (2010) [74] developed the $M_{t,\alpha}$ model, the latent histories in the EV model could be made explicit. Yoshizaki gives the likelihood of the latent histories. The first time an individual is captured, it cannot be misidentified because the marks cannot have changed from a previous situation.

[64]: Yoshizaki et al. (2009), 'Modeling misidentification errors in capture–recapture studies using photographic identification of evolving marks'

I suggest that in the latent history, each time an individual is misidentified and a new history is created, the capture is noted with a new number. Then the probability of latent history 1223 is $p_1p_2(1-\alpha)p_3\alpha p_4(1-\alpha)$. The capture process is the same as model M_t, so $[\mathbf{z}|N,\mathbf{p}]$ from Section 1.2.3 does not change. The identification likelihood for history 1223 is then $(1-\alpha)\alpha(1-\alpha)$.

Let n_j be the number of captures occurring in the history v_j, and c_j the sequence of numbers different from 0 in v_j (i.e. 122 for history 10202). The identification likelihood is:

$$[\mathbf{x}|\mathbf{z},\alpha] = I(\mathbf{z}=\mathbf{B}\mathbf{x})\prod_j\prod_{l=2}^{n_j}\alpha^{I(c_{j,l}=c_{j,l-1})}(1-\alpha)^{I(c_{j,l}\neq c_{j,l-1})}.$$

Then, to estimate the parameters, we just need to change step 5 of Section 1.2.4 so that it can correctly combine the histories from the same individual.

It is likely that the longer the time between two captures of an individual, the greater the probability that the mark has changed. Thus, the identification probability could potentially be modelled as a function of passed time since the last capture.

This model would be very close to the LMM developed in this thesis project. It could easily be implemented and applied to simulations to test its effectiveness. Moreover, many datasets should already be available to test the model on real data. Scar patterns are widely used to study marine mammals. In amphibians, visual implant elastomer (VIE), a visual tag of coloured liquid injected subcutaneously, is widely used. Grant (2008) [118] shows that these marks can be subject to changes, causing

[118]: Grant (2008), 'Visual implant elastomer mark retention through metamorphosis in amphibian larvae'

misidentification of the individuals by observers. Such data illustrate the interest of the model proposed here.

Other extensions and conclusion

For studies of reproductive success, multi-event models are often used because individuals can be spotted with uncertainty concerning their reproductive state. To deal with this problem, multi-event models (Pradel (2005) [119]) include uncertainty in the observed state. Using photographic tags for such a study would additionally lead to the challenge of misidentification. Thus, extending the LMM to multi-event cases would prove useful.

[119]: Pradel (2005), 'Multievent'

The model description would remain relatively straightforward. The observation process of the LMM would be replaced by the observation process outlined in Pradel (2005) [119]. It is very likely that uncertainty in the observation process added to the misidentification process would lower the precision of the estimate, especially if misidentification depends on the state. Nonetheless, I think that with reasonable capture probability values, the uncertainty could be kept at a reasonable level.

Lastly, seeing how the capture-recapture paradigm has shifted to spatial data in recent years, there is a need for models such as the LMM that will allow the collection of more data of lower quality by dealing with misidentification. Augustine et al. (2020) [120] took a step in this direction by developing a statistical framework that uses spatial proximity to mitigate genotype uncertainty in genetic tagging studies.

[120]: Augustine et al. (2020), 'Spatial proximity moderates genotype uncertainty in genetic tagging studies'

To conclude, the LMM is a generalisation of standard CMR models that uses latent histories to describe hidden processes. All extensions developed in this thesis project are steps towards completing a 'full' LMM-CMR framework. I hope this project has contributed to that aim. When most imaginable cases can be modelled, the limits of research become mainly those of technical feasibility and imagination.

And so basically, yeah.

Synthèse en français

Synthèse

Introduction

L'identification individuelle basée sur des marques naturelles est largement utilisée dans les études de capture-recapture (CR) pour estimer la taille des populations, la survie en conditions naturelles ou bien encore les transitions d'états géographiques ou physiologiques. On parle alors de marquage non invasif. Les marques naturelles peuvent être constitué d'ADN extrait de fèces, plumes ou poils par exemple. Des études utilisant l'ADN comme marque ont été menées sur plusieurs taxa comme des ours ([50]), des lynx roux ([51, 52]), des antilopes ([53]) ou encore des éléphants ([54]). Les motifs visuels sont un autre exemple de marquage naturel. Les motifs peuvent être des rayures naturelles ou bien des cicatrices par exemple. Des études utilisant la reconnaissance par photo ont été menées sur des mammifères tels que les baleines ([46]), les dauphins ([45]), les léopards ([43]) ou encore des insectes comme les coléoptères ([44]). Bien que l'échantillonnage non invasif permette d'étudier des espèces en liberté sans avoir à capturer les individus, les manipuler ou même, dans le cas d'une identification génétique, les observer. Ces méthodes présentent certaines limites. En particulier, par rapport aux méthodes de marquage traditionnelles. En effet le risque d'identification incorrecte des individus est beaucoup plus élevé lorsque les marques sont basées sur des caractéristiques naturelles ([49]). Si les erreurs d'identification sont ignorées, les modèles standard peuvent alors surestimer largement la taille de la population ([65]) puisque de nouveaux individus factices vont être créés.

Au niveau de l'échantillonnage de l'ADN, plusieurs études ont proposé des solutions pour réduire les erreurs d'identification. Ces solutions couvrent les méthodes de terrain et l'amélioration des techniques de laboratoire pour l'analyse génétique ([60, 68]), jusqu'aux logiciels de pré-analyse qui aident à filtrer les données susceptibles de contenir des erreurs ([69]). En ce qui concerne la reconnaissance des schémas visuels, des techniques d'appariement d'images assistées par ordinateur ([47, 70]) ont été mises au point pour faciliter l'identification. De plus un logiciel d'analyse sous R traite les données pour lesquelles des photographies des côtés gauche et droit des individus sont disponibles sans moyen fiable de les faire correspondre ([63]). En outre, diverses approches ont été proposées pour prendre en compte des erreurs d'identification dans les modèles d'estimation de la taille des populations ([67, 71, 72, 74]). Aujourd'hui, la pratique la plus courante consiste encore à filtrer en retirant les photographies de mauvaise qualité ou les échantillons d'ADN mal génotypés.

Cependant, le rejet d'un pourcentage non négligeable de données de mauvaise qualité peut avoir pour conséquence que les données conservées sont trop peu nombreuses pour permettre une estimation fiable des paramètres d'intérêt. Parmi les études utilisant des marques naturelles citées ci-dessus, cinq ont rejeté entre 20 et 40% des échantillons collectés. Dans ces cas, il aurait pu être avantageux d'autoriser un petit degré d'incertitude dans l'identification, environ 1-5% comme le suggère [71], en modélisant le taux d'erreur ([67, 71–74]). Si le coût de l'ajout d'un

[50]: Dreher et al. (2007), 'Noninvasive estimation of black bear abundance incorporating genotyping errors and harvested bear'

[51]: Ruell et al. (2009), 'Estimating bobcat population sizes and densities in a fragmented urban landscape using noninvasive capture–recapture sampling'

[52]: Morin et al. (2018), 'Efficient single-survey estimation of carnivore density using fecal DNA and spatial capture-recapture'

[53]: Woodruff et al. (2016), 'Estimating Sonoran pronghorn abundance and survival with fecal DNA and capture—recapture methods'

[54]: Laguardia et al. (2021), 'Nationwide abundance and distribution of African forest elephants across Gabon using noninvasive SNP genotyping'

[46]: Curtis et al. (2021), 'Abundance, survival, and annual rate of change of Cuvier's beaked whales (Ziphius cavirostris) on a Navy sonar range'

[45]: Labach et al. (2022), 'Distribution and abundance of common bottlenose dolphin (Tursiops truncatus) over the French Mediterranean continental shelf'

[43]: Swanepoel et al. (2015), 'Density of leopards Panthera pardus on protected and non-protected land in the Waterberg Biosphere, South Africa'

[44]: Quinby et al. (2021), 'Estimating population abundance of burying beetles using photo-identification and mark-recapture methods'

[49]: Taberlet et al. (1999), 'Noninvasive genetic sampling'

[65]: Creel et al. (2003), 'Population size estimation in Yellowstone wolves with error-prone noninvasive microsatellite genotypes'

[60]: Waits et al. (2005), 'Noninvasive Genetic Sampling Tools for Wildlife Biologists'

[68]: Paetkau (2003), 'An empirical exploration of data quality in DNA-based population inventories'

[69]: McKelvey et al. (2005), 'dropout'

[47]: Crall et al. (2013), 'HotSpotter - Patterned species instance recognition'

[70]: Bolger et al. (2012), 'A computer-assisted system for photographic mark–recapture analysis'

[63]: McClintock (2015), 'multimark'

paramètre (le taux d'erreur) est compensé par le nombre supplémentaire d'échantillons qui peuvent être conservés, le compromis est intéressant.

Erreurs d'identifications

Nous faisons les hypothèses suivantes sur les erreurs d'identification:

- un individu réel ne peut pas être confondu avec un autre individu réel,
- une erreur d'identification crée un nouvel individu qui n'existe pas réellement, un "fantôme",
- les erreurs sont toutes unique et deux erreurs ne peuvent donc pas être reliées au même fantôme.

Exemple pour un individu vu trois occasions consécutive mais mal identifié à la troisième :

Histoire réelle	Histoires observées
111	110
	001

Parmi les approches qui intègrent un processus d'erreur d'identification dans l'analyse, deux types de modèle sont utilisés.
Le premier type de modèle intègre l'incertitude de génotypage en calculant par exemple, la probabilité que deux échantillons aient réellement le même génotype en sachant les génotypes observés. Dans cette classe de modèles, on trouve le modèle de Wright et al. (2009) [72] ou encore celui de Knapp et al. (2009) [73]. Ces approches sont cependant limitées aux identifications basées sur de l'ADN.
Le second type de modèles consiste à ajouter un paramètre de probabilité d'identification correcte d'un événement de capture. Au sein de cette classe de modèle, le modèle de Yoshizaki et al. (2011) [67] est le plus simple, au prix de l'élimination de toutes les histoires contenant une unique capture. Ainsi, toutes les histoires susceptible de résulter d'une erreur d'identification sont retirées. Le modèle peut donc se concentrer sur l'estimation de la taille de population sans se soucier d'erreurs d'identifications d'histoire avec une unique capture. Cependant, le modèle ne fonctionne pas bien lorsque les recaptures sont peu nombreuses.
Un autre modèle du second type, est le modèle multinomial latent (LMM, [74]). C'est un cadre malléable qui a particulièrement retenu l'attention et a été développé dans plusieurs publications ([75, 78, 79]). Il estime, dans un cadre bayésien, le taux de bonnes/mauvaises identifications sans avoir besoin d'informations supplémentaires que la matrice des histoires de capture. Ainsi il peut être utilisé quelque soit la façon dont les individus sont identifiés, que ce soit à l'aide d'ADN ou de marques visuelles. La thèse se concentre sur ce modèle.

En dépit de l'intérêt qu'il soulève et de son potentiel, le LMM présente plusieurs limitations. Premièrement, les simulations de Vale et al. (2014) [77] montrent que ce modèle n'est pas performant dans les scénarios de faibles recaptures.

Deuxièmement, le LMM n'utilise pas toute l'information disponible. En effet, si les identifications sont faites par génotypage, une mesure de qualité du génotype observé peut être obtenue. Par exemple, le

[67]: Yoshizaki et al. (2011), 'Modeling misidentification errors that result from use of genetic tags in capture–recapture studies'
[71]: Lukacs et al. (2005), 'Estimating population size from DNA-based closed capture-recapture data incorporating genotyping error'
[72]: Wright et al. (2009), 'Incorporating genotype uncertainty into mark-recapture-type models for estimating abundance using DNA samples'
[74]: Link et al. (2010), 'Uncovering a Latent Multinomial'
[67]: Yoshizaki et al. (2011), 'Modeling misidentification errors that result from use of genetic tags in capture–recapture studies'
[71]: Lukacs et al. (2005), 'Estimating population size from DNA-based closed capture-recapture data incorporating genotyping error'
[72]: Wright et al. (2009), 'Incorporating genotype uncertainty into mark-recapture-type models for estimating abundance using DNA samples'
[73]: Knapp et al. (2009), 'Incorporating genotyping error into non-Invasive DNA-based mark—recapture population estimates'
[74]: Link et al. (2010), 'Uncovering a Latent Multinomial'

[77]: Vale et al. (2014), 'Maximum likelihood estimation for model $M_{t,\alpha}$ for capture–recapture data with misidentification'

séquençage Illumina donne un q-score ([92]) et une méthode de mesure de qualité est proposée par Miquel et al. (2006) [91] pour une PCR multitube. Ainsi on peut penser que de prendre en compte une mesure de qualité des génotypes dans le modèle pourrait améliorer les estimations. Bien que McClintock et al. (2014) [75] étende le modèle pour incorporer l'hétérogénéité de capture ou d'identification individuelle, son modèle ne prend pas en compte une hétérogénéité au niveau de l'échantillon.

