Statistics with Data Science Project 5

Investigating Demographic Parameters in Birds: Effect of Fitness, Species' Interaction and Climate Change

1 Project Team

The project team for this Statistics with Data Science (SwDS) research project is as follows:

External expert: Dr Blanca Sarzo, Margarita Salas Fellow, University of Valencia;

Project leader: Prof Ruth King;

Project helper: Dr Cecilia Balocchi;

Project assistant: Ms Nina Fischer.

2 Project Organisation

Support for the project will be organised as follows:

Initial information session: 14.00 Monday, 10th July, Nucleus: Elm Lecture Theatre.

Drop-in sessions: The drop-in sessions will be the primary opportunity for asking for advice in relation to the project. There will be a total of 9 drop-in sessions. These will be primarily held in JCMB room 1206C (with one exception on the 24th July). The sessions will be held from 10.00-11.00 on the following days (room in bracket for the 24th July):

  Week 1: Thursday 13th July;

  Week 2: Monday 17th July; Thursday 20th July;

  Week 3: Monday 24th July (Room: JCMB 3217); Thursday 27th July;

  Week 4: Monday 31st July; Thursday 3rd August;

  Week 5: Monday 7th August; Wednesday 9th August.

Piazza: This on-line discussion forum will be available outside of the main drop-in sessions to supplement the advice provided in the drop-in sessions. Piazza will be monitored regularly, but it is encouraged and expected that other students in the project group will contribute to posted queries.

To sign-up for the project Piazza page use:

    https://piazza.com/ed.ac.uk/summer2023/dissertationsds202220235

The link to the Piazza page for the project is:

    https://piazza.com/ed.ac.uk/summer2023/dissertationsds202220235/home

Contact E-mail: For any other queries, please contact Prof Ruth King: Ruth.King@ed.ac.uk.

Deadline: The deadline for submitting your project is 12.00 on Friday 11th August.

3 Project Details

3.1 Introduction

Over the last century, many wildlife species have significantly declined and many others face potential extinction due to threats such as climate change, invasive species, and illegal hunting, among many other factors. There are many underlying factors that may affect demographic parameters, and hence, understanding the factors affecting wildlife populations allow us to improve the management and conservation of animal populations. In particular, within this project we will focus on passerine bird species. For such bird population, factors such as fitness (understood as the ratio weight/size) and environmental conditions at breeding grounds or wintering areas may affect survival probabilities. Further, the species do not live in isolation but may interact, for example, by competing for the same resources. In recent decades, conservation research has considered both an individual species-level approach considering different species in isolation as well as developing additional multi-species approaches that permit a more intricate understanding of the ecosystem of interest.

This project will focus on capture-recapture studies, which involve collecting data by repeatedly sampling the population under study with the aim of obtaining inference on the underlying ecological processes of interest (McCrea and Morgan, 2014; King, 2014; Seber and Schofield, 2019). In particular, we will focus on the long-term capture-recapture database (1997-2020) of three different bird species wintering in Valencia (Spain): blackcap, chiffchaff and robin.

3.2 Data

Data are collected on three species: blackcap, chiffchaff and robin. The data are collected at the wintering grounds of these species in Southern Spain (monthly visits from October to April). Around spring, the birds migrate to breeding areas elsewhere in Europe, returning to the wintering grounds in the autumn. We initially discuss the data collection process before describing the mathematical notation used for the data, data files and finally some ecological aspects of the data that are potentially of interest, in general.

3.2.1 Data collection

For capture-recapture studies, the data collection protocol consists of observers going into the field at a series of capture events (or occasions), denoted $t = 1, \ldots, T$. At each capture event, previously unobserved individuals are uniquely marked (in our case a ring is applied to the leg of an individual with a unique ID code), previously observed individuals are recorded, and all are released back into the population. The data then correspond to the capture histories of each individual observed throughout the study, detailing at which capture occasions they are (or are not) observed. In practice, for general capture-recapture studies additional information may also be recorded on capture, including for example, age, sex, weight, breeding status etc. Importantly, we note that the capture process is imperfect and individuals in the study at a given time may not be observed at that time.

