Rcapture: Loglinear Models for Capture-Recapture in R

April 2007

Sophie Baillargeon
Université Laval, Québec

Louis-Paul Rivest
Université Laval, Québec

Abstract

This article introduces **Rcapture**, an R package for capture-recapture experiments. The data for analysis consists of the frequencies of the observable capture histories over the t capture occasions of the experiment. A capture history is a vector of zeros and ones where one stands for a capture and zero for a miss. **Rcapture** can fit three types of models. With a closed population model, the goal of the analysis is to estimate the size N of the population which is assumed to be constant throughout the experiment. The estimator depends on the way in which the capture probabilities of the animals vary. **Rcapture** features several models for these capture probabilities that lead to different estimators for N. In an open population model, immigration and death occur between sampling periods. The estimation of survival rates is of primary interest. **Rcapture** can fit the basic Cormack-Jolly-Seber and Jolly-Seber model to such data. The third type of models fitted by **Rcapture** are robust design models. It features two levels of sampling; closed population models apply within primary periods and an open population model applies between periods. Most models in **Rcapture** have a loglinear form; they are fitted by carrying out a Poisson regression with the R function glm. Estimates of the demographic parameters of interest are derived from the loglinear parameter estimates; their variances are obtained by linearization. The novel feature of this package is the provision of several new options for modeling capture probabilities heterogeneity between animals in both closed population models and the primary periods of a robust design. It also implements many of the techniques developed by R. M. Cormack for open population models.

Keywords: loglinear models, mixture models, multinomial distribution, profile likelihood confidence intervals, residuals.

1. Introduction

The goal of a classical capture-recapture experiment is to study the demographic characteristics of an animal population. It is carried out by capturing animals, marking them with an animal specific tag and releasing them. This operation is repeated several times. Afterwards, each captured animal is associated with a capture history, which is a vector of zeros and ones giving the capture status at each capture occasion. A 1 is a catch and a 0 is a miss. The frequencies of the observable capture histories form the data set to be analyzed.

The parameters of interest depend on whether the population is assumed to be closed, open, or both. Births and deaths, together with immigration and emigration, can occur in an open

population, but not in a closed one. Therefore, for closed populations, survival rates are supposed equal to one and we want to estimate a population size. On the other hand, open population models specialize in survival rates estimation. Moreover, a capture-recapture experiment can be constructed in a hierarchical way, i.e. by dividing capture occasions into primary periods. This results in two levels of sampling. The population experiences immigration and mortality between primary periods, but it is closed within a primary period, which are typically successive days of capture. This type of sampling is called a robust design. Capture-recapture models for the robust design allow the estimation of abundances for each primary periods and survival rates between periods.

Capture-recapture methods were originally developed in the area of wildlife management (Seber 1982), but they are now used in a variety of applications, including epidemiology (Abeni, Brancato, and Perucci 1994), the evaluation of census undercount (Darroch, Fienberg, Glonek, and Junker 1993) and software testing (Wohlin, Runeson, and Brantestam 1995; Ebrahimi 1997; Briand, Emam, Freimut, and Laitenberger 2000). Therefore, the captured units are no longer animals only. For example, in an epidemiological application, they are humans with a certain disease and capture occasions are reporting lists.

This paper presents a new software for the analysis of capture-recapture data : the R package **Rcapture**. This package uses Poisson regressions to estimate parameters in a capture-recapture experiment. It implements the work of Cormack (Cormack 1985, 1989, 1993b; Cormack and Jupp 1991) and extends it (Rivest and Lévesque 2001; Rivest and Daigle 2004; Rivest and Baillargeon 2007). In **Rcapture**, the Poisson regressions are fitted with the `glm` function; then the loglinear parameters are transformed into demographic parameters.

This article aims to demonstrate the use of the **Rcapture** package. In Section 2, an approach is suggested for the analysis of data from a closed population. Section 3 illustrates how, with **Rcapture**, we can reproduce some of Cormack's data analysis of open populations. The modeling of data from a robust design with **Rcapture** is treated in Section 4. Finally, **Rcapture** is compared to other capture-recapture softwares in Section 5.

2. Closed populations

A population is said to be closed if no mortality nor immigration can occur within the population. Hence, the size of a closed population, noted N, does not vary during the experiment. This assumption is reasonable for capture-recapture experiments held over a short period of time. To estimate this population size, a model is fitted to the data. Following Otis, Burnham, White, and Anderson (1978), the model can incorporate up to three sources of variation among capture probabilities: a temporal effect (subscript t), a heterogeneity between units (subscript h) and a behavioral effect (subscript b). A temporal effect causes the capture probabilities to vary among capture occasions; heterogeneity causes the capture probabilities to vary among units. A behavioral effect means that the first capture changes the behavior of a unit, so the capture probability differs before and after the first capture. These sources of variation lead to eight fundamental closed population models: M_0 (no source of variation), M_t, M_h, M_{th}, M_b, M_{tb}, M_{bh}, M_{tbh}.

The analysis of data from a closed population capture-recapture experiment amounts to finding the best fitting model and estimating the population size from the chosen model. Here we propose steps to follow for such an analysis. Figure 1 schematizes these steps and links

them to relevant functions of the **Rcapture** package. The first step is to explore the data with descriptive statistics. This helps to identify the factors associated to the variability of the capture probabilities. Next, several models are fitted and compared based on standard criteria such as the deviance of the model and the AIC. Ultimately, a model is chosen and the population abundance N is estimated from this model. The following paragraphs describe the **Rcapture** functions associated to each steps.

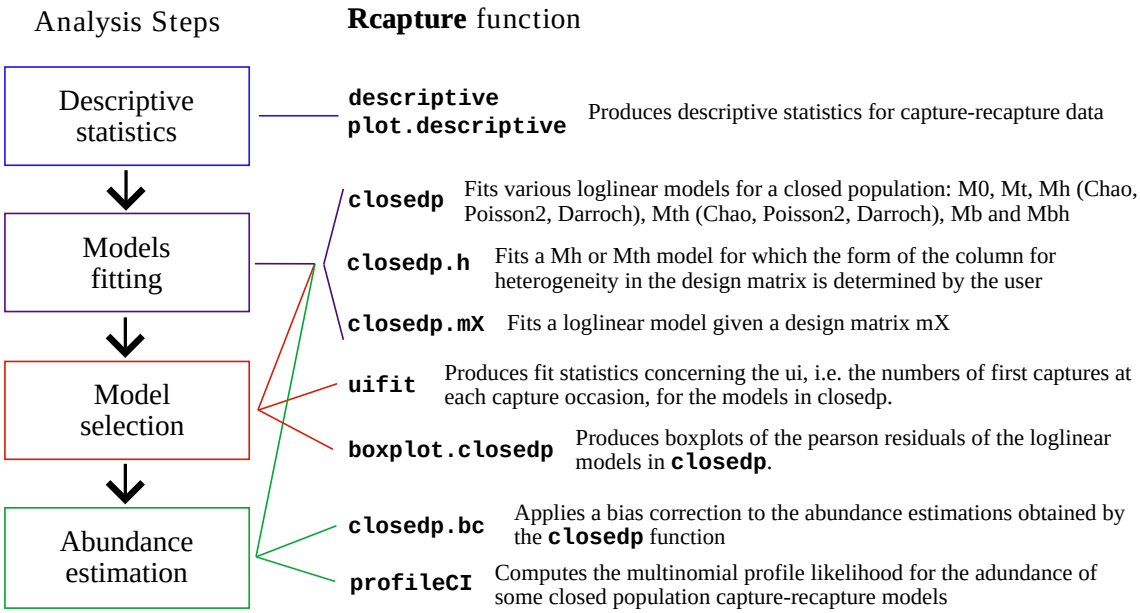

Figure 1: *Analysis approach for closed population data linked to relevant* ***Rcapture*** *functions*

First note that a capture history will be expressed as a $t \times 1$ vector $\boldsymbol{\omega} = (\omega_1, \ldots, \omega_t)$, where $\omega_j = 1$ if the unit is captured at the jth occasion and 0 if not. There are two accepted formats for a capture-recapture data set in the package **Rcapture**. The first one is an R matrix or data frame whose rows are the capture histories of each animal caught. The number of columns in the data matrix is then the number of capture occasions in the experiment (noted t). In the alternative format, the data matrix contains one row per capture history followed by its frequency. In that case, it has $t+1$ columns. The first t columns identify the capture histories. They must contain only zeros and ones. In **Rcapture** functions, the format of the data set is specified with the `dfreq` argument. This argument is set to FALSE for the first format; it is set to TRUE for the alternative format. The function `histpos.t` generates the $(2^t - 1) \times t$ matrix of the observable capture histories in a capture recapture experiment; it also defines the order of the capture histories for the Poisson regression. This order is relevant when using the `closedp.mX` function and the `keep` option of the `openp` functions.