[92]: Illumina (2011), 'Quality scores for next-generation sequencing'

[91]: Miquel et al. (2006), 'Quality indexes to assess the reliability of genotypes in studies using noninvasive sampling and multiple-tube approach'

[75]: McClintock et al. (2014), 'Probit models for capture-recapture data subject to imperfect detection, individual heterogeneity and misidentification'

Troisièmement, le LMM suppose qu'on ne peut pas observer un individu plus d'une fois par occasion. Or de nombreuses études utilisent l'ADN extrait des fèces. Dans ce cadre expérimental, les individus peuvent être "capturés" plusieurs fois à la même occasion. Ainsi le LMM n'est pas adapté à ce cas de figure. Enfin, le LMM a été étendu en population ouverte par Bonner et al. (2015) [79] mais pour un type d'erreur d'identification différent. Dans cette étude, les individus peuvent être confondus, contrairement à la première hypothèse que l'on fait sur les erreurs d'identification. De plus le modèle n'a pas été étendu aux cas multi-état pour l'étude des changements d'états géographiques ou physiologiques.

[79]: Bonner et al. (2015), 'Extending the latent multinomial model with complex error processes and dynamic markov bases'

Dans cette thèse, j'étends le LMM dans trois directions. Une première extension permet d'intégrer une covariable de qualité d'identification. La deuxième extension permet l'analyse de jeux de données où les individus sont potentiellement vus plus d'une fois par occasion. Ces deux extensions sont considérées dans le cadre d'une population fermées, comme le modèle d'origine. Trois autres extensions permettent de considérer plusieurs modèles en population ouverte: le modèle à un seul état sans covariable, à un seul état avec une covariable d'identification et le modèle en multi-état sans covariable. Ce dernier modèle sera utilisé dans le dernier chapitre de la thèse qui propose et compare deux protocoles de terrains pour l'étude de la survie de larves de moustiques. Nous considérerons cette étude à travers des simulations.

Modéliser les erreurs d'identification

Le modèle multinomial latent ($M_{t,\alpha}$)

Une erreur d'identification est le fait de ne pas reconnaître un individu lorsqu'il est capturé, et d'assigner la capture à une histoire qui n'est pas la sienne. Nous faisons les hypothèses suivantes concernant les erreurs d'identifications:

- une erreur d'identification conduit toujours à la création d'une histoire fantôme,
- deux individus ne peuvent pas être confondus.
- deux erreurs ne peuvent pas être liées à la même histoire fantôme, elles créent chacune un fantôme différent.

Par exemple, si un individu a l'histoire de capture réelle 111 mais qu'une erreur a été commise à la deuxième occasion, alors on observera les histoires 101 et 010. Dans cet exemple, l'histoire 010 est une histoire fantôme. Aucun individu réel ne correspond à cette histoire et il ne peut pas y avoir d'autres captures qui seront liées à l'histoire fantôme 010.

Pour tenir compte des erreurs d'identification, Link et al. (2010) [74] ont développé le modèle latent multinomial (LMM). Le modèle utilise des histoires d'erreurs latentes dans lesquelles les erreurs d'identification sont notées par des 2. Dans l'exemple précédent, l'histoire latente est 121. Les histoires d'erreurs latentes sont notées $\boldsymbol{\nu}_j = (\nu_{j,1}, ..., \nu_{j,T})$. La fréquence de l'histoire latente ν_j est notée x_j, et le vecteur de toutes les fréquences des histoires latentes d'erreurs est $\mathbf{x} = (x_1, ..., x_{3^T})$.
Pour faciliter les développements futurs du modèle, Bonner et al. (2015) [79] ont divisé la vraisemblance en deux parties, le processus de capture et le processus d'identification. En plus des histoires latentes d'erreurs, ils ont introduits des histoires latentes de capture $\boldsymbol{\xi}_k = (\xi_{k,1}, ..., \xi_{k,T})$. La fréquence de l'histoire de capture latente ξ_k est notée z_k, et le vecteur de toutes les fréquences de capture latentes est $\mathbf{z} = (z_1, ..., z_{2^T})$.

[74]: Link et al. (2010), 'Uncovering a Latent Multinomial'

[79]: Bonner et al. (2015), 'Extending the latent multinomial model with complex error processes and dynamic markov bases'

Dans la suite de ce document, $\boldsymbol{\theta}_1$ désigne les paramètres liés au processus de capture (ici N et $\mathbf{p}$) et $\boldsymbol{\theta}_2$ désigne les paramètres liés au processus d'identification (ici α). La vraisemblance jointe des données et données latentes est

$$[\mathbf{y}, \mathbf{x}, \mathbf{z}|\boldsymbol{\theta}_1, \boldsymbol{\theta}_2] = I(\mathbf{y} = \mathbf{A}\mathbf{x})[\mathbf{x}|\mathbf{z}, \boldsymbol{\theta}_2][\mathbf{z}|\boldsymbol{\theta}_1] \tag{5.1}$$

Je décrirais chacune des composantes de cette vraisemblance dans le modèle original. Pour décrire les extensions du modèle, je considérerais seulement les composantes qui changent et non plus l'ensemble du modèle .

- $I(\mathbf{y} = \mathbf{A}\mathbf{x})$

$I(\mathbf{z} = \mathbf{B}\mathbf{x})$ vaut 1 si $\mathbf{z} = \mathbf{B}\mathbf{x}$, et 0 sinon.

- $[\mathbf{x}|\mathbf{z}, \boldsymbol{\theta}_2]$

Ici, $\boldsymbol{\theta}_2 = \alpha$. Conditionnellement aux captures qui ont été réalisées et à la probabilité d'identification, la vraisemblance d'identification est

$$[\mathbf{x}|\mathbf{z}, \alpha] = I(\mathbf{z} = \mathbf{B}\mathbf{x}) \frac{\prod_k z_k!}{\prod_j x_j!} \prod_j \left[\prod_{t=1}^{T} \alpha^{I(\nu_{j,t}=1)} (1-\alpha)^{I(\nu_{j,t}=2)} \right]^{x_j} \tag{5.2}$$

- $[\mathbf{z}|\boldsymbol{\theta}_1]$

Ici, $\boldsymbol{\theta}_1 = (N, \mathbf{p})$. $\mathbf{p} = (p_1, ..., p_T)$ est le vecteur des probabilités de capture à chaque occasion. La vraisemblance de capture est celle du modèle classique en population fermée (M_t) :

$$[\mathbf{z}|N, \mathbf{p}] = \frac{N!}{\prod_k z_k!} \prod_k \pi_k^{z_k} \tag{5.3}$$

avec

$$\pi_k = \prod_{t=1}^{T} \left[p_t^{I(\xi_{k,t}=1)} (1-p_t)^{I(\xi_{k,t}=0)} \right] \tag{5.4}$$

Extension avec covariable de qualité d'identification (M_{t,α_n})

Le LMM est un modèle conçu pour des études utilisant de l'ADN. L'obtention de l'identité des individus se fait en génotypant les échantil-

lons d'ADN. Les erreurs d'identification sont dues à la qualité imparfaite des génotypes observés (principalement les loci manquants). La qualité de ces génotypages n'est pas utilisée dans le LMM malgré le fait qu'elle peut être évaluée. Utiliser une telle information pourrait permettre d'améliorer les estimations du modèle. Nous avons donc étendu le modèle pour prendre en compte une covariable de qualité d'identification.

Par rapport au modèle précédent, l'utilisation d'une covariable implique quelques changements. Comme la probabilité d'une identification correcte α n'est plus constante, nous devons travailler au niveau de l'échantillon et les fréquences $\mathbf{y}$, $\mathbf{x}$ et $\mathbf{z}$ ne sont plus des statistiques suffisantes. Nous utilisons donc les ensembles détaillés d'histoires $\mathbf{Y}$, $\mathbf{X}$ et $\mathbf{Z}$. Nous définissons $\alpha_{n,t}$ comme la probabilité d'identifier correctement l'individu n au moment t, pour un individu capturé à ce moment-là. Si n n'a pas été capturé au moment t, nous fixons $\alpha_{n,t} = 1$. Nous notons $\boldsymbol{\alpha} = (\alpha_{n,t})_{n\in[1,N],t\in[1,T]}$, le vecteur des probabilités d'identification associées à chaque capture réalisée. Afin de modéliser $\alpha_{n,t}$ en fonction d'une covariable, nous introduisons $\boldsymbol{\theta}_\alpha$, l'ensemble des paramètres définissant $\boldsymbol{\alpha}$. Nous écrivons $[\boldsymbol{\alpha}|\boldsymbol{\theta}_\alpha]$. Comme α dépend de n, nous appelons ce modèle Mt, α_n.

la vraisemblance s'écrit alors

$$[\mathbf{Y}, \mathbf{X}, \mathbf{Z}|N, \mathbf{p}, \boldsymbol{\alpha}, \boldsymbol{\theta}_\alpha] = I(\mathbf{Y}|\mathbf{X})[\mathbf{X}|\mathbf{Z}, \boldsymbol{\alpha}][\mathbf{Z}|N, \mathbf{p}][\boldsymbol{\alpha}|\boldsymbol{\theta}_\alpha]. \tag{5.5}$$

Nous modifions ensuite les différentes parties de la vraisemblance.

- $I(\mathbf{Y}|\mathbf{X})$

Définissons une fonction f telle que, pour une histoire d'erreurs latentes v_j, elle aboutisse à l'ensemble correspondant d'histoires observés (ω_i) : $f(v_j) = (\omega_i)$. Par exemple, $f((1,1,2)) = \{(1,1,0),(0,0,1)\}$. Si nous appliquons f à toutes les histoires latentes dans $\mathbf{X}$, l'ensemble d'histoires résultant (ω_i) doit être égal à $\mathbf{Y}$, à l'exception des inversions d'index. Ainsi, $I(\mathbf{Y}|\mathbf{X})$ vaut 1 si $f(\mathbf{X}) = \cup_j f(v_j) = \mathbf{Y}$ et 0 sinon. Nous l'écrivons sous la forme $I(\mathbf{Y} = f(\mathbf{X}))$.

- $[\mathbf{X}|\mathbf{Z}, \boldsymbol{\alpha}]$

Tout d'abord, nous réécrivons la partie $I(\mathbf{z} = \mathbf{B}\mathbf{x})$. Comme pour $I(\mathbf{Y}|\mathbf{X})$, définissons une fonction g qui, pour une histoire d'erreurs latentes v_j, aboutit à l'histoire de capture latente correspondant ξ_j : $g(v_j) = \xi_j$. Par exemple, $g((1,1,2)) = (1,1,1)$. Si nous appliquons g à toutes les histoires de $\mathbf{X}$, l'ensemble d'histoires résultant (ξ_j) doit être égal à $\mathbf{Z}$. Ainsi, $I(\mathbf{X}|\mathbf{Z})$ vaut 1 si $g(\mathbf{X}) = \cup_j g(v_j) = \mathbf{Z}$ et 0 sinon. Nous l'écrivons sous la forme $I(\mathbf{Z} = g(\mathbf{X}))$.