For the particular capture-recapture study that we will consider, standardized protocols were used to observe the birds in a crop area located in Valencia (Spain; coordinates: 39°41'52.0"N 0°14'11.9"W) for each year of the study period from 2007 to 2018. For each year, from 2007 to 2018, monthly capture occasions are defined from October to April (i.e. data are collected from October in 2007 to April in 2018). Note that the winter period from October in year j to April in year $j + 1$ is referred to as the winter in year j. The years are typically referred to as the primary capture occasions; and the months within each year

as the secondary capture occasions. At each capture occasion, nets were used to capture the birds with the study area. The same number of nuts and net placements were used throughout for all capture occasions.

Notes:

1. Inter-winter refers to between primary capture occasions (i.e. across years); so that, for example, inter-winter survival corresponds to annual survival.

2. Intra-winter refers to within a single primary capture occasion, (or between the secondary capture occasions within that primary capture occasion); so that intra-winter survival corresponds to weekly/monthly survival.

3. Additional information about the age of the individual at their first capture may also be recorded, in terms of juvenile (first year of life), adult (second year of life onwards). However, for some individuals, this information is not recorded and hence is unknown (or missing).

3.2.2 Data description

The data correspond to the capture histories for each individual observed within the study (i.e. at which capture occasions each individual is observed). For simplicity we label the monthly capture occasions over all years as the capture occasions $t = 1, \ldots, T$. There are a total of 7 capture occasions per year and 7 years, giving a total of $T = 49$ capture occasions. Notationally, for each individual, $i = 1, \ldots, n$; and capture occasions $t = 1, \ldots, T$, we let,

$$x_{it} = \begin{cases} 0 & \text{individual } i \text{ is not observed at time } t; \\ 1 & \text{individual } i \text{ is observed at time } t; \end{cases}$$

The capture history for individual $i = 1, \ldots, n$ is denoted $\boldsymbol{x}_i = \{x_{it} : t = 1, \ldots, T\}$. For example, consider $T = 7$ (i.e. 7 capture occasions), the following capture history:

$$1 \quad 0 \quad 1 \quad 0 \quad 0 \quad 0 \quad 0$$

denotes an individual observed at capture occasions 1 and 3; and unobserved at capture occasions 2, 4, 5, 6 and 7. The set of capture histories over all individuals is denoted $\boldsymbol{x} = \{\boldsymbol{x}_i : i = 1, \ldots, n\}$.

We note that the construction of the capture histories assumes that individuals are uniquely identifiable following initial capture (for the birds, a unique ring is applied to the leg of each bird on initial capture); there is no loss of the unique identifier; and that all identifiers are correctly read and matched across capture occasions. In other words we assume that there is no mis-identification across individuals and/or times.

3.2.3 Data summary

The data for each species are provided in the form of a "*CR.birdname_FixRing.csv*" file, (where *birdname = blackcap/chifchaf/robin*) detailing the monthly capture events for each individual for the duration of the study period. The first row of the spreadsheet contains the header information detailing the capture occasion in terms of a combined number corresponding to the year + month (e.g. 200701 corresponds to January 2007); with the final term relating to the age of the bird at initial ringing. Each subsequent row of the spreadsheet corresponds to a unique individual observed within the study with associated data entries

corresponding to their given capture occasion (0 = not observed at the given capture occasion; 1 = observed) and associated age at initial ringing (juvenile/adult). The total number of ringed individuals in the work database is: (i) blackcap - 861; (ii) chiffchaff - 581; and (iii) robin - 447 birds.

