Research Topic 2:

An Overview of Capture-Recapture in Closed and Open Populations

L. M. Howell

STOR-i | Lancaster University

1 Introduction

Some of the oldest models in statistical ecology are models designed to analyse capture recapture data. The aim of these models is to get an estimate of the size of a population of interest. The book *Analysis of Capture-Recapture Data* by McCrea and Morgan (2015) is a recommended read, and heavily influences §2.

At a given capture time, animals are caught and, if not already, marked in some way. Often this is done with a tag such as a ring around the leg. Then at a later date a 'recapture' event is attempted. This can be repeated multiple times, with each new recapture attempt adding newly marked individuals. This creates observation data for an individual of the form

$$\begin{array}{lcccc} \text{Individual 1:} & 1 & 1 & 0 & 1 & \dots \\ \text{Individual 2:} & 0 & 1 & 0 & 0 & \dots \\ \text{Individual 3:} & 0 & 1 & 0 & 1 & \dots \end{array}$$

where a 1 corresponds to an individual being observed at that capture/recapture event and a 0 otherwise. The goal is then to estimate the number of unobserved individuals with capture histories

$$0 \quad 0 \quad 0 \quad 0 \quad \dots.$$

This report will begin with some background in closed capture-recapture models and their adaptations, including an expansion into spatially explicit capture-recapture models with the aim of gaining insight into population density. This will include an example using `R`, and also a look at an example which adopts a Bayesian framework. This will be expanded upon by looking at the open population model known as the Jolly-Seber Model. Following this is a discussion of open challenges within capture-recapture models and the wider field of statistical ecology that still need researching.

2 Capture-Recapture Models

In the most basic form of the data, with just two capture occasions, the traditional method is the Lincoln-Petersen Estimator. Intuitively, the idea is that the ratio between the marked individuals and the total individuals at the second capture event is equal to the proportion of individuals caught at the first capture event and the total population. Say n_1 individuals are caught at capture occasion 1, which are then marked and released. Then n_2 individuals are caught at the second capture occasion, where m_2 of them are some of the marked population from the previous capture. Then

$$\frac{m_2}{n_2} = \frac{n_1}{N} \implies \hat{N} = \frac{n_1 n_2}{m_2}$$

creates an estimate of the total population N (Pollock et al., 1990). The full assumptions under which this estimator is valid will be discussed later. To expand the model to multiple capture occasions, let p be the probability that an individual is captured at an event. Then, assuming

that the probability of an individual being captured at an event is independant of all other events, the probability of observing that capture history is

$$\begin{array}{lccccc} \text{Individual 1:} & 1 & 1 & 0 & 1 & \dots & pp(1-p)p\dots \\ \text{Individual 2:} & 0 & 1 & 0 & 0 & \dots & (1-p)p(1-p)(1-p)\dots \\ \text{Individual 3:} & 0 & 1 & 0 & 1 & \dots & (1-p)p(1-p)p\dots \end{array}$$

The corresponding unobserved individuals then have an encounter history probability of

$$(1-p)(1-p)(1-p)(1-p)\dots$$

2.1 Likelihood

Formally, denote the data matrix $\mathbf{X}$ with N rows and T columns. N is the total population size and T is the number of capture/recapture events. The ith row is the observation history of an individual, with the bottom $N-D$ rows being the unobserved histories, D being the observed population. Often multiple individuals will have the same capture history. Denote this history $\tilde{x}_h$. Let f_t be the number of individuals caught t times, and n_t be the number of individuals caught at capture event $t \in \{1, 2, \dots, T\}$. Throughout this report, $\{1, 2, \dots, T\}$ will be shortened to $[T]$. Therefore

$$f_0 = N - D \tag{1}$$

and $D = \sum_{t=1}^{T} f_t$. Call $\mathbf{f}$, the vector of f_t's, the census data.

Due to the assumed independence the likelihood function is of the form

$$L(N, p; \mathbf{X}) = \frac{N!}{\left(\prod_h \tilde{x}_h\right)(N-D)!} \prod_{i=1}^{N} \prod_{t=1}^{T} p^{x_{it}} (1-p)^{(1-x_{it})}.$$