Chaque échantillon contribue à la vraisemblance par sa probabilité d'identification $\alpha_{n,t}$ si aucune erreur n'est commise ou par son complément en cas d'erreur.

$$[\mathbf{X}|\mathbf{Z}, \boldsymbol{\alpha}] = I(\mathbf{X}|\mathbf{Z}) \prod_{n=1}^{N} \prod_{t=1}^{T} \alpha_{n,t}^{I(v_{n,t}=1)} (1-\alpha_{n,t})^{I(v_{n,t}=2)} \; ; . \tag{5.6}$$

- $[\mathbf{Z}|N, \mathbf{p}]$

La vraisemblance de capture est un produit catégorielles donnant à chaque individu d'avoir une histoire bien définie :

$$[\mathbf{Z}|N, \mathbf{p}] = N! \prod_{n=1}^{N} \prod_{t=1}^{T} p_t^{I(\xi_{n,t}=1)} (1-p_t)^{I(\xi_{n,t}=0)}, \tag{5.7}$$

- $[\boldsymbol{\alpha} \mid \boldsymbol{\theta}]$

Pour cette partie, à l'instar de Mc clintock et. al (2014) [75], nous avons choisi de développer un modèle probit. D'autres liens pourraient être utilisés, d'autant plus qu'il n'y a pas de covariables manquantes. Le modèle probit nous donne

[75]: McClintock et al. (2014), 'Probit models for capture-recapture data subject to imperfect detection, individual heterogeneity and misidentification'

$$\alpha_{n,t} = \boldsymbol{\phi}(a \cdot \tau_{n,t} + b)$$

où $\boldsymbol{\phi}$ est la fonction de distribution cumulative normale standard. Ainsi, $\boldsymbol{\theta} = (a, b)$.

Pour spécifier complètement le modèle probit, nous définissons $u_{n,t}$ comme un indicateur binaire du succès de l'identification de la capture de l'individu n à l'occasion t. Autrement dit, $u_{n,t} = 1$ si l'échantillon n, t a donné lieu à une identification correcte de l'individu, et 0 sinon. Nous définissons également $\tilde{u}_{n,t}$, un processus latent continu de $u_{n,t}$. Nous fixons $\tilde{u}_{n,t} \sim \mathcal{N}(a\tau_{n,t} + b, 1)$ et si $\tilde{u}_{n,t} < 0$ alors $u_{n,t} = 0$, ou bien si $\tilde{u}_{n,t} > 0$ alors $u_{n,t} = 1$.

On note $\mathbf{u} = (u_{n,t})_{n\in[1,N],t\in[1,T]}$ et $\tilde{\mathbf{u}} = (\tilde{u}_{n,t})_{n\in[1,N],t\in[1,T]}$ Puisque toutes les covariables sont connues, conditionnellement à $\mathbf{X}$, tous les $u_{n,t}$ sont connus. La définition de $\tilde{u}_{n,t}$ n'est donc pas vraiment nécessaire, mais elle permet l'échantillonnage de Gibbs de a et b.

Extension pour capture multiple au cours d'une occasion ($M_{\lambda,\alpha}$)

Dans le cadre d'observations multiples, les histoires de capture peuvent contenir n'importe quels nombres, indiquant combien de fois l'individu a été capturé à chaque fois. Nous ne pouvons donc pas utiliser "2" pour indiquer une erreur d'identification. En outre, une histoire peut contenir plusieurs erreurs d'identification pour la même occasion. Nous commençons donc par modifier les notations utilisées pour les erreurs d'identification. Pour représenter les histoires d'erreurs latentes ν_j, nous notons le nombre total d'observations d'un individu à chaque occasion avec, en exposant, le nombre de ces observations qui ont donné lieu à une erreur d'identification. Par exemple, l'histoire observée $(0, 2, 0, 3, 1)$ peut avoir été générée par l'histoire d'erreurs latentes $(1^{(1)}, 2, 0, 3, 3^{(2)})$. Dans cet exemple, l'observation à la première occasion et deux observations à la cinquième occasion ont été mal identifiées, ce qui a donné zéro observation à l'occasion 1 et une observation à l'occasion 5. L'histoire latente de capture ξ_k est la même que l'histoire latente d'erreurs sans les exposants. Dans notre exemple, l'histoire latente de capture est $(1, 2, 0, 3, 3)$.

Soit $\boldsymbol{\lambda} = (\lambda_1, ..., \lambda_T)$ l'ensemble des paramètres intervenant dans le processus de capture, modélisé par un processus de Poisson (chaque λ_t étant

le paramètre du processus de Poisson à l'occasion t). La vraisemblance du modèle est

$$[\mathbf{y}, \mathbf{x}, \mathbf{z}|N, \boldsymbol{\lambda}, \alpha] = I(\mathbf{y} = \mathbf{A}\mathbf{x})\,[\mathbf{x}|\mathbf{z}, \alpha]\,[\mathbf{z}|N, \boldsymbol{\lambda}] \tag{5.8}$$

Nous décrivons maintenant les deux éléments qui changent par rapport au modèle $M_{t,\alpha}$, à savoir $[\mathbf{z}|N, \boldsymbol{\lambda}]$ et $[\mathbf{x}|\mathbf{z}, \alpha]$.

- $[\mathbf{z}|N, \boldsymbol{\lambda}]$

Nous modélisons le nombre réel d'observations d'un individu à une occasion par une distribution de Poisson. La probabilité π_k qu'un individu ait une histoire de capture latent donné ξ_k est donnée par :

$$\pi_k = \prod_{t=1}^{T} \frac{\lambda_t^{\xi_{k,t}}}{\xi_{k,t}!} e^{-\lambda_t} \tag{5.9}$$

La vraisemblance de capture a une forme multinomiale :

$$[\mathbf{z}|N, \boldsymbol{\lambda}] = \frac{N!}{\prod_k z_k!} \prod_k \pi_k^{z_k} \tag{5.10}$$

- $[\mathbf{x}|\mathbf{z}, \alpha]$

Toutes les captures réalisées sont potentiellement sujettes à des erreurs d'identification. Soit $o_{j,t}$ le nombre de bonnes identifications pour les individus ayant une histoire v_j à l'occasion t. Alors, connaissant le nombre réel de captures, la probabilité qu'il ait été correctement identifié $o_{j,t}$ fois est Binomiale. Ainsi :

$$[\mathbf{x}|\mathbf{z}, \alpha] = \prod_j \prod_{t=1}^{T} \binom{v_{j,t}}{o_{j,t}} o_{j,t}^{\alpha} (v_{j,t} - o_{j,t})^{1-\alpha} \tag{5.11}$$

Population ouverte

Cette section je décrit les principaux développements en population ouverte réalisés au cours de cette thèse. Par souci de légèreté, cette section se contente de lister les modèles développés et implémentés, en décrivant quelle partie du modèle change par rapport au modèle LMM. J'encourage le lecteur à se reporter au Chapitre Chapter 4 pour plus de détails.

Le LMM a d'abord été étendu pour estimer la survie en population ouverte avec un seul état. Pour ce modèle, la vraisemblance de capture $[\mathbf{z}|\boldsymbol{\theta}_1]$ est celle du modèle de Cormack-Jolly-Seber (CJS [27–29]). Nous le notons CJS_α. Les paramètres de capture sont $\boldsymbol{\theta}_1 = (\mathbf{p}, \phi)$, où $\mathbf{p}$ est le vecteur des probabilités de capture par occasion et ϕ est la probabilité de survie supposée constante des individus d'une occasion à la suivante. Le modèle pourrait facilement être modifié pour tenir compte d'une éventuelle variabilité temporelle de la survie. La vraisemblance d'identification est aussi supposée constante, comme pour le modèle $M_{t,\alpha}$.

[27]: Cormack (1964), 'Estimates of Survival from the Sighting of Marked Animals'
[28]: Jolly (1965), 'Explicit estimates from capture-recapture data with both death and immigration-stochastic model'
[29]: Seber (1965), 'A note on the multiple-recapture census'

Le modèle CJS_α a ensuite été étendu pour tenir compte d'une covariable de qualité d'identification comme dans le modèle probit présenté

précédemment. Pour ce modèle noté CJS_{α_n}, la vraisemblance de capture $[\mathbf{z}|\theta_1]$ est celle du CJS ; et la vraisemblance d'identification $[\mathbf{x}|\mathbf{z}, \theta_2]$ est celle du modèle M_{t,α_n}.

Enfin, le LMM a été étendu au cas multi-état en population ouverte. Pour ce modèle, les paramètres du processus d'observation sont $\boldsymbol{\theta}_1 = (\mathbf{p}, \phi, \psi)$, où $\mathbf{p}$ est le vecteur des probabilités de capture par occasion, ϕ est la probabilité de survie supposée constante des individus d'une occasion à la suivante, et ψest la matrice de probabilité de transition des états aux autres. La vraisemblance de capture est celle du modèle d'Arnason-Schwarz (AS [30, 31]). La vraisemblance d'identification est la même que celle du modèle $M_{t\alpha}$.

[30]: Neil Arnason (1972), 'Parameter estimates from mark-recapture experiments on two populations subject to migration and death'
[31]: Neil Arnason (1973), 'The estimation of population size, migration rates and survival in a stratified population'

Estimation des paramètres des modèles

L'estimation des paramètres du modèle ne peut pas être faite par maximum de vraisemblance car il faudrait sommer la vraisemblance jointe $[\mathbf{y}, \mathbf{x}, \mathbf{z}|\boldsymbol{\theta}_1, \boldsymbol{\theta}_2]$ sur l'ensemble de $\mathbf{x}$ possible. Cet ensemble est trop complexe pour être énuméré. Link et al. (2010) [74], et Bonner et al. (2015) [79] montrent comment estimer les paramètres grâce à un algorithme de Monte Carlo Markov Chain (MCMC). Dans ce MCMC, un ensemble latent $\mathbf{x}$ est proposé à chaque itération et accepté ou refusé à travers une procédure de Métropolis-Hastings (MH).

[74]: Link et al. (2010), 'Uncovering a Latent Multinomial'
[79]: Bonner et al. (2015), 'Extending the latent multinomial model with complex error processes and dynamic markov bases'

Pour chaque modèle, j'ai proposé un algorithme de MCMC tel que la proposition de l'ensemble latent $\mathbf{x}$ permette d'accéder à tout ensemble d'histoires latentes possible. Spécifiquement, pour le modèle probit intégrant la covariable de qualité d'identification, l'algorithme a notamment été optimisé pour proposer des ensembles latents en fonction de la qualité d'identification des histoires susceptibles d'être des erreurs. Pour le modèle de Poisson (observations répétées), l'algorithme a été notamment modifié de sorte de pouvoir proposer des histoires latentes où plusieurs erreurs d'identification ont pu être faites pour un même individu à une même occasion. Le ratio de Metropolis-Hasting a aussi dû être calculé de manière différente pour chaque modèle.

Code

Tous les modèles présentés dans cette thèse ont été implémentés avec le langage R, en utilisant la bibliothèque NIMBLE [80]. Le choix du langage R permet de rendre accessible les codes développés à une très large proportion de la communauté utilisant la CR car ce langage y est largement maîtrisé. La bibliothèque NIMBLE permet de coder l'ensemble des algorithmes MCMC et de compiler le code en C++ pour le rendre rapide à exécuter.

[80]: Valpine et al. (2017), 'Programming With Models'

Le développement des codes a été conséquent mais ouvre aujourd'hui une porte d'accès au modèle qui n'existait pas jusque-là. Les codes développés peuvent permettre à des chercheurs de faire tourner facilement les modèles à un état et constituent une bonne base pour le développement d'une bibliothèque dans le langage R qui permettrait une large diffusion du LMM.

Validation des modèles par simulations

Nous avons réalisé plusieurs études à l'aide de simulations afin de tester et comparer les différents modèles disponibles.

Réplication des résultats de Vale et al. (2014)

La première étude a répliqué les résultats de Vale et al. (2014) [77]. Dans cette étude, les auteurs montrent que le modèle $M_{t,\alpha}$ produit des estimations très biaisées si le taux de capture est faible ($p \leq 0.1$). En plus de reproduire les résultats, j'ai ajouté des simulations avec une probabilité d'identification plus faible que ce qui avait été testé ($\alpha = 0.8$).

[77]: Vale et al. (2014), 'Maximum likelihood estimation for model $M_{t,\alpha}$ for capture–recapture data with misidentification'

Population fermée, pas de capture répétées, estimation de la taille de population

Nous avons simulées des données suivant le modèle M_{t,α_n}. Nous avons ensuite estimé la taille des populations à partir des simulations à l'aide des modèles suivants:

- M_t: modèle classique en population fermée,
- $M_{t,\alpha}$: LMM publié par Link et al. (2010) [74],
- M_{t,α_n}: LMM avec une covariable de qualité d'identification,
- Yoshizaki: modèle proposé par Yoshizaki et al. (2011) [67].

Nous rappelons que le modèle suggéré par Yoshizaki demande à retirer les histoires ne contenant qu'une unique capture, et donc susceptibles d'être de faux individus.