3.2.4   Ecological context

Ecological data are collected in order to obtain insight into the ecological processes of interest. This may include, for example, abundance (both in terms of the absolute number as well as potential trends over time), and demographic quantities such as survival probabilities or fecundity rates. Such quantities are often of particular interest as they not only provide information regarding the ecosystem of interest but also inform conservation management. Further, the impact of climate change on species is becoming increasingly important to assess, with rising global temperatures and changes to the onset of seasons. For many bird species, the changing environmental conditions may lead to changes in migration times and locations such that they may migrating earlier/later in the year and replace previous winter grounds with closer areas.

For the bird species within this project, mortality during the winter period is typically one of the main causes regulating the population. There are many factors that may influence survival: age, individual quality, environmental conditions (e.g. precipitations, temperature), predators and food resources (among many others). For all three species (and many bird species in general), survival within the first year of life is often lower than for older birds (first year birds are referred to as juveniles; older birds as adults). This implies that age is likely to be an important factor to include if investigating the survival of individuals over years (i.e. age is likely to be an important covariate for survival probabilities). The species only interact with each other in Spain during winter time, as birds migrate long distances in the spring to their breeding areas elsewhere in Europe (precise locations are not known) before returning to their wintering grounds in the autumn. Thus there are external ecological reasons for varying inter-winter survival across species (due to varying environmental conditions during the breeding season in the summer months) in addition to species variation. However, the environmental conditions for the species are similar at the breeding grounds, and so any differences identified relating to intra-winter survival are more likely to be due to other factors (or species differences). For example, two of the three species (blackcap and chiffchaff) are not territorial to their winter grounds, i.e. birds may either remain within a small area (*resident* strategy) or move frequently over a large area looking for locally abundant food (*transient* strategy) (Cuadrado *et al.*, 1994; Tellería and Pérez-Tris, 2007). Further, effects of individual quality on survival (often referred to as individual heterogeneity) as well as multi-species relationships can often be a factor of interest (see for example, Lahoz-Montfort *et al.*, 2011).

3.3   Statistical models

We provide a description of some of the basic (standard) models for capture-recapture data. In particular we focus on the derivation of the likelihood function of the data, given the parameters for these models. The likelihood can then be maximised to obtain the maximum likelihood estimates (MLEs) of the parameters within a classical analysis (see Section 14 of Statistical Programming notes for how to do this in R using the inbuilt functions `optim` and `nlm`), or combined with the associated prior specification of the models to form the posterior distribution within a Bayesian analysis (see Bayesian Theory and Bayesian Data Analysis). The data are discrete valued (observed/not observed), and so the likelihood function is the

probability of observing the given data (i.e. capture histories) given the parameter values. We specify the likelihood as $f(\boldsymbol{x}|\boldsymbol{\theta})$, where $\boldsymbol{\theta}$ denotes the parameters for the given model. We note that the associated models fitted to capture-recapture data account for imperfect detection through the estimation of detection probabilities, providing the link between what we observe and the true underlying state of the system. The other parameters within the model depends on the specific model.

We consider two separate categories of models dependent on whether the population is assumed to be closed or open or over the duration of the study period:

Closed populations: These assume that the population is constant throughout the study period, so that there are no births, deaths, or migrations. In other words all individuals in the population are present in the study area and available for capture at each capture occasion.

Open populations: These assume that the population may change within the study period, with individuals arriving into the population (via birth or immigration) or exiting the population (via death or emigration). Thus not all individuals in the population may be available at each capture occasion.

We consider models for both closed and open populations.

3.3.1 Closed populations

In general, populations within the study will change over time. However, if the duration of the study period is short (e.g. one week or one month), it may be assumed that the population is (approximately) closed over the study period. In other words the number of births, deaths or migrations will be very small and hence negligible.