Descriptive Statistics

The `descriptive` function of the **Rcapture** package computes basic capture-recapture frequency statistics. It displays, for $i = 1, \ldots, t$, the number of units captured i times (f_i), the

number of units captured for the first time on occasion i (u_i), the number of units captured for the last time on occasion i (v_i) and the number of units captured on occasion i (n_i). If the n_i statistics vary among capture occasions, there is a temporal effect. The `descriptive` function also gives the m-array matrix, which contains recapture frequencies for units released on each occasion.

An interesting tool to explore a possible heterogeneity in the capture probabilities are the graphs of $\log\left(f_i/\binom{t}{i}\right)$ and $\log(u_i)$ versus i generated by the `plot.descriptive` function. Table 1 gives the form of the two graphs for some models. Some elements of Table 1 are easy to justify. For model M_0, the number of captures follows the $Binomial(t, p)$ distribution and the number of capture occasions before the first capture follows the $Geometric(p)$ distribution, where p is the capture probability of a unit at any capture occasion. This latter result also holds under model M_b. So, in these cases,

$$\log\left(\frac{f_i}{\binom{t}{i}}\right) \simeq \log\left(\frac{N \times \Pr(i \text{ captures})}{\binom{t}{i}}\right) = \log(N(1-p)^{t-i}p^i) = \log(N(1-p)^t) + i\log\left(\frac{p}{1-p}\right)$$

and

$$\log(u_i) \simeq \log(N \times \Pr(\text{first capture on occ } i)) = \log(N(1-p)^{i-1}p) = \log\left(\frac{Np}{1-p}\right) + i\log(1-p)$$

where N is the population abundance we want to estimate. Therefore, the graphs produced by `plot.descriptive` are linear. Moreover, Rivest (2007) shows that the f_i graph should be concave downward when there is a temporal effect. This effect is typically small and the graph of the f_i stays almost linear for model M_t. Furthermore, from the work of Lindsay (1986) on mixing distributions in an exponential family, the f_i graph for model M_h and the u_i graph for models M_h and M_{bh} should be convex, up to sampling errors. The shape of the f_i graph for model M_{th} depends on the relative importance of the temporal effect and the heterogeneity. So the `plot.descriptive` function can bring out heterogeneity among capture probabilities in a data set through graphs with a convex shape.

Graph	M_0	M_t	M_h	M_{th}	M_b	M_{bh}
f_i	L	L*	C	L*/C	?	?
u_i	L	?	C	?	L	C

Table 1: Form of the graphs produced by `plot.descriptive` for different models (The letter L means linear, L* means almost linear, C means concave upward or convex and a question mark indicates that the graph has no definitive form.)

Models Fitting

The main **Rcapture** function for fitting a model to a closed population data set is `closedp`. It fits M_0, M_t, M_h, M_{th}, M_b, and M_{bh} through Poisson regressions. Since M_{tb} and M_{tbh} do not have a loglinear form, `closedp` does not produce abundance estimations for these models. All models are fitted using the `glm` function; it produces maximum likelihood estimates of the loglinear parameters. The maximization is done through an iteratively reweighed least-squares algorithm which is simple and numerically stable. An estimate of the population size N is then derived from the loglinear parameters.

The estimator of N is obtained by maximizing a Poisson loglikelihood. Cormack and Jupp (1991) showed that this Poisson estimator is almost identical to the conditional multinomial estimator. A variance, valid under multinomial sampling, is derived in Sandland and Cormack (1984). It is given by $\text{var}_m(\hat{N}) = \text{var}_p(\hat{N}) - N$ where subscripts m and p refer to multinomial and Poisson sampling.

To illustrate the use of a loglinear model in a closed population experiment, let's detail the case of model M_0. This is the simplest model; it has a single capture probability p common to all units, at every capture occasion, which does not change after a first capture. For an experiment including t capture occasions, $2^t - 1$ capture histories $\boldsymbol{\omega}$ are observable. The probability for a unit to experience a capture history $\boldsymbol{\omega}$ is $\Pr(\boldsymbol{\omega}) = (1-p)^{t-\sum\omega_j} p^{\sum\omega_j}$ where $\sum\omega_j$ is the number of times the unit is caught. Therefore, the expected number of units in the population having capture history $\boldsymbol{\omega}$ is $\mu_{\boldsymbol{\omega}} = N(1-p)^{t-\sum\omega_j} p^{\sum\omega_j}$. This expected frequency can be reexpressed in the form of a loglinear model as

$$\mu_{\boldsymbol{\omega}} = \exp\left(\underbrace{\log(N(1-p)^t)}_{\gamma} + \sum \omega_j \underbrace{\log\left(\frac{p}{1-p}\right)}_{\beta}\right).$$

Thus, model M_0 is fitted in `closedp` by fitting a loglinear model $E(\boldsymbol{Y}) = \exp(\boldsymbol{X}\boldsymbol{\beta})$ with $\boldsymbol{Y}$ equal to the $(2^t - 1) \times 1$ vector of the observed frequencies $n_{\boldsymbol{\omega}}$ (including zero frequencies), $\boldsymbol{X}$ is a $(2^t - 1) \times 2$ design matrix with a first column of ones and a second column defined by $\sum\omega_j$, and $\boldsymbol{\beta} = (\gamma, \beta)^t$. Then, the abundance is estimated as $\hat{N} = n + \exp(\hat{\gamma})$ where n is the total number of units caught during the experiment. This is indeed an estimator of the population size because $\exp(\gamma) = \exp(\log(N(1-p)^t)) = N(1-p)^t = N \times \Pr(\boldsymbol{\omega}_0) = \mu_0$ where $\boldsymbol{\omega}_0$ is the unobservable capture history of zero capture and μ_0 is the expected number of units never captured. A loglinear presentation of the other models fitted by `closedp` can be found in Rivest and Lévesque (2001) and Rivest and Baillargeon (2007). Note that in `closedp` model M_b is as presented in (Cormack 1989) while M_{bh} allows the probability of first capture at occasion 1 to differ from the probability of first capture after occasion 1. It is suitable when the u_i plot of `descriptive` is linear except for occasion 1.

The **Rcapture** package specializes in modeling heterogeneity. The `closedp` function suggests three types of models for M_h and M_{th} : Chao, Darroch and Poisson2. Chao's models estimate a lower bound for the abundance. The estimate obtained under $M_{h\ \text{Chao}}$ is Chao's (1987) moment estimator. Rivest and Baillargeon (2007) exhibit a loglinear model underlying this estimator and provide a generalization to M_{th}. Some loglinear parameters of Chao's models, the η parameters, should theoretically be greater or equal to zero. So when the argument `neg` of the function `closedp` is set to TRUE (the default), negative η parameters are fixed to zero. For Darroch's models, a column defined as $(\sum\omega_j)^2/2$ is added to the design matrix for either M_0 or M_t. These models for M_h and M_{th} are considered by Darroch *et al.* (1993) and Agresti (1994). For Poisson2 models, the column for heterogeneity in the design matrix is $2^{\sum\omega_j} - 1$. For these two models, the logits of the individual capture probabilities are assumed to be random variables. These variables are distributed according to a mixed normal distribution under Darroch's model or to a mixed Poisson distribution under a Poisson model. Details can be found in Rivest and Baillargeon (2007). The Poisson model typically yields smaller corrections for heterogeneity than Darroch's model since the capture probabilities are bounded from below under this model.