This can be first simplified to

$$L(N, p; \mathbf{X}) = \frac{N!}{\left(\prod_h \tilde{x}_h\right)(N-D)!} \prod_{t=1}^{T} p^{n_t} (1-p)^{(N-n_t)}. \tag{2}$$

as p is assumed the same for all individuals in the population N. Then since the probability of being captured at time t is assumed to be the same for all $t \in [T]$ this further reduces to

$$L(N, p; \mathbf{X}) \propto \frac{N!}{(N-D)!} \left(p^S (1-p)^{(NT-S)}\right) \tag{3}$$

where S denotes the total number of captures, or $S = \sum_{t=1}^{T} t f_t$. The proportionality comes from dropping the $\prod_h \tilde{x}_h$ term. Then taking the log-likelihood this becomes

$$l(N, p; \mathbf{X}) \propto \log\left(\frac{N!}{(N-D)!}\right) + S\log(p) + (NT - S)\log(1-p) \implies$$

$$\frac{\mathrm{d}}{\mathrm{d}p} l(N, p; \mathbf{X}) \propto \frac{S}{p} - \frac{NT - S}{1-p} \implies$$

$$0 = \frac{S(1-p) - (NT - S)p}{p(1-p)} \implies$$
$$0 = S - pS - pNT + pS \implies$$
$$\hat{p} = \frac{S}{NT}$$

as a Maximum Likelihood Estimate for p. To calculate this however, N is still needed.

2.2 Zero Truncated Poisson Distribution

Let Y be a Poisson distributed random variable with parameter λ that models the census data $\mathbf{f}$. Since f_0 is unknown, formulate a zero truncated conditional probability function. Then

$$\begin{aligned} P(Y = y \mid y > 0) &= \frac{P(Y = y)}{P(y > 0)} \\ &= \frac{e^{-\lambda}\lambda^y}{y!(1 - e^{-\lambda})} \end{aligned}$$

Constructing a likelihood for n observations of Y with standard independence assumptions results in

$$L(y_1, \ldots, y_n; \lambda) = (1 - e^{-\lambda})^{-n} \prod_{i=1}^{n} \frac{e^{-\lambda}\lambda^{y_i}}{y_i!}.$$

Then to get an MLE the log-likelihood becomes

$$l(y_1, \ldots, y_n; \lambda) = -n \log(1 - e^{-\lambda}) + \sum_{i=1}^{n} [-\lambda + y_i \log(\lambda) - \log(y_i!)]$$

$$\implies \frac{\mathrm{d}l}{\mathrm{d}\lambda} = \frac{ne^{-\lambda}}{1 - e^{-\lambda}} + \sum_{i=1}^{n} \left[\frac{y_i}{\lambda} - 1\right]$$

Setting this equal to zero and solving for λ gives

$$\begin{aligned} \frac{n}{(e^{\lambda} - 1)} - n + \sum_{i=1}^{n} \frac{y_i}{\lambda} &= 0 \\ \frac{1}{(e^{\lambda} - 1)} - 1 + \frac{1}{\lambda}\frac{1}{n}\sum_{i=1}^{n} y_i &= 0 \\ \frac{\lambda}{(e^{\lambda} - 1)} - \lambda + \frac{1}{n}\sum_{i=1}^{n} y_i &= 0 \\ \bar{y} &= \lambda - \frac{\lambda}{(e^{\lambda} - 1)} \end{aligned}$$

Thus λ must be computed numerically (Cohen, 1960). It is more intuitive to see the log-likelihood expressed using the census data;

$$l(\lambda; \mathbf{f}) = constant + \log(\lambda) \sum_{t=1}^{\infty} t f_t + \log(e^{\lambda} - 1) \sum_{t=1}^{\infty} f_t.$$

Again this is maximised numerically. Then an estimate for N can be achieved by inflating D by dividing by the probability of being observed, which is equivalent to one minus the probability of being unobserved, $1 - \hat{p}_0$;

$$\hat{N} = \frac{D}{1 - e^{-\hat{\lambda}}}.$$

2.3 Chao's Lower Bound Estimator

When captures follow a Poisson model with parameter λ, it is simple to see that

$$p_0 = e^{-\lambda}, \quad p_1 = \lambda e^{-\lambda}, \quad p_2 = \frac{\lambda^2 e^{-\lambda}}{2}, \quad \dots.$$

Since $p_1^2 = \lambda^2 (e^{-\lambda})^2 \implies e^{-\lambda} = \frac{p_1^2}{\lambda^2 e^{-\lambda}}$. This clearly links with p_2 giving

$$p_0 = e^{-\lambda} = \frac{p_1^2}{2p_2}$$

which can be easily replaced with the proportions from the census data. Thus emerges an estimator for f_0;

$$\hat{f}_0 = \frac{f_1^2}{2f_2} \tag{4}$$

and consequent estimator for N using Equation 1

$$\hat{N} = D + \frac{f_1^2}{2f_2}. \tag{5}$$

It is often useful to assume heterogeneity in the capture probabilities (discussed more in §2.5), in which case it is assumed that λ follows some distribution. The Cauchy-Schwarz inequality gives the fractional inequalities

$$\frac{p_1}{p_0} \leq \frac{2p_2}{p_1} \leq \frac{3p_3}{p_2} \leq \dots$$

and the first inequality here results in

$$p_0 \geq \frac{p_1^2}{2p_2}.$$

This shows that Eq. 4 provides a lower bound on the number of unobserved individuals, and similarly Eq. 5 is a lowerbound on the population size. Eq. 5 is known as Chao's Lower Bound Estimator (Chao, 1987; McCrea and Morgan, 2015).