For closed populations, the quantity of primary interest is usually the total population size, which we denote by N. However, due to imperfect capture, we also need to take into account the associated capture probabilities within the study. In particular, for $t = 1, \ldots, T$ we define:

$p_t = \mathbb{P}(\text{an individual is observed at time } t).$

Here, we have indexed the recapture probability to be potentially dependent on time t. We discuss the capture probabilities further below in terms of what the terms may be dependent on. We let $\boldsymbol{p} = \{p_t : t = 1, \ldots, T\}$. The set of model parameters is then $\boldsymbol{\theta} = \{N, \boldsymbol{p}\}$. The recapture probabilities, $\boldsymbol{p}$ are essentially nuisance parameters, with N the parameter of primary interest. However, the dependence specified on the recapture probabilities will typically influence the estimate of N and hence is important to consider.

We assume that individuals behave independently of each other, and so the associated likelihood function of the data can be written in the form:

$$f(\boldsymbol{x}|\boldsymbol{\theta}) \quad \propto \quad \frac{N!}{(N-n)!} \underbrace{\left[\prod_{i=1}^{n} f(\boldsymbol{x}_i|\boldsymbol{\theta})\right]}_{\substack{\text{Observed}\\\text{individuals}}} \times \underbrace{[f(\boldsymbol{0}|\boldsymbol{\theta})]^{N-n}}_{\substack{\text{Unobserved}\\\text{individuals}}},$$

where $f(\boldsymbol{x}_i|\boldsymbol{\theta})$ denotes the probability of observing capture history $\boldsymbol{x}_i$ given the model parameters; $f(\boldsymbol{0}|\boldsymbol{\theta})$ is the probability of not observing an individual; and the first term corresponds to the Multinomial coefficient the same capture histories are interchangeable with each other.

Considering capture history, $\boldsymbol{x}_i$, the associated probability of the history is simply whether or not the individual is observed or not at each capture occasion,

$$f(\boldsymbol{x}_i|\boldsymbol{\theta}) = \prod_{t=1}^{T} p_t^{x_{it}}(1 - p_t)^{1-x_{it}}.$$

In other words we consider each capture occasion and observe an individual ($x_{it} = 1$) with probability p_t; or fail to observe the individual ($x_{it} = 0$) with probability $(1 - p_t)$.

Similarly for an individual that is not observed (and so has capture history $= \boldsymbol{0}$) is given by,

$$f(\boldsymbol{0}|\boldsymbol{\theta}) = \prod_{t=1}^{T}(1 - p_t).$$

The model described above is often called model M_t, where the subscript denotes the dependence of the capture probability on time (i.e. temporal heteroegeneity). Alternative models include:

M_0: constant model where $p_t = p$ for all t;

M_b: behavioural heterogeneity where $p_t = p$ for initial capture; and $p_t = c$ for subsequent recaptures;

M_h: individual heterogeneity with capture probabilities p_i for $i = 1, \ldots, N$.

See for example, Otis *et al.* (1978); Borchers *et al.* (2002); King and Brooks (2008); King and McCrea (2019) for further discussion. Models may be described that have multiple dependencies (e.g. M_{tb}: temporal and behavioural heterogeneity). Note that individual heterogeneity is the most complex model and additional assumptions are required to permit estimation of the parameters and total population size. For example using a finite mixture model (so that $p_i = p_{d(i)}$ for $d(i) \in \{1, ..., M\}$, where M is to be defined/estimated but is usually small; Pledger (2000)); or (continuous) infinite mixture model (so that $p_i \sim N(\mu, \sigma^2)$, where μ and σ^2 are to be estimated - in this case the likelihood is expressed as an analytically intractable likelihood; Gimenez and Choquet (2010)).

Computer code

Sample `R` code is provided for the log-likelihood for model M_t in Appendix A.1. The code is also provided in the file `closedlik.R` with an example (simulated) dataset in the file `closed_sample.data` in Learn.