In addition to Chao, Darroch and Poisson2 heterogeneity models, other M_h and M_{th} models can be fitted with the `closedp.h` function. This function can fit general Poisson models with heterogeneity columns equal to $a^{\sum \omega_j} - 1$. When a is large, the Poisson estimator is close to the one obtained under models M_0 or M_t and as a goes to 1, the Poisson estimator becomes close to Darroch's estimator. The family of Poisson estimator for M_h and M_{th} provides a wide range of corrections for heterogeneity. The function `closedp.h` can also fit models with the form of the column for heterogeneity in the design matrix defined by the user. For the log-gamma model of Rivest and Baillargeon (2007), this column is $-\log(\lambda+\sum\omega_j)+\log(\lambda)$ for some $\lambda > 0$. **Rcapture** also features a function, `closedp.mX`, that has a user defined design matrix. `closedp.mX` allows, for instance, fitting a model with an interaction between two capture occasions. Adding interactions between successive occasions results in a trap effect because the probability of being captured at occasion i depends on the capture at occasion $i-1$. The function `closedp.mX` estimates the population size as $\hat{N} = n + \exp(\hat{\gamma})$, where $\hat{\gamma}$ is the estimated intercept. Therefore, it is not suited for models with behavioral effects.

Model selection

When several models have been fitted, they must be compared and one has to be selected. The functions `closedp`, `closedp.h` and `closedp.mX` generate deviances, degrees of freedom and Akaike Information Criteria (AIC). These statistics are useful tools to compare models and to assess the goodness of their fit. Under the assumption of a good fit, the deviance of a model follows a chi-square distribution with the model's degrees of freedom. Also, likelihood ratio tests can be constructed to compare nested models and a smaller AIC indicates a better model. Note however that for model $M_{h\ Chao}$ and $M_{th\ Chao}$ a small deviance means that there is a heterogeneity in capture probabilities; it does not mean that the lower bound estimates calculated for these models are unbiased.

The fit of a model can also be judged through its residuals. The functions `boxplot.closedp` and `boxplot.closedp.custom` produces boxplots of the Pearson residuals for the different fitted models. These graphs bring out badly fitted data.

Rcapture also contains a function aimed at studying the model's fit from the u_i statistics. The `uifit` function focuses on what is most important to model accurately, i.e. the number of new captures at each occasion. It displays the observed u_i statistics and the u_i predicted by each model in `closedp`. It also forecasts the u_i for 5 additional hypothetical capture occasions for models M_0, $M_{h\ Poisson2}$, $M_{h\ Darroch}$ and M_b. The predicted and observed u_i-statistics are compared using chi-square statistics. Moreover, the mean and variance of the day of first capture are calculated with the predicted u_i for each model. All these statistics generated by `uifit` are further tools to assess the fit of a model.

Abundance estimation

The functions `closedp`, `closedp.h` and `closedp.mX` give an estimate for the abundance and its standard error. For small samples, the estimation can be improved by a bias correction. The function `closedp.bc` performs, for the models in `closedp`, a bias correction through frequency modifications as presented in Rivest and Lévesque (2001) and Rivest and Baillargeon (2007). These frequency modifications also stabilize the the standard errors estimates for $\hat{N}$.

Abundance can also be estimated through confidence intervals. A naive $100(1-\alpha)\%$ confidence interval assuming asymptotic normality is $\hat{N} \pm Z_{\alpha/2} se(\hat{N})$. Better confidence intervals are

obtained using a profile loglikelihood. This can be done with the `profileCI` function which follows the methodology of Cormack (1992). This function calculates the value of N that maximizes the multinomial likelihood. It also plots the the profile likelihood for N and calculates a $100(1-\alpha)\%$ profile likelihood confidence interval. It works for every model fitted by `closedp`, `closedp.h` or `closedp.mX`, except models M_b and M_{bh}.

2.1. Snowshoe hare example

We now fit closed population models to the snowshoe hare data considered in Cormack (1989) and Agresti (1994). This data set is included in the **Rcapture** package. It has the default format, i.e. each row represents the capture history of one animal. Hence, the argument `dfreq` of **Rcapture** functions doesn't have to be specified as it is set to FALSE by default.

```
> library(Rcapture)
> data(hare)
> desc<-descriptive(hare)
> plot(desc)
> closedp(hare)

Number of captured units: 68

Abundance estimations and model fits:
              abundance  stderr  deviance  df      AIC
M0                 75.4     3.5    68.516  61  154.707
Mt                 75.1     3.4    58.314  56  154.505
Mh Chao            79.8     6.4    58.023  58  150.214
Mh Poisson2        81.5     5.7    59.107  60  147.298
Mh Darroch         90.4    11.6    61.600  60  149.791
Mh Gamma3.5       100.6    21.7    62.771  60  150.961
Mth Chao           79.6     6.3    47.115  52  151.305
Mth Poisson2       81.1     5.6    48.137  55  146.327
Mth Darroch        90.5    11.7    50.706  55  148.896
Mth Gamma3.5      101.6    22.4    51.956  55  150.147
Mb                 81.1     8.3    67.027  60  155.217
Mbh                74.2    14.6    63.257  59  153.447

Note: 1 eta parameter has been set to zero in the Mh Chao model
```

The f_i plot of the function `descriptive` in Figure 2 shows that the two animals caught on all occasions create some heterogeneity in the capture probabilities. Therefore, it is not surprising that the best fitting model is heterogenous. Indeed, the model with the smallest AIC (146.327) is $M_{\text{th Poisson2}}$. It leads to an estimate $\hat{N}$ equals to 81.1 ($s.e. = 5.7$). The estimate for $M_{\text{th Darroch}}$ is equal to that reported in Agresti (1994).

Another approach to take care of the heterogeneity would be to remove the 2 hares caught 6 times, as Cormack (1989) did. With **Rcapture**, the best way to discard these hares is to add a column to the design matrix for M_t taking the value 1 for the capture history $(1, 1, 1, 1, 1, 1)$ and 0 otherwise.

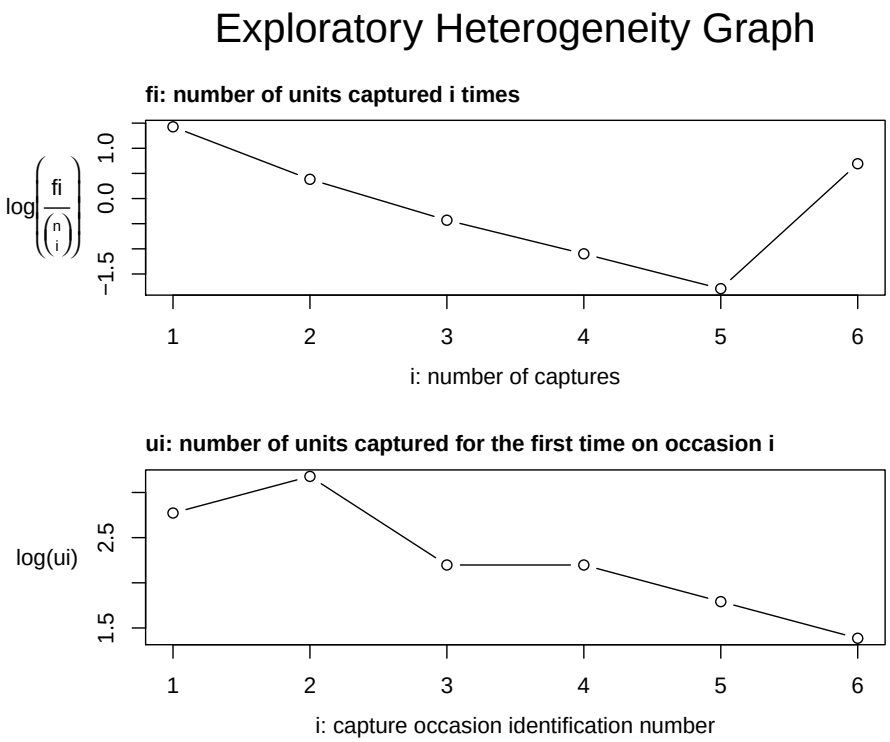

Figure 2: Plot of the descriptive object of the snowshoe hare data

```
> col<-rep(0,2^6-1)
> mat<-histpos.t(6)
> col[apply(mat,1,sum)==6]<-1
> cp.m2<-closedp.mX(hare,mX=cbind(mat,col),mname="Mt without 111111")
> cp.m2$results

                 abundance   stderr deviance df      AIC
Mt without 111111  76.77761 3.911153 47.89417 55 146.0846
```