Furthermore, the inequalities above can give some insight into a way of 'testing' for heterogeneity within capture-recapture datasets. As t increases, the ratios $\frac{tp_t}{p_{t-1}}$ are increasing. Thus if a plot of $\frac{tf_t}{f_{t-1}}$ against t for census counts shows an increasing function, this implies the heterogeneity assumption that causes the inequalities (McCrea and Morgan, 2015, pg. 47).

2.4 Conditional Likelihood

Another way of calculating N in the multinomial case is using conditional likelihoods. Equation 2 can be re-expressed in terms of the census data.

$$L(N, \mathbf{p}; \mathbf{f}) \propto \frac{N!}{(N-D)!} \prod_{t=0}^{T} p_t^{f_t} \equiv \binom{N}{D} p_0^{N-D} (1-p_0)^D D! \prod_{t=1}^{T} \left(\frac{p_t}{1-p_0} \right)^{f_t}$$

Simply conditioning on D, this gives a multinomial likelihood

$$L_c(N, \mathbf{p}; \mathbf{f}) \propto D! \prod_{t=1}^{T} \left(\frac{p_t}{1 - p_0} \right)^{f_t} \tag{6}$$

which is a factor of $L(N, \mathbf{p}; \mathbf{f})$. Maximising 6 produces model parameters $\mathbf{p}$, in turn giving an estimate for p_0. Then the population size estimate conditional on the number of observed individuals is

$$\hat{N}_c = \frac{D}{1 - \hat{p_0}}$$

There is a discrepancy between the values of the estimates $\hat{N}$ and $\hat{N}_c$. However both have been used; as Fewster and Jupp (2009) discuss in *Inference on population size in binomial detectability models*, in a capture-recapture setting Otis et al. (1978) use $\hat{N}$ whereas in a distance sampling setting Buckland et al. (2001) prefer $\hat{N}_c$. (Distance sampling is where an observer exists at a point and records the distance to observed objects, e.g. birds nests, or individuals of interest.) Indeed, Fewster and Jupp (2009) show that the difference of $\hat{N}$ and $\hat{N}_c$ is of order 1 when p is constant.

Other ways of estimating the parameter p_0 exist, for example by assuming that the census counts follow a negative binomial distribution. For more, as well as a comparison of these methods, see McCrea and Morgan (2015, pg. 32).

2.5 Model Assumptions

As mentioned above, this model has been used in various applications. However it has a lot of limitations due to it's simplicity. The assumption was made that the probability of capture p was constant at all capture times. This model often is denoted M_0 as there is no other effects taken into account. The other types of model in this case are:

- M_b - behavioural effects where the first capture probability is different from subsequent capture probabilities.
- M_t - temporal effects induced by environmental factors such as weather.
- M_g - group effects for example individuals of a certain age or colour being less likely to be caught.
- M_h - heterogeneous effects, where each member of the population has a different capture probability.

The behavioural effects can come from individuals becoming 'trap happy', where they learn that they get free food and are more likely to be caught at a later occasion. There is also an assumption that whatever marking method is used for later identification has no effect on the individuals survival or chance of recapture. The other major assumption of this model was that of a closed, static population. In practice, immigration can occur where individuals enter or

leave the area being measured. Moreover, this can effect the population of interest in different ways. For example, say an individual is observed once and then never again, that individual could have emigrated or died, and these cases might be treated differently depending on what the population of interest is. If the population of interest is the number of adults able to reproduce, then the death or migration of a younger animal may want to be observed differently.

3 Spatially Explicit Capture-Recapture

One criticism of the above models is a non well-defined spatial element. It is generally unclear what area a population estimate covers, and there is no insight in the model as to the density or spread of the population of interest. Efford (2004) introduced a simple way of including density in the model, expanded upon in Efford et al. (2009). A summary is presented here.

Assume every individual in the population of interest has some unknown centre of activity in a region. This will be denoted $\mathbf{s}$ and is a point in euclidean space. Then traps are spread at known locations throughout the region and can capture at most one individual. The probability of an individual i being caught in trap j is then a function of the distance between the trap and the centre of activity of the individual. A well used distance function is the half normal (Efford, 2004; Royle et al., 2009), and assuming some baseline encounter rate λ_0, the capture probability of individual i in trap j is

$$P_{ij} = \lambda_0 g_{ij} = \lambda_0 \exp\left(\frac{-d_{ij}^2}{\sigma}\right)$$

where d_{ij} is the distance between individual i's centre of activity and the location of trap j. The σ here is a tuning parameter which serves as a measure of how far away from a centre of activity an individual will go. It it then proposed that the observations (population size, capture probability and distance between activity centre and trap) are a function of the parameters (density, σ, and λ_0) plus some noise (assumed multivariate normal with mean zero). The function thus represents the sampling process for a particular layout of traps. An estimate of the parameters is then achieved by simulation and inversion.