3.3.2 Open populations

Open populations permit the population to change over time with new individuals entering or leaving the study during the study period. This will often be the case when survey periods are over a prolonged period of time, with many capture-recapture studies spanning years, with, for example, annual capture occasions. For such studies, the statistical models developed for the data traditionally focus on the estimation of *survival probabilities*, as opposed to the changing population size over time. We will consider the most commonly used model typically referred to as the Cormack-Jolly-Seber (CJS) model that permits the estimation of survival probabilities, taking into account the imperfect detection within the studies (but due to the construction of the associated likelihood does not estimate abundance). Assuming a temporal dependence, the model parameters are defined to be:

$\phi_t = \mathbb{P}(\text{individual is alive at time } t + 1 \mid \text{alive at time } t)$ for $t = 1, \ldots, T - 1$;

$p_t = \mathbb{P}(\text{individual is recaptured at time } t \mid \text{alive at time } t)$ for $t = 2, \ldots, T$.

We let $\boldsymbol{\phi} = \{\phi_t : t = 1, \ldots, T - 1\}$ and $\boldsymbol{p} = \{p_t : t = 2, \ldots, T\}$. The set of model parameters is then given by $\boldsymbol{\theta} = \{\boldsymbol{\phi}, \boldsymbol{p}\}$. In the definition of the parameters we have assumed that they are both dependent on time (i.e. capture occasion). In practice there may be many different dependence structures that we may wish to consider, including for example, individual covariates (age, sex, weight, breeding status, etc.), environmental covariates, density dependence, behavioural effects etc.

The likelihood is constructed as the probability of the observed capture histories, conditional on their initial capture (and parameter values). Assuming that individuals behave independently of each other, we can write:

$$f(\boldsymbol{x}|\boldsymbol{\theta}) = \prod_{i=1}^{n} f(\boldsymbol{x}_i|\boldsymbol{\theta}),$$

where $f(\boldsymbol{x}_i|\boldsymbol{\theta})$ denotes the probability of capture history $\boldsymbol{x}_i$, given their initial capture. Let $f(i)$ denote the initial capture time for individual i; and $l(i)$ the final capture time for individual i. Then, we can write,

$$f(\boldsymbol{x}_i|\boldsymbol{\theta}) = \prod_{t=f(i)}^{l(i)-1} \phi_t p_{t+1}^{x_{i,t+1}} (1 - p_{t+1})^{1-x_{i,t+1}} \times \chi_{l(i)},$$

where χ_t denotes the probability of not being observed again within the study. The initial product from times $t = f(i)$ to $l(i) - 1$ represents the probability of the capture history after initial capture to their final capture, consisting of (i) surviving between each capture occasion (the ϕ_t term, since an individual once "dead" remains dead) combined with (ii) the probability of being observed, p_{t+1} (i.e. $x_{i,t+1} = 1$) or not observed, $1 - p_{i,t+1}$ (i.e. when $x_{i,t+1} = 0$). The null product is defined such that $\prod_{t=f(i)}^{f(i)-1} \equiv 0$ (this occurs if an individual is seen only once so that $l(i) = f(i)$). The probability of not being observed again within the study, given that an individual is observed at time t, is calculated recursively using,

$$\chi_t = \underbrace{(1 - \phi_t)}_{\mathbb{P}(\text{dying})} + \underbrace{\phi_t(1 - p_{t+1})\chi_{t+1}}_{\substack{\mathbb{P}(\text{survive, not seen at } t+1, \\ \text{and not seen after } t+1)}},$$

with $\chi_T = 1$. Alternatively we can use the equivalent recursion:

$$\underbrace{1 - \chi_t}_{\substack{\mathbb{P}(\text{observed} \\ \text{after time } t)}} = \underbrace{\phi_t p_{t+1}}_{\substack{\mathbb{P}(\text{survive time } t \\ \text{and seen at } t+1)}} + \underbrace{\phi_t(1 - p_{t+1})(1 - \chi_{t+1})}_{\substack{\mathbb{P}(\text{survive, not seen at time } t+1 \\ \text{but seen after } t+1)}}.$$

The above likelihood assumes that the recapture and survival probabilities are dependent on the capture occasions. These are specified as an arbitrary time dependence, but there is often interest in explaining such temporal dependence via environmental conditions, regressing the capture and/or survival probabilities on given environmental covariates. Alternative dependence structures include, for example, a constant probability (no dependence), age-dependence and individual covariates (e.g. sex, breeding status).