This gives $\hat{N} = 76.8$ ($s.e. = 3.9$) with an AIC of 146.085. These results match Cormack's results in Table 4 (1989, p.406). Besides the point estimates for N, profile likelihood confidence intervals can easily be calculated for both models.

```
> CI1<-profileCI(hare,m="Mth",h="Poisson",a=2)
> CI1$results

             abundance    InfCL    SupCL
Mth Poisson2        80 71.84073 93.84254

> CI2<-profileCI(hare,mX=cbind(mat,col),mname="Mt without 111111")
> CI2$results
```

```
                   abundance    InfCL    SupCL
Mt without 111111         76 70.08663 85.41181
```

The upper bound of the confidence interval for N depends on the interpretation given to the two hares caught at all occasions. It is large when they are assumed to be associated with a small heterogeneity in the capture probabilities. It is small when the two trap happy hares are assumed to be unrepresentative of the unsampled part of the population.

2.2. HIV example

We now analyze epidemiological capture-recapture data on HIV in Abeni *et al.* (1994). The capture histories are obtained by linking the records of four reporting centers in Rome, Italy. The data set's format is the alternative one, i.e. each row represents an observed capture history followed by its frequency. Therefore, the argument `dfreq` of the **Rcapture** functions has to be set to TRUE.

```
> data(HIV)
> descriptive(HIV,dfreq=TRUE)

Number of captured units: 1896

Frequency statistics:
       fi    ui    vi    ni
i = 1  1774   466   403   466
i = 2   115   593   578   630
i = 3     7   632   679   693
i = 4     0   205   236   236
fi: number of units captured i times
ui: number of units captured for the first time on occasion i
vi: number of units captured for the last time on occasion i
ni: number of units captured on occasion i
```

The function descriptive shows that 1774 out of 1896 individuals (94%) appear on one list only. The f_i plot in Figure 3 is linear showing that heterogeneity is not a problem; the u_i plot is not interpretable since it depends on the arbitrary ordering of the 4 centers. The model with a time (or a list) effect and the six possible pairwise dependencies between lists is fitted.

```
> mat<-histpos.t(4)
> mX1<-cbind(mat,mat[,1]*mat[,2],mat[,1]*mat[,3],mat[,1]*mat[,4],mat[,2]*mat[,3],mat[,2]*m
> cp.m1<-closedp.mX(HIV,dfreq=TRUE,mX=mX1,mname="Mt double interactions")
> cp.m1$results

                       abundance   stderr deviance df      AIC
Mt double interactions  23443.54 9594.879 3.036804  4 92.07266
```

The above model fits well. We need to find out the dependencies that are important; their estimates are given by parameters `mX5` to `mX10` in the output.

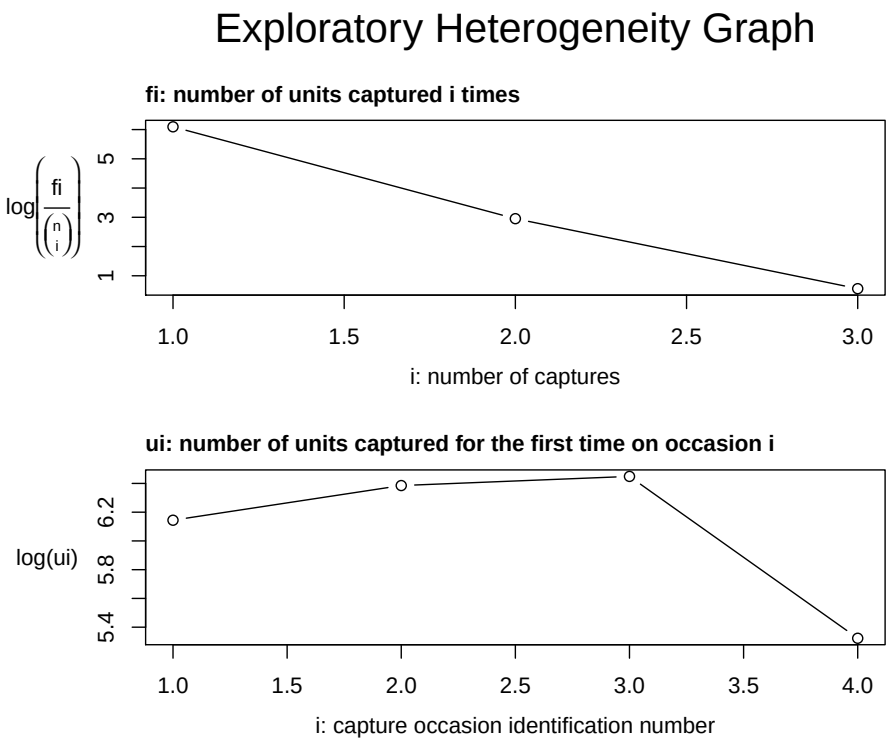

Figure 3: Plot of the descriptive object of the HIV data

```
> summary(cp.m1$glm)$coefficients
```

```
             Estimate Std. Error     z value      Pr(>|z|)
(Intercept)  9.9780167  0.4452368  22.4105827 3.103499e-111
mX1         -3.9758604  0.4438850  -8.9569599  3.337529e-19
mX2         -3.6785194  0.4437085  -8.2903961  1.128719e-16
mX3         -3.5469201  0.4440686  -7.9873239  1.378994e-15
mX4         -4.6582017  0.4453831 -10.4588655  1.334439e-25
mX5          1.1545857  0.4329136   2.6670119  7.652896e-03
mX6          0.4810600  0.4305346   1.1173552  2.638425e-01
mX7          0.3339371  0.5168483   0.6461027  5.182129e-01
mX8          0.8266913  0.4291786   1.9262176  5.407721e-02
mX9          0.7884198  0.4612659   1.7092522  8.740424e-02
mX10         0.6951611  0.4705025   1.4774867  1.395452e-01
```

Eliminating the non-significant interactions stepwise shows that only the [1,2] interaction is important. The results for the final model are the following. Figure 4 shows the 95% profile likelihood confidence interval of the abundance. The results are close to the results in Abeni *et al.* (1994, p.413), but not equal due to differences in the estimation method.

```
> mX2<-cbind(mat,mat[,1]*mat[,2])
```

```
> cp.m2<-closedp.mX(HIV,dfreq=TRUE,mX=mX2,mname="Mt interaction 1,2")
> cp.m2$results

                   abundance   stderr deviance df      AIC
Mt interaction 1,2  12318.47 1188.722 7.613759  9 86.64962

> CI<-profileCI(HIV,dfreq=TRUE,mX=mX2,mname="Mt interaction 1,2")
> CI$results

                   abundance    InfCL    SupCL
Mt interaction 1,2     12308 10286.85 14977.69
```