3.1 Example: with `ipsecr` in `R`

Data for snowshoe hares (*Lepus americanus*) was obtained in 1972 north of Fairbanks, Alaska. One hundred traps were set on a grid 200ft apart, with traps being done on 9 consecutive days. However, traps were unbaited on the first three days and thus the final six days are considered here. The data is available here. Over the days, 145 detections were made of 68 distinct animals. The movements can can be seen roughly across the 100 detectors in Fig. 1.

Then the function `ipsecr.fit` fits a model to the data using inverse prediction. A core assumption of this method is that, on a small scale around the trap, the data behaves linearly. Initial input values are n; the total number of distinct individuals detected, p_t; a non spatial detection probability, and σ; the scale of movement are required to start. A good starting

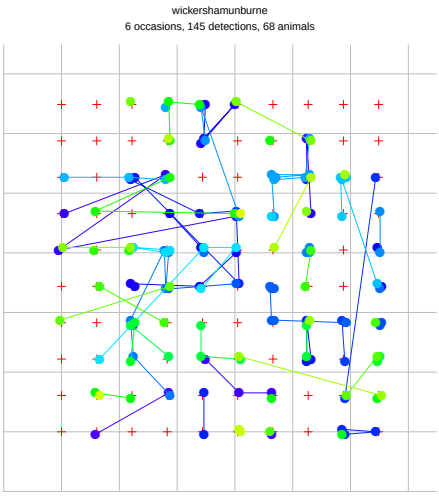

Figure 1: The movement of individuals across the traps. Some colours were reused in the plotting.

estimate for σ can be found in Calhoun and Casby (1958, eq. 8a) or Efford (2023) and is effectively a pooled standard deviation of the capture locations weighted by how frequently a trap was effective in captures. Let d_i denote the number of captures for individual i, and the trap locations $\mathbf{s}_j = (x_j, y_j)$. Then (x_{ij}, y_{ij}) are the coordinates of the j^{th} capture of the i^{th} animal. This individual i will have an average capture location $(\bar{x}_i, \bar{y}_i)$. Then

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n} \sum_{j=1}^{d_i} (x_{ij} - \bar{x}_i)^2 + (y_{ij} - \bar{y}_i)^2}{2 \sum_{i=1}^{n} (d_i - 1)}}$$

serves as a naive estimate for the spatial parameter.

The method is then well described by Efford (2023) in *`ipsecr`: An R package for awkward spatial capture–recapture data*:

> "Datasets are simulated for known levels of the parameters in the model and a vector of proxy statistics is computed from each dataset (a proxy is a measure closely correlated with the parameter of interest). A multivariate multiple linear regression [...] is fitted with parameter vectors as the predictor variable and proxy vectors as the response variable. From the inverted linear model we can infer parameter values from a single proxy vector."

Fitting this model to the snowshoe hares then gives an estimate of 1.37 hares per 0.37 hectares, which equates to 3.68 per hectare. Given the space is $\approx$ 30ha, the final population estimate is 110.69 (or 111, rounded to the nearest hare), based on the averages. Full code for this can be found in Appendix A.

3.2 Example: Tigers in Nagarahole National Reserve

Figure 2: A tiger caught on camera by a photo trap in Nagarahole Reserve (Royle et al., 2009). Photo credit: WCS/K. U. Karanth.

One of the features of the model as proposed by Efford (2023) is that, since individuals can only be caught by a trap once per capture occasion, the parameter estimates are a convolution of the actual probabilities of being caught according to location and the relative competition between individuals. This issue is circumnavigated by Royle et al. (2009) in their application to Bengal tigers (*Panthera tigris tigris*, Fig. 2) in the Nagarahole National Reserve in Karnataka, India. In this experiment, the traps were cameras and the individual tigers could be identified by their unique stripes, a trap could catch multiple individuals in a capture occasion, and an individual could be caught by multiple traps in the same capture occasion. Therefore in this setting capture frequency y_{ijt} was modelled using a Poisson random variable

$$y_{ijt} \sim \text{Po}(\lambda_0 g_{ij})$$

and thus the probability of individual i being observed by trap j during capture occasion t could be modelled using a Bernoulli random variable, where

$$P(y_{ijt} = 1) = 1 - \exp(-\lambda_0 g_{ij}),$$

arising from the positive mass of the Poisson distribution. Then Royle et al. (2009) adopt a hierarchical Bayesian framework; first assume that the centres of activity $\mathbf{s} \sim Unif(S)$ are uniformly spread over the available state space S. Then data augmentation (Royle et al., 2007) is used. This is where $M - D$ all zero observations are added to the observation matrix, with $M > N$. The next step is to assume a binomial prior on $N \sim \text{Bi}(M, \psi)$ and $\psi \sim \text{Unif}(0, 1)$. In an ecological sense, M is the size of the "super-community" of the population, and ψ is the probability that an individual in super-community is in the local community and thus had a chance of being observed. Then MCMC can be used to yield a posterior distribution for $\mathbf{s}$. Fig.