Notes:

7

1. The model *conditions* on the first time an individual is observed. This means that there is no initial capture probability within the likelihood, as we condition on seeing the individual at their initial capture.

2. The survival probability, ϕ_t is the *apparent* survival probability as mortality and permanent migration from the study site are confounded. Within the CJS model, if an individual dies or leaves the study site we assume that they are no longer available for capture within the study (i.e. there is no temporary migration). For simplicity we refer to all individuals that depart from the study site (and hence are no longer available for capture) as "dying" (which is an absorbing state, once an animal dies it cannot be observed again).

3. Transient behaviour is a particular form of permanent migration that can be estimated separately from mortality. In particular, transients are individuals that are not able to be recaptured following their initial capture, as they leave the study area (with probability 1). In the context of the given study, transients are birds who are not "local" to the wintering ground study site, and are likely to be searching for food away from their home territory. For more information, and statistical modelling of transients, see for example, Pradel *et al* (1997).

4. For many animal (and in particular bird) populations, the corresponding survival probabilities are often dependent on the age of an individual. In particular, a first year survival probability is often necessary to reflect that young individuals have a higher mortality rate compared to older individuals.

5. In ecological studies it is often the case that certain variables of interest have missing observations (particularly individual covariate values). If incorporating such covariate information, consideration needs to be taken regarding how to deal with such missing data.

6. The CJS model with full (arbitrary) time dependence on the recapture and survival probabilities is parameter redundant, which means that the parameters ϕ_{T-1} and p_T cannot be separately estimated (only their product $\phi_{T-1}p_T$ can be estimated). Mathematically, there is a ridge in the likelihood such that there is no unique value for ϕ_{T-1} and p_T that maximises the likelihood, and hence no unique MLE for these terms. For such a model, the parameters should be reparameterised and the product $\phi_{T-1}p_T$ estimated (as opposed to the individual terms).

More recent models (often referred to as stopover models) have been developed to additionally estimate the total population size (Pledger *et al.*, 2013). These models remove the conditioning on the initial capture probability, and require the estimation of additional arrival probability of individuals into the study site (this can be considered to be the opposite/symmetric term to the survival probability). The arrival probability is required to take into account that individuals are not available for capture prior to their arrival at the study site in the analogous way that individuals are not available for capture once they have departed from the study site, i.e. "die".

Computer code

R code is provided for the log-likelihood for the CJS model with a constant capture probability and time-dependent survival probability for a simple dataset in the file CJSlik.R on Learn. The code is also provided in Appendix A.2.

3.4 Potential research questions

There are numerous research questions of interest in relation to the given data, relating to estimation of abundance (i.e. number of individuals at a given time); or the survival probabilities of the different species. Example research areas that could be investigated include:

- Estimation of inter-winter (i.e. annual) juvenile and adult survival probabilities for the three species separately and/or jointly.

- Estimation of intra-winter (within year) survival probabilities for the three species separately and/or jointly.

- Development of approaches for jointly estimating inter-winter and intra-winter survival probabilities.

- Estimation of annual population sizes at the winter grounds.

- Investigation of whether there is evidence of individual heterogeneity in the survival and/or capture probabilities within/across years.

- Investigation of whether the different species have synchronous inter-winter survival probabilities.

The above list provides some ideas of research questions that can be considered, but this is not intended, and should not be interpreted as, an exhaustive list. The aim of the project is to interrogate the data, fitting relevant statistical models to extract ecological quantities of interest to answer a given research question or hypothesis. Note that providing a discussion of the results in the context of the given ecological data is an important aspect of the statistical analysis.

References

Borchers, D. L., Buckland, S. T. and Zucchini W. (2002) *Estimating Animal Abundance, Closed Population.* Springer.