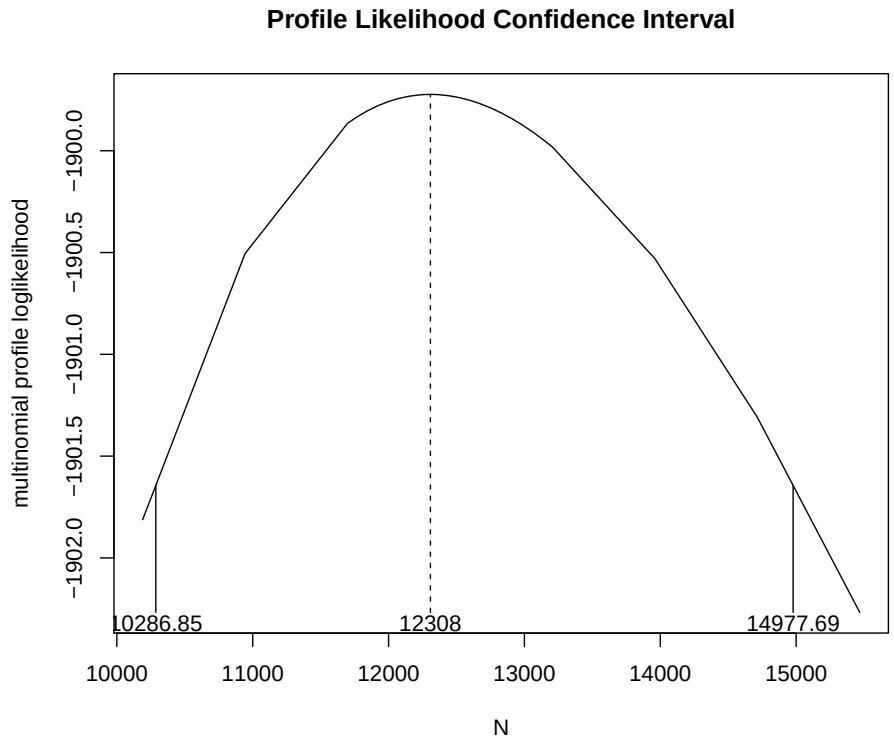

Figure 4: Plot of the 95% profile likelihood confidence interval of the abundance for the HIV data

2.3. Meadow vole period 3 example

The last closed population example concerns the third primary sampling period of the meadow vole data set presented in Chapter 19 of Williams, Nichols, and Conroy (2002). The data is in columns 11 to 15 of the data set `mvole` included in the **Rcapture** package. The complete data set will be analyzed with a robust design model in Section 4.1. Descriptives statistics are not presented here, but they suggest that heterogeneity is present in the data for the third period.

```
> data(mvole)
> cp<-closedp(mvole[,11:15])
> cp

Number of captured units: 49

Abundance estimations and model fits:
              abundance  stderr  deviance  df      AIC
M0                 51.1     1.6    66.964  29  122.895
Mt                 50.9     1.6    61.208  25  125.138
Mh Chao            71.9    14.2    33.556  26   95.486
Mh Poisson2        61.0     6.3    37.902  28   95.833
Mh Darroch         93.2    26.7    34.611  28   92.541
Mh Gamma3.5       203.2   119.1    33.984  28   91.915
Mth Chao           71.0    13.7    26.120  22   96.051
Mth Poisson2       60.5     6.1    30.652  24   96.582
Mth Darroch        93.1    26.6    27.178  24   93.108
Mth Gamma3.5      209.9   124.3    26.539  24   92.470
Mb                 51.0     2.0    66.964  28  124.894
Mbh                52.6     9.1    66.256  27  126.187
```

Model M_h gives the best fit; the abundance estimator can vary by up to 33% according to the model selected. Very large estimates are possible; for instance one can try the log gamma model discussed in Rivest and Baillargeon (2007).

```
> psi<-function(x){-log(3.5+x)+log(3.5)}
> lgmodel<-closedp.h(mvole[,11:15],h=psi)
> lgmodel$results

       abundance   stderr deviance df      AIC
Mh psi  203.2393 119.0627 33.98449 28 91.91481
```

This gives a very small AIC. However the estimate of 203 is too large. To help select an estimate one can use the function `uifit` that assesses the fit of each model for the number of new captures at each occasion and forecasts, for some models, the number of new captures if the experiment were continued.

```
> xx<-uifit(cp)
> xx$predicted[,c(1,4,5,6)]
```

```
    observed Mh Chao Mh Poisson2 Mh Darroch
u1        26    24.2  24.2000000  24.200000
u2        12     9.9  10.1340984   9.900000
u3         3     6.2   6.5764481   6.340635
u4         6     4.7   4.6588525   4.747957
u5         2     4.0   3.4306010   3.811407
u6        NA      NA   2.5849972   3.180338
u7        NA      NA   1.9785928   2.720707
u8        NA      NA   1.5319327   2.368835
u9        NA      NA   1.1965924   2.089944
u10       NA      NA   0.9411818   1.863153
```

There is not much ground for discriminating between the $M_{h\ Poisson2}$ and the $M_{h\ Darroch}$ estimator; still the $M_{h\ Poisson2}$ predicted values for u_i are somewhat closer to the observed u_i than those for $M_{h\ Darroch}$. One can also wonder whether to predict that 1.86 new unmarked animals will be caught on a hypothetical 10th day of capture is realistic. In the model selection process, it might also be useful to look at the models' Pearson residual.

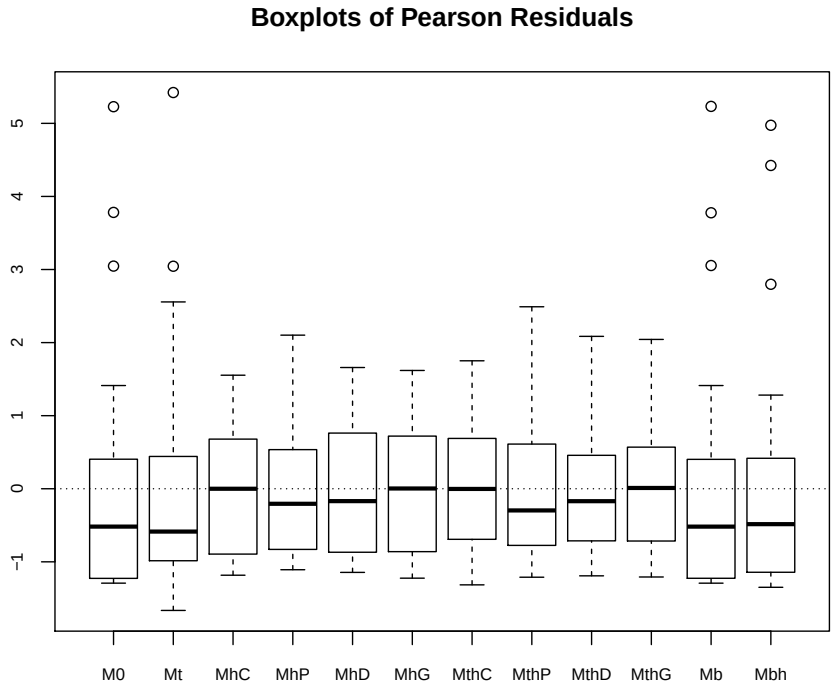

Figure 5: Boxplots of the Pearson residuals of the models fitted by `closedp` for the third period of the meadow vole data

The boxplots in Figure 5 present another argument for selecting $M_{h\ Poisson2}$ over $M_{h\ Darroch}$ since $M_{h\ Poisson2}$ residuals are more concentrated around zero. The selection of a model for

M_h is settled using the robust design in Section 4.1. It turns out that $M_{h\ Darroch}$ is not appropriate.

3. Open populations

Open population models apply when animals are released and recaptured or resighted at future capture occasions. Typically the capture occasions are distant in time and mortality occurs between them. When the animals released are not a random sample of the animals in the population at a given capture occasion, the analysis focuses on the estimation of survival rates of the animals that were released. The Cormack-Jolly-Seber model applies in such situations. When marked and unmarked animals undergo the same sampling process, both the population sizes and the survival rates can be estimated. This is the Jolly-Seber model. Open population models are often used for capture-recapture experiments held over a long period of time. Therefore, the capture occasions are called periods; they are indexed by the subscript i ranging from 1 to I.