3 (Royle et al., 2009) shows the estimated density of the Bengal tigers. The posterior mean was 13.4 tigers per 100km^2, with 95% confidence interval (9.3, 19.6). This summary doesn't include the full detail where there was use of generalised linear models and also inclusion of a behavioural aspect in the capture probability, for the full detail see Royle et al. (2009).

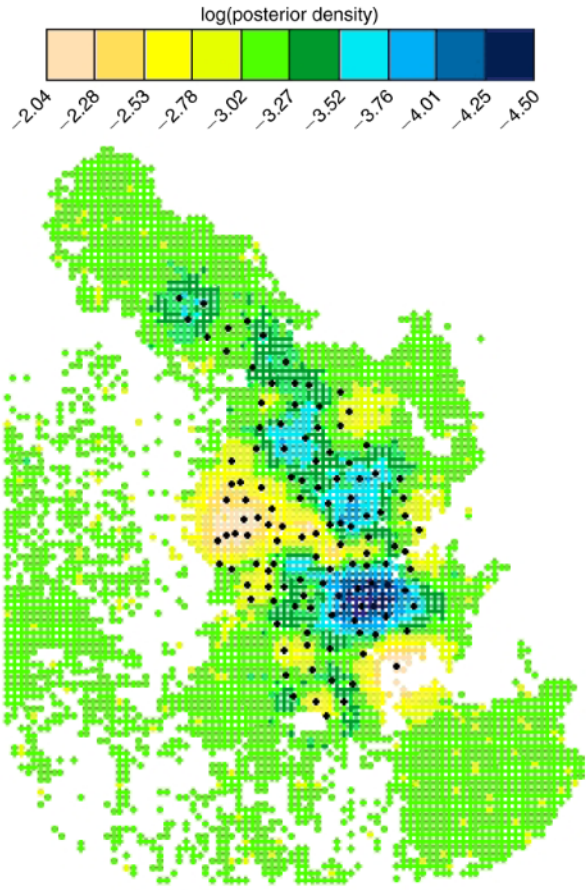

Figure 3: Posterior Distribution of activity centres $\mathbf{s}$. The scale is logged, with red and yellow being areas of higher density and blues being areas of lower density. Areas in white were removed from the state space S prior to computation as they were areas known to be unsuitable habitat.

One drawback on the model is the significant area of the state space where there were no traps placed. The initial assumption of uniform density then has no data with which to update it's parameters, resulting in large areas of middling density and little way of assessing the accuracy of the model in those places. Initial trap placement was led by workers who were familiar with the park, and so attempted to put them in places that were likely to get observations. Since a limited number of traps could be placed, a trade off has to be made between putting cameras where there is known to be little activity and confirming that against putting traps in areas of high traffic to get more observations. A smaller number of observations leads to less power in the model, and this trade off is delicate.

4 Jolly-Seber Model

The assumption of closure for population models can be shown to be violated in some cases (see §5), and it is obvious in an intuitive sense why such models might not be fully realistic. Jolly-Seber models assume an open population and thus allow for estimating another parameter of interest; **survival probability.** The model takes into account individuals entering and leaving the population. One drawback however is that it does not allow for a meaningful separation between animals entering/leaving due to birth/death versus migration. Therefore, emigration away from the population is assumed permanent. The parameters of the model are detailed in Table 1, and the observed values in the model are detailed in Table 2;

Symbol	Range	Definition
M_t	$t \in [T]$	The number of individuals that have been marked when the t^{th} sample is done. Therefore $M_1 = 0$.
N_t	$t \in [T]$	The total number of individuals in the population when the t^{th} sample is done.
B_t	$t \in [T-1]$	The number of new individuals that enter the population between the t^{th} and $(t+1)^{\text{th}}$ samples (that are still in the population at the $(t+1)^{\text{th}}$ time).
ϕ_t	$t \in [T-1]$	The probability that an individual survives between the t^{th} and $(t+1)^{\text{th}}$ sample. This is assumed the same for all individuals in this interval.
p_t	$t \in [T-1]$	The capture probability for all animals in the t^{th} sample.

Table 1: Parameters in the Jolly-Seber Model (Pollock et al., 1990).