Cuadrado, M., Senar, J.C., and Copete, J.L. (1994) Do all blackcaps *Silvia atricapilla* show winter site fidelity? *Ibis*, **137**, 70-75.

Gimenez O. Choquet R. (2010) Individual heterogeneity in studies on marked animals using numerical integration: capture–recapture mixed models. *Ecology*, **91**, 951– 957.

King, R. (2014) Statistical Ecology. *Annual Review of Statistics and its Application*, **1**, 410-426.

King, R. and Brooks, S. P. (2008) On the Bayesian estimation of a closed population size in the presence of heterogeneity and model uncertainty. *Biometrics*, **64**, 816-824.

King, R. and McCrea, R. S. (2019) Capture-recapture: Methods and Models, in *Handbook of Statistics*, Volume 40, 33-83.

Lahoz-Monfort, J. J., Morgan, B. J. T, Harris, M. P., Wanless, S. Freeman, S. (2011). A capture-recapture model for exploring multi-species synchrony in survival. *Methods in Ecology and Evolution*, **2**, 116-124.

McCrea, R. S. and Morgan, B. J. T. (2014) *Analysis of Capture-Recapture Data*. CRC Press.

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*, **62**, 1-135.

Pledger, S., Efford, M., Pollock, K., Collazo, J. A., and Lyons, J. E. (2009). Stopover duration analysis with departure probability dependent on unknown time since arrival. In *Modeling Demographic Processes in Marked Populations*, 349– 363. Springer.

Pledger S. (2000) Unified maximum likelihood estimates for closed capture-recapture models using mixtures. *Biometrics*, **56**, 434-42

Pradel, R., Hines, J. E., Lebreton, J-D., Nichols, J. D. (1997) Capture-recapture survival models taking account of transients. *Biometrics*, **53**, 60-72.

Seber, G. A. F. and Schofield, M. R. (2019) *Capture-Recapture: Parameter Estimation for Open Animal Populations*. Springer.

Tellería, J.L., and Pérez-Tris, J. (2007). Habitat effects on resource tracking ability: do wintering Blackcaps *Sylvia atricapilla* track fruit availability? *Ibis*, **149**, 18-25.

A   Appendix A: Basic computer codes

In this section some basic computer codes are provided for both the closed population model and open population model (Cormack-Jolly-Seber model). Sample data are also provided for both (in files `closed_sample.data` and `open_sample.data`. Note that these are basic models that will require development within the project, dependent on the statistical model to be fitted to the data.

A.1   Closed population

We initially consider the closed population model M_t, where the capture probabilities are dependent on the capture occasion. We assume that the capture recapture data are stored in the array x which is of dimension $(n \times T)$ (i.e. the rows represent the observed individuals; and the columns the capture occasions).

   We consider a function that calculates the log-likelihood function. We let `theta` denote the vector of parameter values. Note that it is often useful to transform the parameters to be specified such that they are on the real line. Thus for this code we transform the capture probabilities on the logistic scale; and transform the number of unobserved individuals, denoted `unobs` (as this must be positive) to be on the log scale. The log-likelihood function is specified to be a function of the parameters (`theta`); data x; number of individuals observed, `n`, and number of capture occasions, `T`. Sample code for calculating the log-likelihood (also provided in the file "`closedlik.R`" in Learn) is given by:

```
closedlik <- function(theta, x, n, T) {

  # The data are stored in the array x;
  # n = number of observed individuals; T = number of capture occasions
  # Theta stores the set of parameter values - specified on the real line.
  # Define the parameter values in terms of capture probs and population size.
  # Use the transformations: logit p = theta[1:T]; log N = theta[T+1]
```

```
    p <- exp(theta[1:T])/(1+exp(theta[1:T]))
    unobs <- exp(theta[T+1])

    N <- unobs + n

    # Initialise the log-likelihood value:

    lik <- 0

    # Calculate the (log-)likelihood component for each individual capture history

    for (i in 1:n){
      for (t in 1:T){
        lik <- lik + x[i,t]*log(p[t]) + (1-x[i,t])*log(1-p[t])
      }
    }

    # Calculate the (log) probability of not being observed within the study

    noprob <- sum(log(1-p[]))

    # Add the log-likelihood contribution of the probability of unobserved individuals

    lik <- lik + (N-n)*noprob

    # Add the Multinomial coefficient likelihood component:

    lik <- lik + lgamma(N+1) - lgamma(N-n+1)

    # Output the log-likelihood value:

    lik
}
```