The function `openp` of **Rcapture** fits both the Cormack-Jolly-Seber and the Joly-Seber model following the loglinear approach of Cormack (1985, 1989), see also Rivest and Daigle (2004). If the interest focuses only on estimating survival rates, the abundance estimators are simply discarded. Besides the survival rates ϕ_1 to ϕ_{I-1}, these functions estimate the capture probabilities $p*_1$ to $p*_I$, the population sizes N_1 to N_I, the number of new units entering the population B_1 to B_{I-1} and the total number or units who ever inhabited the survey area $Ntot$. In some applications of the Jolly Seber model, births are arrivals to the colony and deaths are departures (see Schwarz and Stobo 1997). In those cases, the total number of visitors to the colony $Ntot$, is the parameter of interest. By default, the argument `m` of the function `openp` is set to "up"; which means that the capture probabilities vary between periods (up = unconstrained probabilities). Because of the well known lack of identifiability for the Jolly-Seber model (see Pollock, Nichols, Brownie, and Hines 1990), the parameters $p*_1$, $p*_I$, the survival rate ϕ_{I-1} between periods $I-1$ and I, N_1 and N_I are not estimable with the function `openp.up`. On the other hand, all the parameters are estimable when `m` is given the value "ep" because it sets the capture probabilities equal to a common value (ep = equal probabilities). The function `openp` insures that the estimated survival probabilities belong to $[0, 1]$ and that the births B_i are positive by imposing constraints to the loglinear parameters. Setting the argument `neg` of this function to FALSE removes these constraints.

The steps we propose to follow in the analysis of an open population data set differ from the ones for a closed population data set. A single model is fitted; refitting the model to a subset of the data can be attempted if it doesn't fit well. Figure 6 summarizes the procedure. First, if the experiment follows a robust design, the data matrix must be converted to between primary session data. This is done with the function `periodhist` which pools the capture histories for several occasions into a single entry having the value 1 for a unit caught at least once during these occasions and 0 otherwise. Next, descriptive statistics can be produced to explore the data; the `m.array` matrix output by `descriptive` is of interest. The Cormack-Jolly-Seber or the Jolly-Seber model is fitted with `openp` which produces estimates of the demographic parameters. The presence of a trap effect is tested by including additional loglinear parameters in the model. Unfortunately, estimates of demographic parameters accounting for a significant trap effect cannot be calculated at this time. The model's quality of fit is assessed with the deviance of the standard model (without a trap effect) and with `plot.openp`. This function

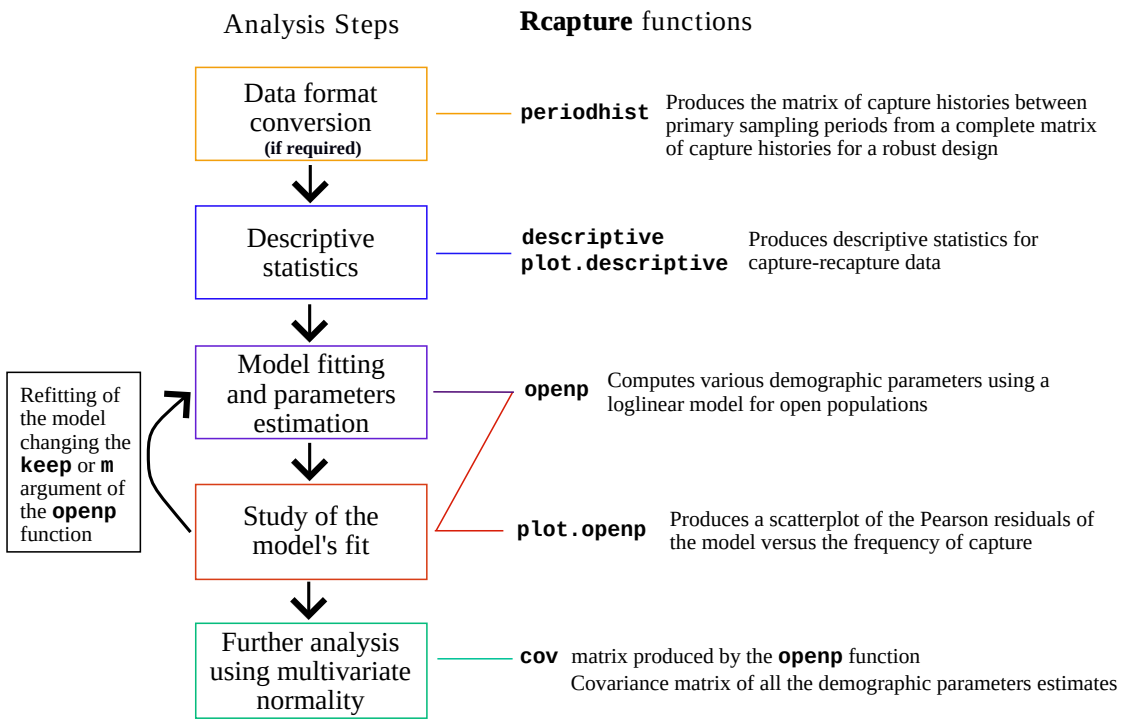

Figure 6: *Strategy for the analysis of open population data linked to relevant* ***Rcapture*** *functions*

plots the Pearson residuals versus the number of captures. Large residuals bring out badly fitted data. To pursue its analysis, the user can choose to remove some units from the data set and refit the model. This is done with the `keep` argument of `openp`. Typically, one wishes to omit the units caught too often or, on the contrary, the units caught only once which can be considered as transients. The removal of units is a good diagnostic tool to assert the stability of the results, but it is not advised as a general strategy. When some units are omitted, **Rcapture** brings them back into the final abundance estimates with a correction proposed by (Cormack 1989, p.410). The difference between observed and expected frequencies for the omitted groups is added to the model's estimations of abundance. To complete the analysis, it is possible to construct tests on the parameters under the assumption of multivariate normality of the estimators as will be shown in the following examples. For this, the covariance matrix of all the demographic parameters estimates is used. This matrix is returned by the function `openp` under the name `cov`. As noted before, the abundance output presents standard errors adjusted to be valid under multinomial sampling. However, the `code` matrix contains unadjusted variances of abundance estimators. So these variances are valid under Poisson sampling and not multinomial sampling.

3.1. Lazuli bunting example

Let's analyze the lazuli bunting data treated in Cormack (1993a). The data comes from a eight-year (1973 to 1980) study by Allen W. Stokes of lazuli bunting wintering in Logan,

Utah.

```
> data(bunting)
> descriptive(bunting,dfreq=TRUE)

Number of captured units: 1681

Frequency statistics:
       fi    ui    vi    ni
i = 1  1430   168   132   168
i = 2   180   367   359   398
i = 3    37    65    64    88
i = 4    23   230   213   264
i = 5     7   255   232   304
i = 6     2   256   246   322
i = 7     1   240   247   323
i = 8     1   100   188   188
fi: number of units captured i times
ui: number of units captured for the first time on occasion i
vi: number of units captured for the last time on occasion i
ni: number of units captured on occasion i
```

The descriptives statistics show that 1430 birds out of a total of 1681 birds seen (85%) were caught only once. This suggests the presence of transient birds at each capture occasion. This might bias the survival probabilities downward since, in the presence of transient animals, these represent the probabilities of not being a transient and of surviving. The Jolly-Seber model is fitted by the following command.

```
> op.m1<-openp(bunting,dfreq=TRUE)
> op.m1$model.fit[1,]

 deviance        df       AIC
 219.4100  234.0000  456.0147

> plot(op.m1)
```

The residuals plot in Figure 7 shows large residuals for the birds caught twice or more while the residuals are small for birds caught once. The Jolly-Seber model does not fit well and the likely presence of transients might cause that. To remove the birds caught only once from the analysis, one uses the `keep` argument as follows.

```
> keep2<-apply(histpos.t(8),1,sum)>1
> op.m2<- openp(bunting,dfreq=TRUE,keep=keep2)
> op.m2$model.fit[1,]

 deviance        df       AIC
 125.1796  228.0000  302.5521
```