Nous avons simulé des données pour les valeurs de paramètres suivantes: $N = 500$, $p = 0.1, 0.2, 0.3, 0.4$, $\bar{\alpha} = 0.8, 0.9, 0.95$ et $T = 5, 7, 9$. Les valeurs des paramètres a et b du modèle probit sont calculée automatiquement en fonction de la probabilité d'identification moyenne $\bar{\alpha}$ voulue. En tout 36 scénarios sont testés (12 scénarios avec différentes combinaisons de probabilités de capture et d'identification pour 3 nombres d'occasions différents).

Les résultats montrent que le modèle M_t surestime toujours la taille de la population, jusqu'à 1.5 fois (pour $\alpha = 0.8$).
Pour les scénarios avec $p \geq 0.3$, tous les autres modèles estiment très bien la taille de la population. Le plus précis est le modèle M_{t,α_n}, les intervalles de confiances estimés sont plus faibles et contiennent plus souvent la vraie valeur de la population en moyenne. Le modèle de Yoshizaki estime des intervalles de confiances plus étroits que le modèle $M_{t,\alpha}$ mais qui contiennent moins souvent la vraie valeur.
Pour les scénarios avec $p = 0.2$, les trois modèles intégrant les erreurs d'identification sont biaisées si la probabilité d'identification et le nombre d'occasions sont faibles ($\alpha = 0.8$ et $T = 5$). Si la probabilité d'identification ou le nombre d'occasions est suffisamment élevé, alors ces modèles produisent des estimations non biaisées. Le modèle M_{t,α_n} est celui qui produit des estimations non biaisées dans le plus grand nombre de scénarios pour $p = 0.2$ (tous sauf le scénario $\alpha = 0.8 \& T = 5$).
Pour les scénarios où $p = 0.1$, les modèles $M_{t,\alpha}$ et Yoshizaki sont toujours biaisées. Ils sous estiment la taille de la population jusqu'à 50%. Le modèle

M_{t,α_n} estime correctement la taille de la population quand $\alpha = 0.95$. Pour les autres scénarios, il est le modèle le moins biaisé. Lorsque $\alpha = 0.9$ et $T \geq 7$, en moyenne le modèle M_{t,α_n} surestime la taille de population mais pas systématiquement. Et la vraie taille de population est comprise dans l'intervalle de confiance pour plus de 90% des simulations.

Population fermée, capture répétées, estimation de la taille de population

Nous avons simulé des données selon le modèle $M_{\lambda,\alpha}$. Nous avons ensuite utilisé les trois modèles suivants pour estimer la taille de population:

- M_t: Modèle classique en population fermée,
- $M_{\lambda,\alpha}$: LMM intégrant les captures multiples au cours d'une occasion,
- Yoshizaki: modèle proposé par Yoshizaki et al. (2011) [67].

[67]: Yoshizaki et al. (2011), 'Modeling misidentification errors that result from use of genetic tags in capture–recapture studies'

Les scénarios simulés sont les mêmes que dans la section précédente. Les λ simulés sont ceux qui permettent d'observer au moins une fois un pourcentage de la population proche du taux de capture voulu. Pour simuler une probabilité de capture $p = 0.4$, j'ai choisi un λ qui conduira 40% de la population à être observé au moins une fois ainsi $\lambda = 0.11, 0.23, 0.36, 0.51$.

Les résultats montrent que le modèle M_t surestime la taille de la population plus encore que lorsque les individus ne peuvent être vus qu'une seule fois et jusqu'à 2 fois la taille réelle de la population.
Pour les deux autres modèles, dans les scénarios où $\lambda \geq 0.23$ (correspondant à $p \geq 0.2$), un faible biais est observé quand $T = 5$. Lorsque $T \geq 7$, l'estimation de taille de population n'est pas biaisée.
Enfin, quand $\lambda = 0.11$ (correspond à $p = 0.1$), la taille de la population est systématiquement sous-estimée (jusqu'à être sous-estimée de moitié).

Population fermée, estimation des transitions

Une étude en multi-état et population fermée a porté sur l'impact des erreurs d'identification sur l'estimation des probabilités de transition d'un état à un autre. Nous avons montré que si la probabilité d'identification ne dépend ni de l'état ni de l'occasion de capture, alors un modèle classique estime les probabilités de transitions sans biais. Si la probabilité d'identification dépend de l'état ou de l'occasion de capture, les estimations du modèle classique sont biaisées, et à plus forte raison si la probabilité d'identification dépend à la fois du stade et de l'occasion de capture. Dans chaque situation, le LMM permet d'estimer correctement les probabilités de transition.

Population ouverte, estimation de la survie

Une autre courte étude a porté sur l'estimation de la probabilité de survie par le modèle CJS_α, ainsi que l'importance de ne pas ignorer les erreurs d'identification. Nous avons montré que le CJS classique sous-estime la survie en présence d'erreurs d'identification et qu'une

survie relativement faible ($\phi = 0.6$) peut être estimée correctement par le modèle CJS_α, même avec une probabilité de capture de seulement 0.3.

Étude de la taille de la population de la loutre européenne

Nous avons appliqué les modèles M_t, $M_{\lambda,\alpha}$, et le modèle de Yoshizaki à un jeu de données d'une étude sur la loutre européenne (*Lutra lutra*). La collecte des données a consisté à recueillir des fèces de loutre (épreintes) sur cinq jours consécutifs ($T = 5$) lors de 6 années, de 2006 à 2012 (à l'exception de 2009).

Les auteurs ont estimé qu'il était peu probable que la PCR répétée puisse éliminer complètement toutes les erreurs de génotypage en raison des taux d'erreur de génotypage et de loci non-observés relativement élevés. Ils ont utilisé le modèle d'erreur d'identification proposé par [71] (ci-après modèle L&B), mis en œuvre dans le programme MARK [95]. Cependant le modèle L&B ne prend pas correctement en compte les erreurs d'identification car il ignore la dépendance complète entre la paire d'histoires créées chaque fois qu'un génotype est incorrectement identifié.

[71]: Lukacs et al. (2005), 'Estimating population size from DNA-based closed capture-recapture data incorporating genotyping error'

[95]: White et al. (1999), 'Program MARK'

L'utilisation des 3 modèles (M_t, $M_{\lambda,\alpha}$ et celui de Yoshizaki), révèle que les jeux de données de 4 années sur 6 contenaient des erreurs d'identification. Pour ces années-là, le modèle $M_{\lambda,\alpha}$ et celui de Yoshizaki donnent les mêmes estimations de taille de population. Mais l'écart-type estimé de la taille de la population par le modèle $M_{\lambda,\alpha}$ est plus faible que celui estimé par le modèle de Yoshizaki. Pour les deux autres années, le modèle M_t et celui de Yoshizaki donnent les mêmes estimations. Le modèle $M_{\lambda,\alpha}$ donne une estimation plus faible de la taille de population comparée aux autres modèles.

Étude de deux protocoles pour l'estimation de la survie des larves de moustique

Les moustiques transmettent certaines des plus importantes maladies infectieuses de l'homme. Le paludisme, la filariose bancroftienne et les virus tels que la dengue, le chikungunya, le Zika ou la fièvre jaune continuent de poser des défis majeurs à la santé publique. Ces défis sont encore aggravés par les changements globaux de l'environnement et de la société, qui favorisent l'émergence et la résurgence de ces maladies dans le monde entier.

Ce n'est qu'en étudiant la manière dont les transformations globales influencent les traits de leur histoire de vie que nous pourrons anticiper plus précisément les conséquences épidémiologiques de l'expansion de la niche des moustiques vecteurs et limiter la prolifération des maladies qu'ils transmettent.

Afin d'étudier en particulier la survie des larves de moustiques dans des environnements urbains, une équipe de l'UMR MIVEGEC a pensé une expérience de terrain pour le suivi individuel des larves de moustiques.

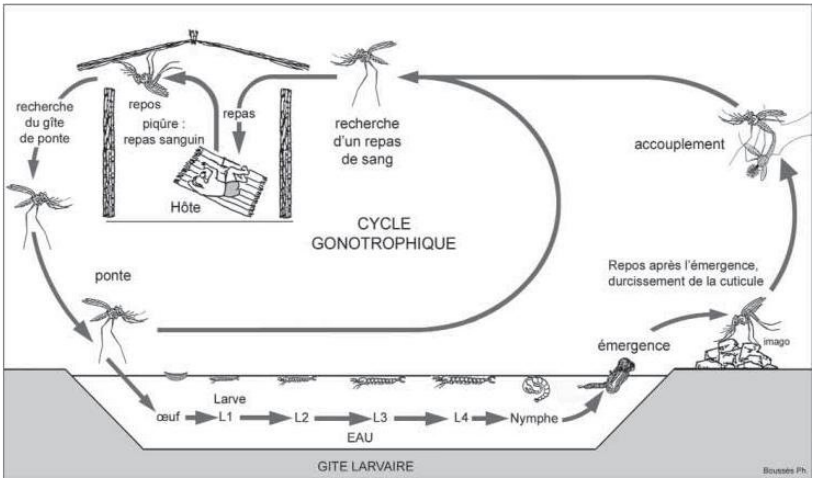

Figure 5.1: Cycle biologique de l'Anophele (Boussès Ph.). Les larves passent par 4 stades larvaires et un stade de nymphe avant de devenir adulte.

Un organisme comme le moustique ne peut faire l'objet d'un marquage traditionnel de part sa petite taille. Le cycle de vie du moustique est montré sur la Figure 5.1. L'idée du protocole est de capturer des larves de moustique et de les isoler dans de l'eau sans contaminant ADN. Ensuite, les larves sont relâchées, l'ADN est extrait de l'eau et les larves sont identifiées individuellement sur la base de cet ADN. La quantité et la qualité de l'ADN ainsi extrait étant faible, des erreurs d'identification sont donc à prévoir.

Sur la base de l'idée qu'ils ont développée, j'ai proposé deux protocoles de terrain distincts.
1) Le premier consiste à capturer les larves aux stades 3 et 4 ainsi que les nymphes. Les nymphes sont tuées car la probabilité de les ré-attraper si elles sont relâchées est très faible. On suppose qu'on ne fait pas d'erreur d'identification du stade nymphe car l'on dispose de beaucoup d'ADN pour identifier individuellement les nymphes. Par contre on ne pourra pas estimer la survie des nymphes puisqu'elles sont tués.
2) Le second protocole consiste à poser un filet au dessus du gîte larvaire. Ainsi on capture tous les adultes (avec une probabilité supposée de 1) en plus des larves aux stades 3-4 et des nymphes. Ce protocole permet de relâcher les nymphes et d'estimer leur survie.

J'ai simulé 100 jeux de données selon chaque protocole. La taille des populations simulées varient aléatoirement entre 500 et 1000 larves tous stades confondus (du stade 1 à la nymphe), afin de correspondre à de petits gîtes larvaires tels qu'un pneu abandonné. La probabilité d'identification simulée est de 0.9 et 100 individus sont capturés à chaque occasion. Les valeurs de survie et de transition utilisées sont empruntées à la littérature afin que les simulations soient le plus proche possible de la réalité.

L'ensemble des simulations a été analysé avec une version multi-état du LMM tenant compte des spécificités des protocoles. Les résultats montrent que la survie des larves aux stades 3 et 4 est estimée sans biais pour les deux protocoles. En revanche contrairement au premier protocole, la survie des nymphes est aussi estimée assez précisément pour le second protocole. Le second protocole permet d'estimer correctement toutes les probabilités de transition d'un stade à un autre du stade 3 à l'adulte. En comparaison, le premier protocole ne permet pas l'estimation

des probabilités de transition de la nymphe à l'adulte et les estimations de transitions qu'il fait sont biaisées.

Cette étude, même si elle ne se base que sur des simulations, est une première tentative pour définir un protocole de CR sur des larves de moustiques. A l'aide de cette étude, je montre que le LMM peut être utilisé pour des situations complexes qui ne pourraient pas être étudiées sans modéliser les probabilités d'identification. D'autres espèces pourraient ainsi être étudiées à l'aide d'un protocole précis, évalué par simulation et d'une des extensions du LMM.

Conclusion

Dans cette thèse, j'ai abordé de nombreux modèles de CR en présence d'erreurs d'identifications individuelles, allant jusqu'à définir un LMM multi-état en population ouverte.