The assumptions under which this model are valid are;

1. Every individual in the population at the time of sample t has the same capture probability, p_t.
2. Marks are never lost by the individual, or missed by the observer taking the sample.
3. Samples are taken at an instant in time.
4. Every individual that is observed during sample t has the same survival probability up until the next sample.

It is worth noting that the first three assumptions have been underlying the earlier models discussed (except where explicitly refuted). In particular, they are required for the Lincoln-Petersen estimator.

Symbol	Range	Definition
m_t	$t \in [T]$	The number of individuals in the t^{th} sample that have been previously captured and marked. u_t is the equivalent unmarked individuals.
$n_t = m_t + u_t$	$t \in [T]$	The number of individuals caught in the t^{th} sample.
$R_t \leq n_t$	$t \in [T-1]$	The number of individuals released in the t^{th} sample. This may differ from n_t as some losses may occur in the recapture attempt.
r_t	$t \in [T-1]$	How many individuals released in sample t that are caught again.
z_t	$t \in [T-1] \backslash 1$	How many individuals were caught prior to sample t, and were caught in a sample after t, but were not captured in the t^{th} sample.

Table 2: Observed values in the Jolly-Seber Model (Pollock et al., 1990).

4.1 Survival Probability

The survival probability can be estimated by considering the ratio between the number of animals marked in sample $t+1$, and the number of marked animals in the population after sample t is completed. This is given by $M_t - m_t + R_t$, the sum of the marked individuals *not* captured in sample t, which is $(M_t - m_t)$, and the number of individuals marked and released in sample t, R_t. Therefore

$$\hat{\phi}_t = \frac{\hat{M}_{t+1}}{\hat{M}_t - m_t + R_t}$$

is the survival probability estimate which naturally leads to a recruitment estimate of

$$\hat{B}_t = \hat{N}_{t+1} - \hat{\phi}_t(\hat{N}_t - n_t + R_t).$$

These are all applicable for the interval $(t, t+1)$. As Pollock et al. (1990) explain, this estimate is the difference between the population size at the $(t+1)^{\text{th}}$ sample and the expected number of survivors in the time period.

Estimates of the capture probability can then be made using the same assumption as with the Lincoln-Petersen estimate about ratio of marked individuals to total individuals caught at time t. Therefore

$$\hat{p}_t = \frac{m_t}{\hat{M}_t} = \frac{n_t}{\hat{N}_t}.$$

Since the population is open, the number of marked individuals in the population at time t is unknown, and therefore must be estimated. This is done by seeing that the future recovery rates for the animals marked but not seen at t is equivalent to the recovery rates of the individuals marked at time t and released for possible recapture

$$\frac{z_t}{M_t - m_t} = \frac{r_t}{R_t} \implies \hat{M}_t = m_t + \frac{z_t R_t}{r_t}.$$

These estimators are guided by intuition, though $\hat{\phi}_t$ and $\hat{p}_t$ do turn out to be maximum likelihood estimators. Unfortunately, all of them are biased (Pollock et al., 1990). Non-biased estimates can be found in Seber (1982) for $\hat{\phi}_t, \hat{M}_t, \hat{B}_t, \hat{N}_t$ and Jolly (1982) for $\hat{p}_t$.

5 Closure Tests

One of the challenges of the closed versus open population models is knowing when each is appropriate. In may cases the assumptions of a closed population, whilst not strictly true, are not violated enough to be a problem. As in the case of of the tiger study discussed in §3.2 (Royle et al., 2009), since the time period of the captures was 48 days, which is relatively short compared to a tigers life span, there isn't any meaningful impact on the overall population count due to birth or death. Furthermore, due to the focus of estimating the population within the borders of the national park, immigration isn't an important factor. Therefore modelling the population with a closed model still produces meaningful results. However this isn't always the case, particularly with species known for travelling vast distances or for having short life spans. In this case, the open population models are more appropriate.

Stanley and Burnham (1999a) outline a hypothesis test for closure by comparing M_t, the temporal based closed population model, against the Jolly-Seber model. This is based on the idea that M_t is just a specific case of the Jolly-Seber model, so it is just a restricted version of the parameters, as there is no recruitment. The test expands upon the goodness of fit test of M_t proposed by Stanley and Burnham (1999b).

6 Open Challenges

One of the main assumptions that underlies every model discussed here is that of independence of individual captures. However, there may be cases where this assumption is violated. For example, many species maintain monogamous relationships for a time. Therefore, it may be that the capture probabilities are not independent; if they always travel together, capturing one nearly guarantees capturing the other. In the other extreme, if one individual is always in a nest, den or 'home' where they are looking after their young, then capturing one half of the pair means you may be unlikely to capture the other half of the pair. Whilst there may be workarounds for this in species where the monogamy is time specific (e.g. avoiding taking captures in mating season), this is not a workable assumption when animals are known to have life long pairings. There may be many other factors affecting capture independence as well, and the lack of robust statistical models that take this into account leaves an open area of research.