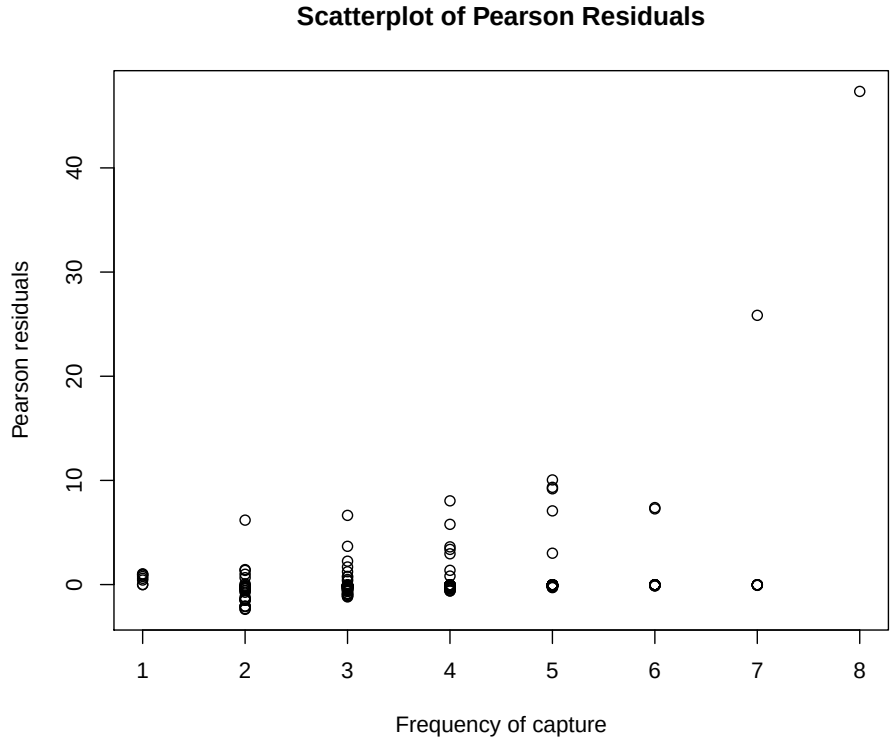

Figure 7: Plot of the Pearson residuals of the Jolly-Seber model fitted to the lazuli bunting data

The deviance drop of 94 for 6 degrees of freedom is highly significant. The residual plot for this model is not presented here; it still has some Pearson residuals larger than 4 that might influence the survival estimates. The next commands use the object `keep3` to identify capture histories with more than one capture and with residuals smaller than 4.

```
> keep3p<-residuals(op.m2$glm,type="pearson")<4
> num3<-((1:255)[keep2])[keep3p]
> keep3<-rep(FALSE,255)
> keep3[num3]<-TRUE
> op.m3<- openp(bunting,dfreq=TRUE,keep=keep3)
> tab<-data.frame(op.m2$survivals,rep("|",7),op.m3$survivals)
> colnames(tab)<-c("estimate.m2","stderr.m2","|","estimate.m3","stderr.m3")
> tab
```

```
              estimate.m2  stderr.m2 | estimate.m3  stderr.m3
period 1 -> 2          NA         NA |          NA         NA
period 2 -> 3   0.4851117 0.13125375 |   0.4815109 0.14167601
period 3 -> 4   0.6742944 0.13207217 |   0.6188964 0.13535331
period 4 -> 5   0.7287239 0.12854646 |   0.7013263 0.12751218
```

```
period 5 -> 6   0.5176471 0.09566484 |   0.5012495 0.09560564
period 6 -> 7   0.5559809 0.08310532 |   0.5512693 0.08368251
period 7 -> 8          NA         NA |          NA         NA
```

The two sets of survival estimates are similar; the large residuals have a small impact. Tables 3 and 4 of Cormack (1993a, p.46) present estimates obtained by fitting the first two models of this Section. They report estimates that are identical to those presented here.

The survival estimates are quite similar between periods. We would like to test the equality of the survival probabilities and estimate their common value. In softwares such as **Mark** and **M-Surge**, this test is easily performed by fitting a model with constant survival probabilities. This model is not loglinear. An asymptotically equivalent to these homogeneity tests can be calculated with **Rcapture** using the output from `openp`.

The vector of estimated survival probabilities $\hat{\boldsymbol{\phi}} = (\hat{\phi}_2, ..., \hat{\phi}_6)$ is `op.m2$survivals[2:6,2]` while its estimated covariance matrix $\hat{\boldsymbol{\Sigma}}$ is `op.m2$cov[8:12,8:12]`. Under the hypothesis of constant survival, $\hat{\boldsymbol{\phi}}$ is distributed as $N_5(\phi\mathbf{1}, \hat{\boldsymbol{\Sigma}})$. The least squares estimates of ϕ and of its standard error are

$$\hat{\phi} = \frac{\mathbf{1}^t\hat{\boldsymbol{\Sigma}}^{-1}(\hat{\phi}_2, ..., \hat{\phi}_6)^t}{\mathbf{1}^t\hat{\boldsymbol{\Sigma}}^{-1}\mathbf{1}} \qquad \text{and} \qquad se(\hat{\phi}) = \frac{1}{\sqrt{\mathbf{1}^t\hat{\boldsymbol{\Sigma}}^{-1}\mathbf{1}}}.$$

In R, we calculate them as follows.

```
> siginv<-solve(op.m2$cov[8:12,8:12])
> phi<-t(rep(1,5))%*%siginv%*%op.m2$survivals[2:6,1]/(t(rep(1,5))%*%siginv%*%rep(1,5))
> se<-1/sqrt(t(rep(1,5))%*%siginv%*%rep(1,5))
> data.frame(estimate=phi,stderr=se,row.names="Common survival: ")

                  estimate    stderr
Common survival:  0.5872904 0.0342289
```

Thus $\hat{\phi} = 0.587$ with $s.e. = 0.034$. Under the assumption of a constant survival, the statistic $(\hat{\boldsymbol{\phi}} - \hat{\phi}\mathbf{1})^t\hat{\boldsymbol{\Sigma}}^{-1}(\hat{\boldsymbol{\phi}} - \hat{\phi}\mathbf{1})$ has a chi-square distribution with 5-1=4 degrees of freedom. So, the chi-square goodness of fit statistic for a constant survival and its pvalue are the following.

```
> chisq4<-t(op.m2$survivals[2:6,1]-phi*rep(1,5))%*%siginv%*%(op.m2$survivals[2:6,1]-phi*re
> data.frame(stat=chisq4,pvalue=1-pchisq(chisq4,df=4),row.names="Chi-square test: ")

                    stat    pvalue
Chi-square test:  2.62006 0.6232736
```

The hypothesis of a constant survival is accepted.

3.2. Eider duck example

This example shows that **Rcapture** can reproduce the analysis of eider duck data set presented in Cormack (1989).

```
> data(duck)
> op.m1<-openp(duck,dfreq=TRUE)
> op.m1$model.fit[1,]

 deviance        df       AIC
 83.35991  49.00000 328.83007

> plot(op.m1)
```