En premier, j'ai développé des modèles visant à estimer la taille de la population en population fermé. J'ai commencé par étendre le LMM en population fermé à plusieurs états. Puis j'ai étendu le modèle LMM pour prendre en compte une covariable de qualité d'identification. Ce modèle permet notamment d'améliorer les estimations et d'estimer les paramètres dans des cas de probabilité de capture faible. Ensuite j'ai étendu le LMM au cas où les individus peuvent êtres capturés plusieurs fois au cours d'une même occasion. Ce modèle offre de nouvelles perspectives dans la mesure où de nombreuses études obtiennent les données en prélevant des fèces. Ces études prélèvent de l'ADN de faible qualité avec plusieurs échantillons d'un même individu.

Ensuite j'ai étendu le LMM aux populations ouvertes pour estimer la survie des individus. En particulier, j'ai développé le modèle à un seul état, avec ou sans covariable de qualité d'identification. Puis j'ai développé le modèle en multi-état, ouvrant la possibilité d'estimer des probabilités de transition entre états conditionnellement à la survie.

Enfin, j'ai décris comment de la CR pourrait être réalisée sur des larves de moustique. Pour cela je propose deux protocoles de terrain différents et je les compare à l'aide de simulations. Je montre, d'abord, que la survie peut être estimée correctement. Ensuite, je montre qu'un protocole est meilleur que l'autre, mais un peu plus contraignant.

Le LMM est une généralisation des modèles CR classiques qui utilise des histoires latentes pour décrire les processus cachés. Toutes les extensions développées dans cette thèse et proposées ici sont des étapes vers l'exhaustivité du cadre de la CR. Lorsque la plupart des cas imaginables sont modélisés, les limites de la recherche sont repoussées vers les limites de la faisabilité technique et de l'imagination. J'espère que ce travail aura contribué à cet objectif.

Appendix

A

Metropolis-Hastings ratio simplification

]

The ratio r_1 given in Section 1.2.4 is

$$r_1 = min\left(1, \frac{[\mathbf{y}, \mathbf{x}', \mathbf{z}'|N', p, \alpha]}{[\mathbf{y}, \mathbf{x}^{(k-1)}, \mathbf{z}^{(k-1)}|N, p, \alpha]} \frac{q(\mathbf{x}^{(k-1)}|\mathbf{x}')}{q(\mathbf{x}'|\mathbf{x}^{(k-1)})}\right) \tag{A.1}$$

The first fraction of Equation A.1 can be simplified in as the product of two fractions:

$$\begin{aligned}
&\frac{[\mathbf{x}'|\mathbf{z}', \alpha]}{[\mathbf{x}^{(k-1)}|\mathbf{z}^{(k-1)}, \alpha]} \frac{[\mathbf{z}'|N', \mathbf{p}]}{[\mathbf{z}^{(k-1)}|N, \mathbf{p}]} \\
=&\frac{\prod_k z'_k! / \prod_j x'_j! \prod_j \left[\prod_{t=1}^T A_{j,t}\right]^{x'_j}}{\prod_k z_k! / \prod_j x_j! \prod_j \left[\prod_{t=1}^T A_{j,t}\right]^{x_j}} \frac{N'! / \prod_k z'_k! \prod_k \pi_k^{z'_k}}{N! / \prod_k z_k! \prod_k \pi_k^{z_k}} \\
=&\frac{\prod_j x_j! \prod_j \left[\prod_{t=1}^T A_{j,t}\right]^{x'_j}}{\prod_j x'_j! \prod_j \left[\prod_{t=1}^T A_{j,t}\right]^{x_j}} \frac{N'! \prod_k \pi_k^{z'_k}}{N! \prod_k \pi_k^{z_k}}.
\end{aligned} \tag{A.2}$$

The only difference in the nominators and denominators are: the values of x_j and x'_j for the histories v_j whose counts have been changed (by the proposal algorithm), and the population size N and N'.
And since we only add or remove one misidentification at the time, the difference between x_j and x'_j, and between N and N' is always 1 (or -1).

When we add a misidentification, we remove 1 from x_j for the histories v_0 and v_1 that were sampled, and we add 1 to x_j for the defined history $v2$ (see Section 1.2.4). Thus,

$$\begin{aligned}
&\frac{[\mathbf{x}'|\mathbf{z}', \alpha]}{[\mathbf{x}^{(k-1)}|\mathbf{z}^{(k-1)}, \alpha]} \frac{[\mathbf{z}'|N', \mathbf{p}]}{[\mathbf{z}^{(k-1)}|N, \mathbf{p}]} \\
=&\frac{x_{v_0} x_{v_1} \prod_{t=1}^T A_{v_2,t}}{x'_{v_2} \prod_{t=1}^T (A_{v_0,t} A_{v_1,t})} \frac{\pi_{v_2}}{N \pi_{v_0} \pi_{v_1}} \\
=&\frac{x_{v_0} x_{v_1} (1-\alpha)}{x'_{v_2} \alpha} \frac{1}{N \prod_{t=1}^T p_t}
\end{aligned} \tag{A.3}$$

When we remove a misidentification, we can show that .

$$\frac{[\mathbf{x}'|\mathbf{z}',\alpha]}{[\mathbf{x}^{(k-1)}|\mathbf{z}^{(k-1)},\alpha]}\frac{[\mathbf{z}'|N',\mathbf{p}]}{[\mathbf{z}^{(k-1)}|N,\mathbf{p}]} = \frac{x_{\nu_2}\alpha}{x'_{\nu_0}x'_{\nu_1}(1-\alpha)}N'\prod_{t=1}^{T}p_t \tag{A.4}$$

B

Forward-backward algorithm

B.1 The forward-backward algorithm

The forward-backward algorithm is an algorithm used to compute posterior marginal probabilities of hidden states. I present the algorithm to compute the probability that an individual was in a state s given it's history, in a closed population experiment.

Say that an individual can be in S different states, and during the experiment it was in states $\mathbf{z} = (z_1, ..., z_T) \in \{1, ..., S\}^T$. This individual had the latent capture history $\boldsymbol{\omega} = (\omega_1, ..., \omega_T)$.

We want to compute the probabilities $P(z_t = s|\theta, \boldsymbol{\omega})$ for $t = 1, ..., T$ and $s = 1, ..., S$. The algorithm proceeds as follows.

1) First, the "forward probabilities" are computed. For each occasion $t = 1, ..., T$, it is the probability $P(z_t = s, \boldsymbol{\omega}_{1,...,t}) = \alpha_t(s)$ that the individual was in state s at t and that the history $\boldsymbol{\omega}_{1,...,t}$ was observed.

$$\begin{cases} \alpha_1(s) & = \delta_s p_{1,s} \\ \alpha_t(s) & = \sum_{r=1}^{S} \left[\alpha_{t-1}(r) \psi_{r,s} p_{r,t}^{I(\omega_t = r)} (1 - p_{r,t})^{I(\omega_t = 0)} 0^{I(\omega_t \notin \{0,r\})} \right] \quad \forall t = 2, ..., T \end{cases}$$

Notation reminder

- $p_{s,t}$: probaility of observing an individual in state s at occasion t (in the manuscript we kept p state independant),
- δ_s: probability that an individual is in state s at occasion 1,
- $\psi_{r,s}$: probability that an individual transition from state s to state r between two occasions.**x**.

2) Second, the "backward probabilities" are computed. For each occasion $t = 1, ..., T - 1$, it is the probability $P(\boldsymbol{\omega}_{t+1,...,T}) = \beta_t(s)$ that the history $\boldsymbol{\omega}_{t+1,...,T}$ was observed, knowing that the individual was in state s at occasion t.

$$\begin{cases} \beta_T(r) & = 1 \\ \beta_t(r) & = \sum_{s=1}^{S} \left[\psi_{r,s} p_{s,t}^{I(\omega_t = s)} (1 - p_{s,t})^{I(\omega_t = 0)} 0^{I(\omega_t \notin \{0,s\})} \beta_{t+1} \right] \quad \forall t = 1, ..., T - 1 \end{cases}$$

3) Lastly, the forward and backward probabilities are multiplied to give the marginal posterior probabilities.

$$P(z_t = s|\theta, \boldsymbol{\omega}) = \alpha_t(s)\beta_t(s)$$

In open population, an additional state "dead" must be considered and the probability that the individual survived or transitioned from state "alive" to "dead" must be added.

B.2 Probabilities of transition

The posterior probability that an individual transitioned from state r at time t to state s at time $t + 1$ can be calculated using the forward

probabilities $\alpha_t(r)$ and the backward probabilities $\beta_t(s)$.

Let τ_t be the realised transition between occasions t and $t+1$.

$$P(\tau_t = (r,s)|\boldsymbol{\omega}, \theta) = \alpha_t(r)\psi r, s p_{s,t+1}^{I(\omega_t=s)}(1-p_{s,t+1})^{I(\omega_t=0)}0^{I(\omega_t \notin \{0,s\})}\beta_{t+1}(s)$$

C

Single state models confidence interval

T	p	M_t	Yoshizaki	$M_{t,\alpha}$	M_{t,α_n}
5	0.1	62	61	15	90
5	0.2	16	92.3	82	93.3
5	0.3	2.33	90.7	92	96
5	0.4	0	93	97	94.7
7	0.1	38	67	26	92.3
7	0.2	0.667	90.7	86.3	93
7	0.3	0	90	92.3	94.7
7	0.4	0	76.7	89.3	93
9	0.1	15.3	71.3	41.7	92.3
9	0.2	0	89.3	88.7	94.7
9	0.3	0	90	96	96.7
9	0.4	0	77.3	94.3	94.3

Table C.1: Percent of simulations for which the various models' 95% confidence interval contain the true population size.

D

NIMBLE code for Yoshizaki's model

- Nimble code:

```
yoshiCode <- nimbleCode({

  # priors
  for(t in 1:S){
    capture[t] ~ dbeta(1.0, 1.0)
  }

  tau[S+1] <- prod(1 - capture[1:S])
  for(t in 1:S) tau[t] <- capture[t] * prod(1 - capture[1:S]) / (1-capture[t])

  tauStar <- 1 - sum(tau[1:(S+1)])
  N <- sum(x[1:nbLatentObs]) / tauStar

  # Likelihood, x
  x[1:nbLatentObs] ~ dYoshi(capture[1:S],  tauStar,
                            latentObservation[1:nbLatentObs, 1:S],
                            latentIndex[1:nbLatentObs], S)

})
```

- Nimble Distribution:

```
dYoshi <- nimbleFunction(
  run = function(x = double(1),
                 capture = double(1),
                 tauStar = double(0),
                 latentObservation = double(2),
                 latentIndex = double(1),
                 S = double(0),
                 log = integer(0, default = 1)
  ) {
    logProbData <- lfactorial(sum(x))
    indexs <- which(x > 0)

    for(i in 1:length(indexs)){
      I <- indexs[i]
      hist <- latentObservation[I,]
      logProbData <- logProbData - lfactorial(x[I])
      probHist <- 1
      for(t in 1:S){
        if(hist[t] == 0)
          probHist <- probHist * (1-capture[t])
        else if(hist[t] == 1)
          probHist <- probHist * capture[t]
      }
      logProbData <- logProbData + x[I] * log(probHist / tauStar)
    }

    if(log) return(logProbData)
    return(exp(logProbData))
    returnType(double(0))
  })
```

Bibliography

[1] Rodolfo Dirzo et al. 'Defaunation in the Anthropocene'. In: *Science* 345.6195 (July 2014). Publisher: American Association for the Advancement of Science, pp. 401–406. DOI: 10.1126/science.1251817 (cited on page 2).

[2] Douglas J. McCauley et al. 'Marine defaunation: Animal loss in the global ocean'. In: *Science* 347.6219 (Jan. 2015). Publisher: American Association for the Advancement of Science, p. 1255641. DOI: 10.1126/science.1255641 (cited on page 2).

[3] Jonathan L. Payne et al. 'Ecological selectivity of the emerging mass extinction in the oceans'. In: *Science* 353.6305 (Sept. 2016). Publisher: American Association for the Advancement of Science, pp. 1284–1286. DOI: 10.1126/science.aaf2416 (cited on page 2).

[4] Jeremy B. C. Jackson et al. 'Historical overfishing and the recent collapse of coastal ecosystems'. In: *Science* 293.5530 (July 2001). Publisher: American Association for the Advancement of Science, pp. 629–637. DOI: 10.1126/science.1059199 (cited on page 2).

[5] Xingli Giam. 'Global biodiversity loss from tropical deforestation'. In: *Proceedings of the National Academy of Sciences* 114.23 (June 2017). Publisher: Proceedings of the National Academy of Sciences, pp. 5775–5777. DOI: 10.1073/pnas.1706264114 (cited on page 2).