Another challenge is the effective use of integrated population models. Capture-recapture is a well studied area, however other methods of observing populations are similarly popular. Already mentioned was distance sampling, but there is also eDNA, quadrants and ring recovery. Ring recovery has some similarities with capture-recapture, though differs in that individuals are only recovered dead. eDNA are 'scraps' of DNA recovered from habitats where an individual

has been, for example skin cells that get left on trees. It is then hoped that DNA can then be extracted from the cells and the individual uniquely identified.

There is also the advancement in citizen science data sets. These are when members of the public are invited to submit their observations of a population to a central data set. This results in a data set of significant size, which is not common in many other ecological data sets, particularly if the population of interest is endangered. This has it's own drawbacks, with the issue of data quality being significant (Johnston et al., 2023). In citizen science projects where the goal is to report on multiple species (for example, the Big Butterfly Count, https://bigbutterflycount.butterfly-conservation.org/), psychological factors around people being overly eager to report particularly rare species, or getting better at identifying species as the study progresses (Sharma et al., 2019), makes population estimates over time periods hard. There can also be spatial bias around members of the general public self selecting areas of known biodiversity or outstanding natural beauty (Tulloch et al., 2013). However, the sheer size of the data sets make them attractive to analyse, not to mention the publicity it does for the important issue of conservation.

A big step forward was to realise that often multiple types of data sets can be achieved on the same species. These are known as integrated population models (IPMs) and can combine the various data types to gain more accurate estimates or be able to estimate parameters that are otherwise conflated within the model. IPMs have their own set of open challenges including overlap and dependence between data sets, computational challenges and model evaluation such as goodness-of-fit tests (Frost et al., 2023). Particularly when reliable data sets are combined with citizen science data sets and therefore the accuracy of the data is not consistent across the whole model and some sort of assessment of this is needed. There is also the challenges of making these more complicated models widely usable to the larger science community in general so that they can be applied by those with less in depth statistical training.

Another interest in statistical ecology is in regards to forecasting. There is a desire to predict how population sizes will look in the future as a way of guiding conservation efforts. However, particularly for forecasts further into the future, imprecise parameter estimates have a cumulative effect on bad estimates. Therefore large data sets are needed to get good estimates. This can be hard to do in capture-recapture settings and therefore IPMs can utilised. Particularly by combining ecological data sets with climate data sets, good inference on the relationship between a population and the climate may be possible. Consequently this allows predicting how a species will respond to the growing effects of climate change. However, in the current setting these models rely on the assumption that a species will respond to future events in the same way as past events. If a data set on an animal population was taken over a relatively small time frame, it is harder to believe that this is representative of the populations on going response, and non-stationary models are needed to be able to deal with the potential dynamics of the situation Frost et al. (2023).

7 Conclusion

Capture-recapture data sets are some of the oldest in statistical ecology. In this report there's been an overview of the most basic capture-recapture models in closed populations, with an extension into including an explicit spatial aspect. Two examples were given with a look at both a computational Frequentist approach and an example with a Bayesian framework. Then there was a summary of the open population Jolly-Seber model and a brief look at a hypothesis test for closure within a data set. §6 shows how capture-recapture fits within the current body of research within statistical ecology. These models can provide a good starting point within integrated population models, which is an emerging field of importance. A current barrier is the lack of approachable software that allows IPMs to have a broad impact on the wider community of conservationists, which include a broad swathe of people with a background in biology or earth sciences as opposed to statistics. Making these models accessible, reliable and explainable could have a broad impact on the field.

References

Buckland, S., Anderson, D. R., Burnham, K. P., Laake, J. L., Borchers, D. L., and Thomas, L. (2001). *Introduction to Distance Sampling.* Oxford University Press, Oxford.

Calhoun, J. B. and Casby, J. U. (1958). Calculation of home range and density of small mammals. *Public Health Service Publication*, 1(592).

Chao, A. (1987). Estimating the population size for capture-recapture data with unequal catchability. *Biometrics*, 43(4):783–791.

Cohen, A. C. (1960). Estimating the parameter in a conditional poisson distribution. *Biometrics*, 16(2):203–211.

Efford, M. (2004). Density estimation in live-trapping studies. *Oikos*, 106(3):598–610.

Efford, M. G. (2023). `ipsecr`: An R package for awkward spatial capture–recapture data. *Methods in Ecology and Evolution*, 00:1–8.

Efford, M. G., Dawson, D. K., and Borchers, D. L. (2009). Population density estimated from locations of individuals on a passive detector array. *Ecology*, 90(10):2676–2682.

Fewster, R. M. and Jupp, P. E. (2009). Inference on population size in binomial detectability models. *Biometrika*, 96(4):805–820.