A.2 Appendix B: Open population

We consider an open population and the Cormack-Jolly Seber model, with time dependent survival, ϕ_t (for $t = 1, \ldots, T-1$) and constant capture probability p. As for the above closed population case, the capture recapture data are assumed to be stored in the array $(n \times T)$ array, x. For efficiency we store the initial time that each individual is observed for the first time in the vector f; and the final capture time for each individual in the vector l.

The likelihood is expressed as a product over all observed individuals of the probability of their given capture histories, conditional on their initial capture. (This means that we do not need to consider the times before initial capture, nor the probability of initial capture). The model parameters are the recapture probability p (constant over time) and the survival probability phi[t] for $t = 1, \ldots$. However, it is often useful to transform the parameters such that they lie on the real line (for example, if we then wish to obtain the MLE of the parameters as we do not need to be concerned about constraints on possible values). We consider a logistic regression for both the recapture and survival probabilities. The log-likelihood function is specified to be a function of the parameters (theta), capture histories

(x), set of initial capture times (f), final capture times (l), number of observed individuals (n) and the number of capture occasions (T). Associated sample code (also provided in the file "CJSlik.R" in Learn) to evaluate the likelihood is:

```
CJSlik <- function(theta, x, f, l, n, T) {

  # The data are stored in the array x;
  # n = number of observed individuals; T = number of capture occasions

  # f - array corresponding to first time each individual is observed
  # f can be calculated using e.g.:
  #   for (i in 1:n){f[i] <- which(x[i,]==1)[1]}

  # l - array corresponding to last time each individual is observed
  # l can be calculated using e.g.:
  #   for (i in 1:n){l[i] <- which(x[i,]==1)[length(which(x[i,]==1))]}

  # Theta stores the set of parameter values - specified on the real line.
  # Define the parameter values in terms of capture probabs and population size.
  # Use the transformations: logit phi[1:T-1] = theta[1:T-1]; logit p = theta[T]

  phi <- exp(theta[1:(T-1)])/(1+exp(theta[1:(T-1)]))
  p <- exp(theta[T])/(1+exp(theta[T]))

  # Calculate the chi terms: probability of not observed after time t
  # Initially set up chi to be an array of length T all elements equal to 1
  # chi[T] <- 1
  # Calculate chi[1:T-1] using recursive formula

  chi <- array(1,T)

  for (t in (T-1):1){
    chi[t] <- 1 - phi[t] + phi[t]*(1-p)*chi[t+1]
  }

  # Initialise the log-likelihood value for each individual:

  lik <- array(0,n)

  # Calculate the (log-)likelihood component for each individual capture history

  for (i in 1:n){

    # Consider the times between initial capture and final capture for each individual
    # Contribution only required when l[i] is not equal to f[i]

    if (f[i] != l[i]){

      for (t in f[i]:(l[i]-1)){
        lik[i] <- lik[i] + log(phi[t]) + x[i,t+1]*log(p) + (1-x[i,t+1])*log(1-p)
```

12

```
      }
    }

    # Note the above function assumes that we do not have any observed
    # histories with a single observation at the final capture occasion
    # I.e. we no not have any histories such that f = l = T.
    # For such histories there is no information (and so can be omitted).

    # Add chi terms (probability of not being observed after time l[i])

    lik[i] <- lik[i] + log(chi[l[i]])
  }

# Calculate log-likelihood over all individuals:

  sumlik <- sum(lik)

# Output the log-likelihood value:

  sumlik
}
```