[6] J. F. Walsh, D. H. Molyneux, and M. H. Birley. 'Deforestation: effects on vector-borne disease'. eng. In: *Parasitology* 106 Suppl (1993), S55–75. DOI: 10.1017/s0031182000086121 (cited on page 2).

[7] Jeffrey A. McNeely. 'The sinking ark: Pollution and the worldwide loss of biodiversity'. en. In: *Biodiversity & Conservation* 1.1 (Mar. 1992), pp. 2–18. DOI: 10.1007/BF00700247 (cited on page 2).

[8] Sabrina J. Lovell, Susan F. Stone, and Linda Fernandez. 'The economic impacts of aquatic invasive species: a review of the literature'. en. In: *Agricultural and Resource Economics Review* 35.1 (Apr. 2006). Publisher: Cambridge University Press, pp. 195–208. DOI: 10.1017/S1068280500010157 (cited on page 2).

[9] Belinda Gallardo et al. 'Global ecological impacts of invasive species in aquatic ecosystems'. en. In: *Global Change Biology* 22.1 (2016). _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/gcb.13004, pp. 151–163. DOI: 10.1111/gcb.13004 (cited on page 2).

[10] Oksana Grente et al. 'Wolf depredation hotspots in France: Clustering analyses adjusting for livestock availability'. In: *Biological Conservation* 267 (Mar. 2022), p. 109495. DOI: 10.1016/j.biocon.2022.109495 (cited on page 2).

[11] L. Jen Shaffer et al. 'Human-Elephant Conflict: A Review of Current Management Strategies and Future Directions'. In: *Frontiers in Ecology and Evolution* 6 (2019) (cited on page 2).

[12] William J. Sutherland et al. 'The need for evidence-based conservation'. In: *Trends in Ecology & Evolution* 19.6 (June 2004), pp. 305–308. DOI: 10.1016/j.tree.2004.03.018 (cited on page 2).

[13] Caroline Moussy et al. 'A quantitative global review of species population monitoring'. en. In: *Conservation Biology* 36.1 (2022). _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/cobi.13721, e13721. DOI: 10.1111/cobi.13721 (cited on page 2).

[14] James D. Nichols and Byron K. Williams. 'Monitoring for conservation'. In: *Trends in Ecology & Evolution* 21.12 (Dec. 2006), pp. 668–673. DOI: 10.1016/j.tree.2006.08.007 (cited on page 2).

[15] Jan Perret et al. 'Plants stand still but hide: Imperfect and heterogeneous detection is the rule when counting plants'. en. In: *Journal of Ecology* 111.7 (2023), pp. 1483–1496. DOI: 10.1111/1365-2745.14110 (cited on page 2).

[16] Sarah Cubaynes et al. 'Importance of accounting for detection heterogeneity when estimating abundance: the case of french wolves'. en. In: *Conservation Biology* 24.2 (2010), pp. 621–626. DOI: 10.1111/j.1523-1739.2009.01431.x (cited on page 2).

[17] Lise Comte and Gaël Grenouillet. 'Species distribution modelling and imperfect detection: comparing occupancy versus consensus methods'. en. In: *Diversity and Distributions* 19.8 (2013). _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/ddi.12078, pp. 996–1007. DOI: 10.1111/ddi.12078 (cited on page 2).

[18] Nigel G. Yoccoz, James D. Nichols, and Thierry Boulinier. 'Monitoring of biological diversity in space and time'. In: *Trends in Ecology & Evolution* 16.8 (Aug. 2001), pp. 446–453. DOI: 10.1016/S0169-5347(01)02205-4 (cited on page 2).

[19] Gurutzeta Guillera-Arroita, Martin S. Ridout, and Byron J. T. Morgan. 'Design of occupancy studies with imperfect detection'. en. In: *Methods in Ecology and Evolution* 1.2 (2010). _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.2041-210X.2010.00017.x, pp. 131–139. DOI: 10.1111/j.2041-210X.2010.00017.x (cited on page 2).

[20] B. K. Williams, J. D. Nichols, and M. J. Conroy. *Analysis and management of animal populations: modeling, estimation and decision making*. en. 2002 (cited on page 2).

[21] P.S. Laplace. 'Sur les naissances, les mariages et les morts'. In: *Histoire de L'Academic Royale des Sciences* (1786) (cited on page 3).

[22] C. G. J. Petersen. 'The yearly immigration of young plaice into the Limfjord from the German Sea, ect'. In: *Report of the Danish Biological Station for 1985* (1986) (cited on page 3).

[23] Lincoln. 'Calculating waterfowl abundance on the basis of banding returns'. In: *US Department of Agriculture* (1930) (cited on page 3).

[24] Douglas G. Chapman. 'Inverse, Multiple and Sequential Sample Censuses'. In: *Biometrics* 8.4 (1952). Publisher: [Wiley, International Biometric Society], pp. 286–306. DOI: 10.2307/3001864 (cited on page 3).

[25] J. N. Darroch. 'The Multiple-recapture census: I. estimation of a closed population'. In: *Biometrika* 45.3/4 (1958), pp. 343–359. DOI: 10.2307/2333183 (cited on pages 3, 18).

[26] David L. Otis et al. 'Statistical inference from capture data on closed animal populations'. In: *Wildlife Monographs* 62 (1978). Publisher: [Wiley, Wildlife Society], pp. 3–135 (cited on pages 3, 18).

[27] R. M. Cormack. 'Estimates of Survival from the Sighting of Marked Animals'. In: *Biometrika* 51.3/4 (1964). Publisher: [Oxford University Press, Biometrika Trust], pp. 429–438. DOI: 10.2307/2334149 (cited on pages 3, 113).

[28] G. M. Jolly. 'Explicit estimates from capture-recapture data with both death and immigration-stochastic model'. In: *Biometrika* 52.1/2 (1965), pp. 225–247. DOI: 10.2307/2333826 (cited on pages 3, 62, 113).

[29] G. A. F. Seber. 'A note on the multiple-recapture census'. In: *Biometrika* 52.1/2 (1965), pp. 249–259. DOI: 10.2307/2333827 (cited on pages 3, 62, 113).

[30] A. Neil Arnason. 'Parameter estimates from mark-recapture experiments on two populations subject to migration and death'. en. In: *Population Ecology* 13.2 (1972), pp. 97–113. DOI: 10.1007/BF02521971 (cited on pages 3, 69, 114).

[31] A. Neil Arnason. 'The estimation of population size, migration rates and survival in a stratified population'. en. In: *Population Ecology* 15.2 (1973), pp. 1–8. DOI: 10.1007/BF02510705 (cited on pages 3, 69, 114).

[32] Carl J. Schwarz, Jake F. Schweigert, and A. Neil Arnason. 'Estimating migration rates using tag-recovery data'. In: *Biometrics* 49.1 (1993), pp. 177–193. DOI: 10.2307/2532612 (cited on pages 3, 69).

[33] Luke Tierney. 'Markov Chains for Exploring Posterior Distributions'. In: *The Annals of Statistics* 22.4 (Dec. 1994). Publisher: Institute of Mathematical Statistics, pp. 1701–1728. DOI: 10.1214/aos/1176325750 (cited on page 5).

[34] Alan E. Gelfand. 'Gibbs Sampling'. In: *Journal of the American Statistical Association* 95.452 (Dec. 2000), pp. 1300–1304. DOI: 10.1080/01621459.2000.10474335 (cited on page 5).

[35] Siddhartha Chib and Edward Greenberg. 'Understanding the Metropolis-Hastings Algorithm'. In: *The American Statistician* 49.4 (1995). Publisher: [American Statistical Association, Taylor & Francis, Ltd.], pp. 327–335. DOI: 10.2307/2684568 (cited on page 5).

[36] Ming-Hui Chen, Qi-Man Shao, and Joseph G. Ibrahim. *Monte Carlo Methods in Bayesian Computation*. Springer Series in Statistics. New York, NY: Springer, 2000 (cited on page 5).

[37] Andrew Gelman et al. *Bayesian Data Analysis*. 3rd ed. New York: Chapman and Hall/CRC, July 2015 (cited on page 5).

[38] Andrew Gelman and Donald B. Rubin. 'Inference from Iterative Simulation Using Multiple Sequences'. In: *Statistical Science* 7.4 (1992). Publisher: Institute of Mathematical Statistics, pp. 457–472 (cited on page 5).

[39] Stephen P. Brooks and Andrew Gelman. 'General Methods for Monitoring Convergence of Iterative Simulations'. In: *Journal of Computational and Graphical Statistics* 7.4 (1998). Publisher: [American Statistical Association, Taylor & Francis, Ltd., Institute of Mathematical Statistics, Interface Foundation of America], pp. 434–455. DOI: 10.2307/1390675 (cited on page 5).

[40] J. M. Meyers. 'Leg bands cause injuries to parakeets and parrots'. en. In: *North American Bird Bander* 19.4 (1994), pp. 133–136 (cited on page 6).

[41] Nancy Burley. 'Leg-band color and mortality patterns in captive breeding populations of Zebra finches'. In: *The Auk* 102.3 (July 1985), pp. 647–651. DOI: 10.1093/auk/102.3.647 (cited on page 6).

[42] Dennis L Murray and Mark R Fuller. 'A critical review of the effects of marking on the biology of vertebrates'. In: *Research techniques in animal ecology: controversies and consequences*. Columbia University Press New York, 2000, pp. 15–64 (cited on page 6).

[43] L. Swanepoel, Micheal J. Somers, and F. Dalerum. 'Density of leopards Panthera pardus on protected and non-protected land in the Waterberg Biosphere, South Africa'. In: 2015. DOI: 10.2981/wlb.00108 (cited on pages 6, 107).

[44] Brandon M Quinby, J Curtis Creighton, and Elizabeth A Flaherty. 'Estimating population abundance of burying beetles using photo-identification and mark-recapture methods'. In: *Environmental Entomology* 50.1 (Feb. 2021), pp. 238–246. DOI: 10.1093/ee/nvaa139 (cited on pages 6, 107).

[45] Hélène Labach et al. 'Distribution and abundance of common bottlenose dolphin (Tursiops truncatus) over the French Mediterranean continental shelf'. en. In: *Marine Mammal Science* 38.1 (2022), pp. 212–222. DOI: 10.1111/mms.12874 (cited on pages 6, 107).

[46] K. Alexandra Curtis et al. 'Abundance, survival, and annual rate of change of Cuvier's beaked whales (Ziphius cavirostris) on a Navy sonar range'. en. In: *Marine Mammal Science* 37.2 (2021), pp. 399–419. DOI: 10.1111/mms.12747 (cited on pages 6, 107).

[47] Jonathan P. Crall et al. 'HotSpotter - Patterned species instance recognition'. In: *Proceedings of the 2013 IEEE Workshop on Applications of Computer Vision (WACV)*. WACV '13. USA: IEEE Computer Society, Jan. 2013, pp. 230–237. DOI: 10.1109/WACV.2013.6475023 (cited on pages 6, 10, 107).

[48] Pierre Taberlet and Gordon Luikart. 'Non-invasive genetic sampling and individual identification'. In: *Biological Journal of the Linnean Society* 68.1-2 (Sept. 1999), pp. 41–55. DOI: 10.1111/j.1095-8312.1999.tb01157.x (cited on pages 6–8, 84).

[49] Pierre Taberlet, Lisette P. Waits, and Gordon Luikart. 'Noninvasive genetic sampling: look before you leap'. en. In: *Trends in Ecology & Evolution* 14.8 (Aug. 1999), pp. 323–327. DOI: 10.1016/S0169-5347(99)01637-7 (cited on pages 6, 9, 107).

[50] Brian P. Dreher et al. 'Noninvasive estimation of black bear abundance incorporating genotyping errors and harvested bear'. en. In: *The Journal of Wildlife Management* 71.8 (2007), pp. 2684–2693. DOI: 10.2193/2006-398 (cited on pages 6, 107).

[51] Emily W. Ruell et al. 'Estimating bobcat population sizes and densities in a fragmented urban landscape using noninvasive capture–recapture sampling'. In: *Journal of Mammalogy* 90.1 (Feb. 2009), pp. 129–135. DOI: 10.1644/07-MAMM-A-249.1 (cited on pages 6, 107).