Frost, F., McCrea, R., King, R., Gimenez, O., and Zipkin, E. (2023). Integrated population models: Achieving their potential. *Journal of Statistical Theory and Practice*, 17(1):6.

Johnston, A., Matechou, E., and Dennis, E. B. (2023). Outstanding challenges and future directions for biodiversity monitoring using citizen science data. *Methods in Ecology and Evolution*, 14(1):103–116.

Jolly, G. M. (1982). Mark-recapture models with parameters constant in time. *Biometrics*, pages 301–321.

McCrea, R. S. and Morgan, B. J. T. (2015). *Analysis of Capture-Recapture Data*. Interdisciplinary statistics. CRC Press, Boca Raton.

Otis, D. L., Burnham, K. P., White, G. C., and Anderson, D. R. (1978). Statistical inference from capture data on closed animal populations. *Wildlife Monographs*, 1(62):3–135.

Pollock, K. H., Nichols, J. D., Brownie, C., and Hines, J. E. (1990). Statistical inference for capture-recapture experiments. *Wildlife Monographs*, 1(107):3–97.

Royle, J. A., Dorazio, R. M., and Link, W. A. (2007). Analysis of multinomial models with unknown index using data augmentation. *Journal of Computational and Graphical Statistics*, 16(1):67–85.

Royle, J. A., Karanth, K. U., Gopalaswamy, A. M., and Kumar, N. S. (2009). Bayesian inference in camera trapping studies for a class of spatial capture–recapture models. *Ecology*, 90(11):3233–3244.

Seber, G. A. F. (1982). *The estimation of animal abundance and related parameters*, volume 8. Blackburn press Caldwell, New Jersey.

Sharma, N., Colucci-Gray, L., Siddharthan, A., Comont, R., and Van der Wal, R. (2019). Designing online species identification tools for biological recording: the impact on data quality and citizen science learning. *PeerJ*, 6:e5965.

Stanley, T. R. and Burnham, K. P. (1999a). A closure test for time-specific capture-recapture data. *Environmental and Ecological Statistics*, 6:197–209.

Stanley, T. R. and Burnham, K. P. (1999b). A goodness-of-fit test for capture-recapture model Mt under closure. *Biometrics*, 55(2):366–375.

Tulloch, A. I., Mustin, K., Possingham, H. P., Szabo, J. K., and Wilson, K. A. (2013). To boldly go where no volunteer has gone before: predicting volunteer activity to prioritize surveys at the landscape scale. *Diversity and Distributions*, 19(4):465–480.

A Appendix A

```
install.packages("secr") #installing packages only needs to be done once
install.packages("ipsecr")
library(ipsecr) # this automatically loads secr

capt <- read.table("hareCH6capt.txt") # read in the capture data for the hares
capt <- capt[,2:4] # remove the first column as it just has the data source
colnames(capt) <- c("ID", "Occasion", "Detector") # better names for the columns

trap <- read.table("hareCH6trap.txt") # load the trap layout for the hares
colnames(trap) <- c("Detector","x", "y") # better names for the columns

# just to see what the trap layout looks like without the animals data
library(ggplot2) # for a nice plot
ggplot(trap, aes(trap[,2],trap[,3])) + geom_point() + labs(x="x coordinate (
    metres)", y="y coordinate (metres)")
# it is just a grid, as expected

# convert data into the data type for the function
hareCH6 <- read.capthist("hareCH6capt.txt", "hareCH6trap.txt", detector = "
    single")
summary(hareCH6)
# Counts by occasion
#                    1   2   3   4   5   6 Total
# n                 16  28  20  26  23  32   145
# u                 16  24   9   9   6   4    68
# f                 25  22  13   5   1   2    68
# M(t+1)            16  40  49  58  64  68    68
# losses             0   0   0   0   0   0     0
# detections        16  28  20  26  23  32   145
# detectors visited 16  28  20  26  23  32   145
# detectors used   100 100 100 100 100 100   600

# plot the traps with animal movements on top
plot(hareCH6, tracks = TRUE)

initialsigma <- RPSV(hareCH6, CC = TRUE) # quick (biased) estimate of sigma is
    63.66117 m

# use a buffer of four, as the half normal detection function generates a close
# to zero probability of being observed that far away
fit <- secr.fit (hareCH6, buffer = 4 * initialsigma, trace = FALSE)
# can plot the detection probability
plot(fit, limits = TRUE)


ipfit <- ipsecr.fit (hareCH6, buffer = 4*initialsigma, trace = FALSE)
summary(ipfit)
```

```
1.36652240/(60.96*60.96*0.0001) # = 3.677281 per hectare
548.64*548.64*0.0001 # = 30.10058 hectares
3.677281*30.10058 # = 110.6883 hares in the area
```