[52] Dana J. Morin et al. 'Efficient single-survey estimation of carnivore density using fecal DNA and spatial capture-recapture: a bobcat case study'. en. In: *Population Ecology* 60.3 (July 2018), pp. 197–209. DOI: 10.1007/s10144-018-0606-9 (cited on pages 6, 107).

[53] Susannah P. Woodruff et al. 'Estimating Sonoran pronghorn abundance and survival with fecal DNA and capture—recapture methods'. In: *Conservation Biology* 30.5 (2016), pp. 1102–1111 (cited on pages 6, 107).

[54] A. Laguardia et al. 'Nationwide abundance and distribution of African forest elephants across Gabon using non-invasive SNP genotyping'. en. In: *Global Ecology and Conservation* 32 (Dec. 2021), e01894. DOI: 10.1016/j.gecco.2021.e01894 (cited on pages 6, 11, 107).

[55] Alain Vignal et al. 'A review on SNP and other types of molecular markers and their use in animal genetics'. en. In: *Genetics Selection Evolution* 34.3 (May 2002), p. 275. DOI: 10.1186/1297-9686-34-3-275 (cited on page 7).

[56] Phillip A. Morin et al. 'SNPs in ecology, evolution and conservation'. en. In: *Trends in Ecology & Evolution* 19.4 (Apr. 2004), pp. 208–216. DOI: 10.1016/j.tree.2004.01.009 (cited on page 7).

[57] G. C. B. Schopen et al. 'Comparison of information content for microsatellites and SNPs in poultry and cattle'. en. In: *Animal Genetics* 39.4 (2008), pp. 451–453. DOI: https://doi.org/10.1111/j.1365-2052.2008.01736.x (cited on page 7).

[58] Pierre Taberlet et al. 'Reliable genotyping of samples with very low DNA quantities using PCR'. In: *Nucleic Acids Research* 24.16 (Aug. 1996), pp. 3189–3194. DOI: 10.1093/nar/24.16.3189 (cited on page 8).

[59] W Navidi, N Arnheim, and M S Waterman. 'A multiple-tubes approach for accurate genotyping of very small DNA samples by using PCR: statistical considerations.' In: *American Journal of Human Genetics* 50.2 (Feb. 1992), pp. 347–359 (cited on page 8).

[60] Lisette P. Waits and David Paetkau. 'Noninvasive genetic sampling tools for wildlife biologists: a review of applications and recommendations for accurate data collection'. en. In: *The Journal of Wildlife Management* 69.4 (2005), pp. 1419–1433. DOI: 10.2193/0022-541X(2005)69[1419:NGSTFW]2.0.CO;2 (cited on pages 9, 10, 107).

[61] Simon Bonner and Jason Holmberg. 'Mark-recapture with multiple, non-invasive marks'. en. In: *Biometrics* 69.3 (2013), pp. 766–775. DOI: 10.1111/biom.12045 (cited on pages 9, 12).

[62] Brett T. McClintock et al. 'Integrated modeling of bilateral photo-identification data in mark–recapture analyses'. en. In: *Ecology* 94.7 (2013). _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1890/12-1613.1, pp. 1464–1471. DOI: 10.1890/12-1613.1 (cited on pages 9, 12).

[63] Brett T. McClintock. 'multimark: an R package for analysis of capture–recapture data consisting of multiple "noninvasive" marks'. en. In: *Ecology and Evolution* 5.21 (2015), pp. 4920–4931. DOI: 10.1002/ece3.1676 (cited on pages 9, 12, 107).

[64] Jun Yoshizaki et al. 'Modeling misidentification errors in capture–recapture studies using photographic identification of evolving marks'. en. In: *Ecology* 90.1 (2009), pp. 3–9. DOI: 10.1890/08-0304.1 (cited on pages 9, 104).

[65] Scott Creel et al. 'Population size estimation in Yellowstone wolves with error-prone noninvasive microsatellite genotypes'. en. In: *Molecular Ecology* 12.7 (2003), pp. 2003–2009. DOI: 10.1046/j.1365-294X.2003.01868.x (cited on pages 10, 107).

[66] Kristopher J. Winiarski and Kevin McGarigal. 'Effects of photo and genotype-based misidentification error on estimates of survival, detection and state transition using multistate survival models'. In: *PLoS ONE* 11.1 (Jan. 2016), e0145640. DOI: 10.1371/journal.pone.0145640 (cited on page 10).

[67] Jun Yoshizaki et al. 'Modeling misidentification errors that result from use of genetic tags in capture–recapture studies'. In: *Environmental and Ecological Statistics* 18 (Mar. 2011), pp. 27–55. DOI: 10.1007/s10651-009-0116-1 (cited on pages 10–12, 18, 45, 55, 58, 99, 107, 108, 115, 116).

[68] D. Paetkau. 'An empirical exploration of data quality in DNA-based population inventories'. en. In: *Molecular Ecology* 12.6 (2003), pp. 1375–1387. DOI: 10.1046/j.1365-294X.2003.01820.x (cited on pages 10, 107).

[69] K. S. McKelvey and M. K. Schwartz. 'dropout: a program to identify problem loci and samples for noninvasive genetic samples in a capture-mark-recapture framework'. en. In: *Molecular Ecology Notes* 5.3 (2005), pp. 716–718. DOI: 10.1111/j.1471-8286.2005.01038.x (cited on pages 10, 107).

[70] Douglas T. Bolger et al. 'A computer-assisted system for photographic mark–recapture analysis'. en. In: *Methods in Ecology and Evolution* 3.5 (2012), pp. 813–822. DOI: 10.1111/j.2041-210X.2012.00212.x (cited on pages 10, 107).

[71] Paul M. Lukacs and Kenneth P. Burnham. 'Estimating population size from DNA-based closed capture-recapture data incorporating genotyping error'. en. In: *The Journal of Wildlife Management* 69.1 (2005), pp. 396–403. DOI: https://doi.org/10.2193/0022-541X(2005)069<0396:EPSFDC>2.0.CO;2 (cited on pages 11, 57, 107, 108, 117).

[72] Janine A. Wright et al. 'Incorporating genotype uncertainty into mark-recapture-type models for estimating abundance using DNA samples'. eng. In: *Biometrics* 65.3 (Sept. 2009), pp. 833–840. DOI: 10.1111/j.1541-0420.2008.01165.x (cited on pages 11, 107, 108).

[73] Shannon M. Knapp, Bruce A. Craig, and Lisette P. Waits. 'Incorporating genotyping error into non-Invasive DNA-based mark—recapture population estimates'. en. In: *The Journal of Wildlife Management* 73.4 (2009). _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.2193/2007-156, pp. 598–604. DOI: 10.2193/2007-156 (cited on pages 11, 107, 108).

[74] William A. Link et al. 'Uncovering a latent multinomial: analysis of mark–recapture data with misidentification'. en. In: *Biometrics* 66.1 (2010), pp. 178–185. DOI: https://doi.org/10.1111/j.1541-0420.2009.01244.x (cited on pages 12–14, 18, 20, 21, 27, 29, 36, 51, 58, 101, 104, 107, 108, 110, 114, 115).

[75] Brett T. McClintock et al. 'Probit models for capture-recapture data subject to imperfect detection, individual heterogeneity and misidentification'. In: *arXiv e-prints* 1401 (Jan. 2014), arXiv:1401.3290 (cited on pages 13, 37, 40, 42, 66, 99, 108, 109, 112).

[76] David A. Fournier et al. 'AD Model Builder: using automatic differentiation for statistical inference of highly parameterized complex nonlinear models'. In: *Optimization Methods and Software* 27.2 (Apr. 2012). Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/10556788.2011.597854, pp. 233–249. DOI: 10.1080/10556788.2011.597854 (cited on pages 13, 101).

[77] R. T. R. Vale et al. 'Maximum likelihood estimation for model $M_{t,\alpha}$ for capture–recapture data with misidentification'. en. In: *Biometrics* 70.4 (2014), pp. 962–971. DOI: https://doi.org/10.1111/biom.12195 (cited on pages 13, 17, 29, 99, 101, 108, 115).

[78] Matthew R. Schofield and Simon Bonner. 'Connecting the latent multinomial'. en. In: *Biometrics* 71.4 (2015), pp. 1070–1080. DOI: https://doi.org/10.1111/biom.12333 (cited on pages 13, 21, 31, 101, 108).

[79] Simon Bonner et al. 'Extending the latent multinomial model with complex error processes and dynamic markov bases'. In: *The Annals of Applied Statistics* 10 (Apr. 2015). DOI: 10.1214/15-AOAS889 (cited on pages 13, 17, 19–22, 36, 37, 61, 101, 108–110, 114).

[80] Perry de Valpine et al. 'Programming with models: writing statistical algorithms for general model structures with NIMBLE'. In: *Journal of Computational and Graphical Statistics* 26.2 (Apr. 2017), pp. 403–413. DOI: 10.1080/10618600.2016.1172487 (cited on pages 17, 30, 45, 56, 73, 114).

[81] Persi Diaconis and Bernd Sturmfels. 'Algebraic algorithms for sampling from conditional distributions'. In: *The Annals of Statistics* 26.1 (1998), pp. 363–397 (cited on page 21).

[82] 4ti2 team. *4ti2—A software package for algebraic, geometric and combinatorial problems on linear spaces* (cited on page 21).

[83] Adrian Dobra. 'Dynamic markov bases'. In: *Journal of Computational and Graphical Statistics* 21.2 (Apr. 2012). Publisher: Taylor & Francis, pp. 496–517. DOI: 10.1080/10618600.2012.663285 (cited on pages 21, 64, 72).

[84] Christian P. Robert and George Casella. *Monte Carlo Statistical Methods*. Springer Texts in Statistics. New York, NY: Springer, 2004 (cited on page 22).

[85] E. Çınlar. *Introduction to stochastic processes*. English. Ed. by Norman J. Sollenberger. Dover edition. Mineola, New York: Dover Publications, Inc, 2013 (cited on page 24).

[86] Jun S. Liu. *Monte Carlo Strategies in Scientific Computing*. Springer Series in Statistics. New York, NY: Springer, 2004 (cited on page 24).

[87] Elizabeth S. Garrett and Scott L. Zeger. 'Latent class model diagnosis'. In: *Biometrics* 56.4 (2000), pp. 1055–1067 (cited on pages 29, 30, 32, 33).

[88] Diana J. Cole and Rachel S. McCrea. 'Parameter redundancy in discrete state-space and integrated models'. en. In: *Biometrical Journal* 58.5 (2016), pp. 1071–1090. DOI: 10.1002/bimj.201400239 (cited on page 29).

[89] Hannah Worthington et al. 'Estimation of population size when capture probability depends on individual states'. en. In: *Journal of Agricultural, Biological and Environmental Statistics* 24.1 (Mar. 2019), pp. 154–172. DOI: 10.1007/s13253-018-00347-x (cited on pages 30, 98).

[90] Bradley P. Carlin and Thomas A. Louis. *Bayes and empirical Bayes methods for data analysis*. Monographs on statistics and applied probability 69. London ; Melbourne: Chapman & Hall, 1996 (cited on page 36).

[91] C. Miquel et al. 'Quality indexes to assess the reliability of genotypes in studies using noninvasive sampling and multiple-tube approach'. en. In: *Molecular Ecology Notes* 6.4 (2006). _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1471-8286.2006.01413.x, pp. 985–988. DOI: 10.1111/j.1471-8286.2006.01413.x (cited on pages 39, 109).

[92] I Illumina. 'Quality scores for next-generation sequencing'. In: *Technical Note: Informatics* 31 (2011) (cited on pages 39, 109).

[93] Simone Lampa et al. 'Non-Invasive Genetic Mark-Recapture as a Means to Study Population Sizes and Marking Behaviour of the Elusive Eurasian Otter (Lutra lutra)'. en. In: *PLOS ONE* 10.5 (May 2015). Publisher: Public Library of Science, e0125684. DOI: 10.1371/journal.pone.0125684 (cited on pages 57, 59).

[94] Hans Kruuk. *Otters: Ecology, behaviour and conservation*. Oxford University Press, Aug. 2006 (cited on page 57).

[95] Gary C. White and Kenneth P. Burnham. 'Program MARK: survival estimation from populations of marked animals'. In: *Bird Study* 46.sup1 (Jan. 1999). Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/00063659909477239, S120–S139. DOI: 10.1080/00063659909477239 (cited on pages 57, 58, 117).

[96] E. A. Catchpole et al. 'Integrated recovery/recapture data analysis'. In: *Biometrics* 54.1 (1998). Publisher: [Wiley, International Biometric Society], pp. 33–46. DOI: 10.2307/2533993 (cited on page 62).

[97] R. King and S. P. Brooks. 'Closed-form likelihoods for Arnason-Schwarz models'. In: *Biometrika* 90.2 (2003), pp. 435–444 (cited on page 69).

[98] Carl James Schwarz and A. Neil Arnason. 'A general methodology for the analysis of capture-recapture experiments in open populations'. In: *Biometrics* 52.3 (1996), pp. 860–873. DOI: 10.2307/2533048 (cited on page 69).

[99] Colince Kamdem et al. 'Anthropogenic habitat disturbance and ecological divergence between incipient species of the malaria mosquito Anopheles gambiae'. en. In: *PLOS ONE* 7.6 (June 2012). Publisher: Public Library of Science, e39453. DOI: 10.1371/journal.pone.0039453 (cited on page 80).

[100] Billy Tene Fossog et al. 'Physiological correlates of ecological divergence along an urbanization gradient: differential tolerance to ammonia among molecular forms of the malaria mosquito Anopheles gambiae'. eng. In: *BMC ecology* 13 (Jan. 2013), p. 1. DOI: 10.1186/1472-6785-13-1 (cited on page 80).

[101] Ranjan Ramasamy and Sinnathamby Noble Surendran. 'Global climate change and its potential impact on disease transmission by salinity-tolerant mosquito vectors in coastal zones'. eng. In: *Frontiers in Physiology* 3 (2012), p. 198. DOI: 10.3389/fphys.2012.00198 (cited on page 80).

[102] Billy Tene Fossog et al. 'Habitat segregation and ecological character displacement in cryptic African malaria mosquitoes'. en. In: *Evolutionary Applications* 8.4 (2015), pp. 326–345. DOI: 10.1111/eva.12242 (cited on page 80).

[103] Ranjan Ramasamy et al. 'Biological differences between brackish and fresh water-derived Aedes aegypti from two locations in the Jaffna peninsula of Sri Lanka and the implications for arboviral disease transmission'. eng. In: *PloS One* 9.8 (2014), e104977. DOI: 10.1371/journal.pone.0104977 (cited on page 80).

[104] Edith O. Lyimo and W. Takken. 'Effects of adult body size on fecundity and the pre-gravid rate of Anopheles gambiae females in Tanzania'. en. In: *Medical and Veterinary Entomology* 7.4 (1993), pp. 328–332. DOI: 10.1111/j.1365-2915.1993.tb00700.x (cited on page 80).

[105] Kathryne L. Dieter, Diana L. Huestis, and Tovi Lehmann. 'The effects of oviposition-site deprivation on Anopheles gambiae reproduction'. In: *Parasites & Vectors* 5.1 (Oct. 2012), p. 235. DOI: 10.1186/1756-3305-5-235 (cited on page 80).

[106] J. C. Hogg and H. Hurd. 'The effects of natural Plasmodium falciparum infection on the fecundity and mortality of Anopheles gambiae s. l. in north east Tanzania'. en. In: *Parasitology* 114.4 (Apr. 1997). Publisher: Cambridge University Press, pp. 325–331. DOI: 10.1017/S0031182096008542 (cited on page 80).

[107] E. S. Nnakumusana. 'The effect of Coelomomyces indicus on the fecundity and longevity of Anopheles Gambiae, Culex fatigans and Aedes aegypti exposed to infection at each larval instar'. en. In: *International Journal of Tropical Insect Science* 7.2 (Apr. 1986). Publisher: Cambridge University Press, pp. 139–142. DOI: 10.1017/S1742758400008869 (cited on page 80).

[108] Hortance Manda et al. 'Effect of discriminative plant-sugar feeding on the survival and fecundity of Anopheles gambiae'. In: *Malaria Journal* 6.1 (Aug. 2007), p. 113. DOI: 10.1186/1475-2875-6-113 (cited on page 80).

[109] M. N. Bayoh and S. W. Lindsay. 'Effect of temperature on the development of the aquatic stages of Anopheles gambiae sensu stricto'. en. In: *Bulletin of Entomological Research* 93.5 (Sept. 2003), pp. 375–381. DOI: 10.1079/BER2003259 (cited on page 82).

[110] M. W. Service. 'Studies on sampling larval populations of the Anopheles gambiae complex'. In: *Bulletin of the World Health Organization* 45.2 (1971), pp. 169–180 (cited on pages 82, 83).

[111] M. W. Service. 'Mortalities of the larvae of the Anopheles gambiae Giles complex and detection of predators by the precipitin test'. en. In: *Bulletin of Entomological Research* 62.3 (Feb. 1973), pp. 359–369. DOI: 10.1017/S0007485300003862 (cited on pages 82, 83).

[112] M Mathis. 'Cycle biologique complet d'Anopheles gambiae Giles élevé en série au laboratoire'. In: *CR Soc. Biol. Paris* 119 (1935), p. 1385 (cited on page 82).

[113] M. H. Holstein. 'Biology of Anopheles gambiae. Research in French West Africa.' English. In: *Biology of Anopheles gambiae. Research in French West Africa.* (1954) (cited on page 82).

[114] M. Bates. 'Field Studies of the Anopheline Mosquitoes of Albania.' In: 1941 (cited on page 83).

[115] M. W. Service. 'Mortalities of the immature stages of species B of the Anopheles gambiae complex in Kenya: comparison between rice fields and temporary pools, identification of predators, and effects of insecticidal spraying'. eng. In: *Journal of Medical Entomology* 13.4-5 (Jan. 1977), pp. 535–545. DOI: 10.1093/jmedent/13.4-5.535 (cited on page 83).

[116] Martin H. Birley. 'The estimation and simulation of variable developmental period, with application to the mosquito Aedes aegypti (L.)' en. In: *Researches on Population Ecology* 21.1 (Oct. 1979), p. 68. DOI: 10.1007/BF02512639 (cited on page 90).

[117] Nathan R. Campbell, Stephanie A. Harmon, and Shawn R. Narum. 'Genotyping-in-Thousands by sequencing (GT-seq): a cost effective SNP genotyping method based on custom amplicon sequencing'. en. In: *Molecular Ecology Resources* 15.4 (2015), pp. 855–867. DOI: https://doi.org/10.1111/1755-0998.12357 (cited on page 102).

[118] Evan H. Campbell Grant. 'Visual implant elastomer mark retention through metamorphosis in amphibian larvae'. en. In: *The Journal of Wildlife Management* 72.5 (2008), pp. 1247–1252. DOI: 10.2193/2007-183 (cited on page 104).

[119] Roger Pradel. 'Multievent: an extension of multistate capture–recapture models to uncertain states'. en. In: *Biometrics* 61.2 (2005), pp. 442–447. DOI: 10.1111/j.1541-0420.2005.00318.x (cited on page 105).

[120] Ben C. Augustine et al. 'Spatial proximity moderates genotype uncertainty in genetic tagging studies'. In: *Proceedings of the National Academy of Sciences* 117.30 (July 2020). Publisher: Proceedings of the National Academy of Sciences, pp. 17903–17912. DOI: 10.1073/pnas.2000247117 (cited on page 105).

Notations

Parameters	
N	Population size
N_t	Number of individual first captured at occasion t (open population models)
$N_{s,t}$	Number of individual first captured at occasion t in state s (open population models)
p_t	Capture probability at occasion t
λ_t	Expected number of observations of an individual at occasion t in the Poisson model
α	identification probability
$\alpha_{n,t}$	identification probability of individual n at occasion t
$\theta_\alpha = (a, b)$	regression parameters of the probit model
ϕ	survival probability of individual n at occasion t
$\boldsymbol{\psi}$	matrix of transition probabilities
$\psi_{r,s}$	transition probability from state r to state s
$\boldsymbol{\delta}$	vector of initial state probabilities
δ_s	probability that an individual is in state s at the beginning of the experiment

Data and latent variables	
ω_i	Observed history i
ν_j	Latent error history j (in which misidentification are noted down)
ξ_k	Latent capture history k (true capture history)
$\mathbf{Y}$	set of observed histories
$\mathbf{X}$	set of latent error histories
$\mathbf{Z}$	set of latent capture histories
y_i	number of observed history i
x_j	number of latent error history j
z_k	number of latent capture history k
$\mathbf{y} = (y_1, ..., y_{2^T-1})$	vector of counts of observed histories
$\mathbf{x} = (x_1, ..., x_{3^T}))$	vector of counts of latent error histories
$\mathbf{z} = (z_1, ..., z_{2^T}))$	vector of counts of latent capture histories
$o_{j,t}$	number of correct identifications in history j at occasion t (for the repeated observation model, in Chapter 3)

Computed values	
π_k	probability of history k
χ_t	probability that an individual, alive at t, is not seen again after
$Q_{(c,d)}(r, s)$	probability that an animal changes from state r at time c to state s at time d, and is unobserved between these times
$O_{(c,d)}(r, s)$	probability that an animal in state r at time c remains unobserved until it is subsequently resighted in state s at time $d + 1$
$O_{(c,d)}(r_1, r_2, s_1, s_2)$	probability that an animal in state r_1 and r_2 at times c and $c + 1$ remains unobserved until it is subsequently resighted in state s_1 and s_2 at times $d + 1$ and $d + 2$

Models	
M_t	Classical closed population model (no misidentifications)
$M_{t,\alpha}$	Closed population model with misidentification (constant across individual)
M_{t,α_n}	Closed population model with identification covariates
$M_{\lambda,\alpha}$	Closed population model with multiple recapture and misidentification
CJS	Cormack-Jolly-Seber (CJS) model (no misidentifications)
CJS_α	CJS model with misidentifications
CJS_{α_n}	CJS model with identifications covariates
AS	Arnason-Schwartz (AS), open population multi-state model (no misidentifications)
AS_α	AS model with misidentifications

Notations

Parameters	
N	Population size
N_t	Number of individual first captured at occasion t (open population models)
$N_{s,t}$	Number of individual first captured at occasion t in state s (open population models)
p_t	Capture probability at occasion t
λ_t	Expected number of observations of an individual at occasion t in the Poisson model
α	identification probability
$\alpha_{n,t}$	identification probability of individual n at occasion t
$\theta_\alpha = (a, b)$	regression parameters of the probit model
ϕ	survival probability of individual n at occasion t
$\boldsymbol{\psi}$	matrix of transition probabilities
$\psi_{r,s}$	transition probability from state r to state s
$\boldsymbol{\delta}$	vector of initial state probabilities
δ_s	probability that an individual is in state s at the beginning of the experiment

Data and latent variables	
ω_i	Observed history i
ν_j	Latent error history j (in which misidentification are noted down)
ξ_k	Latent capture history k (true capture history)
$\mathbf{Y}$	set of observed histories
$\mathbf{X}$	set of latent error histories
$\mathbf{Z}$	set of latent capture histories
y_i	number of observed history i
x_j	number of latent error history j
z_k	number of latent capture history k
$\mathbf{y} = (y_1, ..., y_{2^T-1})$	vector of counts of observed histories
$\mathbf{x} = (x_1, ..., x_{3^T}))$	vector of counts of latent error histories
$\mathbf{z} = (z_1, ..., z_{2^T}))$	vector of counts of latent capture histories
$o_{j,t}$	number of correct identifications in history j at occasion t (for the repeated observation model, in Chapter 3)

Computed values	
π_k	probability of history k
χ_t	probability that an individual, alive at t, is not seen again after
$Q_{(c,d)}(r, s)$	probability that an animal changes from state r at time c to state s at time d, and is unobserved between these times
$O_{(c,d)}(r, s)$	probability that an animal in state r at time c remains unobserved until it is subsequently resighted in state s at time $d + 1$
$O_{(c,d)}(r_1, r_2, s_1, s_2)$	probability that an animal in state r_1 and r_2 at times c and $c + 1$ remains unobserved until it is subsequently resighted in state s_1 and s_2 at times $d + 1$ and $d + 2$

Models	
M_t	Classical closed population model (no misidentifications)
$M_{t,\alpha}$	Closed population model with misidentification (constant across individual)
M_{t,α_n}	Closed population model with identification covariates
$M_{\lambda,\alpha}$	Closed population model with multiple recapture and misidentification
CJS	Cormack-Jolly-Seber (CJS) model (no misidentifications)
CJS_α	CJS model with misidentifications
CJS_{α_n}	CJS model with identifications covariates
AS	Arnason-Schwartz (AS), open population multi-state model (no misidentifications)
AS_α	AS model with misidentifications