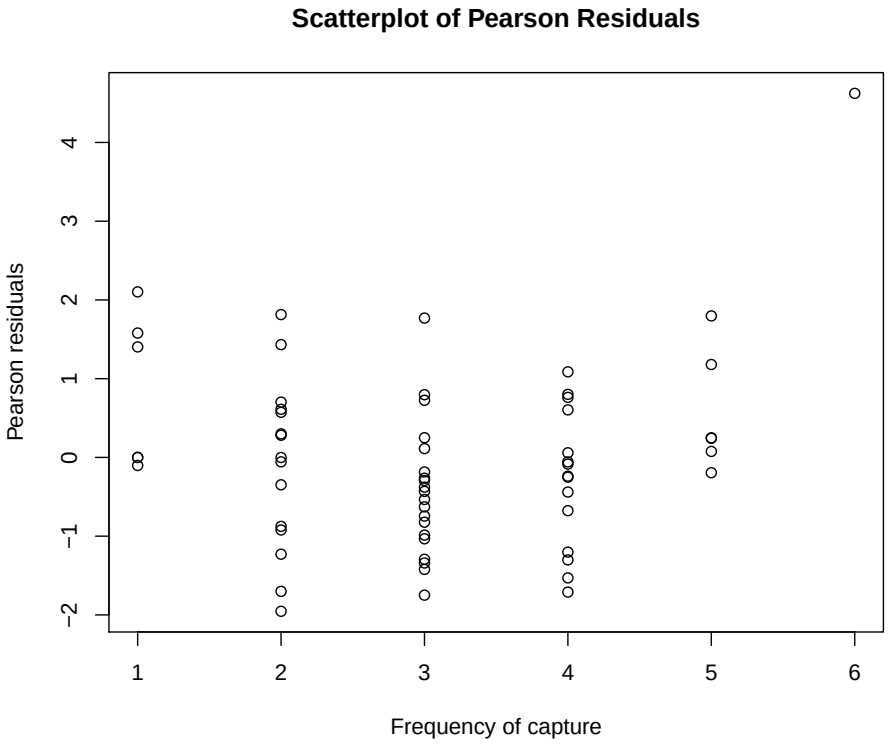

Figure 8: Plot of the Pearson residuals of the model fitted to the eider duck data

The deviance is 83.36 for 49 degrees of freedom. The pvalue of the goodness of fit test based on the deviance is

```
> 1-pchisq(op.m1$model.fit[1,1],df=49)

[1] 0.001592682
```

This is less than 2%. The residual plot in Figure 8 shows a large residual for the 13 ducks captured all the times. We redo the analysis without them.

```
> keep2<-apply(histpos.t(6),1,sum)!=6
> op.m2<-openp(duck,dfreq=TRUE,keep=keep2)
> op.m2$model.fit[1,]
```

```
 deviance        df       AIC
 67.31143  48.00000 308.36595
```

```
> 1-pchisq(op.m2$model.fit[1,1],df=48)
```

```
[1] 0.03427131
```

The fit is still not satisfactory. The residual plot has the convex shape characteristic of heterogeneity in the capture probabilities. We also remove the individuals caught at 5 periods out of 6.

```
> keep3<-apply(histpos.t(6),1,sum)<5
> op.m3<-openp(duck,dfreq=TRUE,keep=keep3)
> op.m3$model.fit[1,]
```

```
 deviance        df       AIC
 56.83066  42.00000 277.20140
```

```
> 1-pchisq(op.m3$model.fit[1,1],df=42)
```

```
[1] 0.06298297
```

The fit is better but there is still heterogeneity in the data. To investigate whether the capture probabilities are homogeneous, one can fit a model with equal capture probabilities.

```
> op.m4<-openp(duck,dfreq=TRUE,keep=keep3,m="ep")
> op.m4$model.fit[1,]
```

```
 deviance        df       AIC
 117.9115   47.0000  328.2822
```

It gives a much larger deviance; so the hypothesis of equal capture probabilities is rejected. In the end, the best model is the one fitted without the animals captured 5 or 6 times. The abundances obtained from that model are the following.

```
> op.m3$N
```

```
          estimate   stderr
period 1       NA       NA
period 2 395.8469 22.96070
period 3 483.6490 32.44455
period 4 386.9222 22.76065
period 5 494.1728 28.99329
period 6       NA       NA
```

These abundances and previously shown deviances reproduce the results in Table 6 of Cormack (1989, p.408).

We now investigate models for the growth rate N_{i+1}/N_i of this population using the multivariate normal distribution for the abundance estimates. If the estimated variance covariance matrix of $(\hat{N}_2, \ldots, \hat{N}_5)$ is $\hat{\Sigma}$, then the variance of the growth rates $(\hat{N}_3/\hat{N}_2, \ldots, \hat{N}_5/\hat{N}_4)$ is $\hat{A}\hat{\Sigma}\hat{A}^t$ where A is the 3×4 matrix of partial derivatives,

$$A = \begin{pmatrix} -\hat{N}_3/\hat{N}_2^2 & 1/\hat{N}_2 & 0 & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & -\hat{N}_5/\hat{N}_4^2 & 1/\hat{N}_4 \end{pmatrix}.$$

The **R** code for the calculations of the the growth rates and their standard errors is as follows.

```
> growth<-op.m3$N[3:5,1]/op.m3$N[2:4,1]
> partial<-matrix(c(-op.m3$N[3,1]/op.m3$N[2,1]^2,1/op.m3$N[2,1],0,0,
+ 0,-op.m3$N[4,1]/op.m3$N[3,1]^2,1/op.m3$N[3,1],0,
+ 0,0,-op.m3$N[5,1]/op.m3$N[4,1]^2,1/op.m3$N[4,1]),3,4,byrow=TRUE)
> sig<-partial%*%op.m3$cov[9:12,9:12]%*%t(partial)
> cbind(estimate=growth,stderr=sqrt(diag(sig)))
```

```
          estimate     stderr
period 3 1.2218081 0.11281160
period 4 0.8000063 0.07563975
period 5 1.2771890 0.08512487
```

As previously mentioned, this standard error is calculated using variances under Poisson sampling. In the same way that we calculated a common survival in Section 3.1, we can now obtain an estimate for the common growth rate.

```
> siginv<-solve(sig)
> growth.e<-t(rep(1,3))%*%siginv%*%growth/(t(rep(1,3))%*%siginv%*%rep(1,3))
> se<-1/sqrt(t(rep(1,3))%*%siginv%*%rep(1,3))
> data.frame(estimate=growth.e,stderr=se,row.names="Common growth rate: ")
```

```
                    estimate     stderr
Common growth rate:  1.037558 0.03187539
```

A chi-square statistic for testing the equality of the growth rates and its pvalue are

```
> chisq2<-t(growth-growth.e*rep(1,3))%*%siginv%*%(growth-growth.e*rep(1,3))
> data.frame(stat=chisq2,pvalue=1-pchisq(chisq2,df=2),row.names="Chi-square test: ")
```

```
                      stat      pvalue
Chi-square test:  13.53338 0.001151498
```

The hypothesis of a common growth rate is rejected at the 5% level. As an alternative to this analysis, one may fit a model with a common growth rate. Such models are not loglinear; they can be fitted by the software **Popan** available in **Mark**.

3.3. Revisiting the snowshoe hare example of Section 2.1

One can use the function `openp` to investigate whether the hare population is closed. The following commands add possible deaths and immigrations to the final model fitted with `closedp.mX` in Section 2.1.

```
> data(hare)
> keep<-rep(TRUE,2^6-1)
> mat<-histpos.t(6)
> keep[apply(mat,1,sum)==6]<-FALSE
> op<-openp(hare,keep=keep)
> op$model.fit[1,]

 deviance        df       AIC
 46.12223  52.00000 145.69894
```

The new deviance of 46.2, $df = 52$, is not significantly smaller than the one for the closed population model, 47.9, $df = 55$, see Section 2.1. The assumption that the population is closed cannot be rejected. Models featuring births and deaths after a particular sampling occasion can be fitted to this data set using the function `robustd.t` discussed in the next section.

4. Robust design

The robust design is a combination of models for closed and open populations introduced by Pollock (1982). Units are captured at different periods between which the population experiences mortality and immigration. Thus, open population models apply at this first level of sampling to estimate survival rates. However, within each primary period, sampling is done more than once; that is, a short term study is conducted. Closed population models are used at this stage to estimate population sizes. By pooling the data of a series of short-term studies, the robust design improves the estimation of the demographic characteristics of the population.

With the package **Rcapture**, one can fit a model for a robust design using either the function `robustd.t` or the function `robustd.0`. These functions implement the loglinear parameterizations presented in Rivest and Daigle (2004). They estimate the same demographic parameters as the `openp` function, without any constrain or unestimable parameters. Within the primary periods, The function `robustd.t` can fit closed population models M_0, M_t, M_h and M_{th} while `robustd.0` only accepts models M_0 and M_h. That is, the function `robustd.0` doesn't fit models with a within period temporal effect. However, it is much less memory consuming than `robustd.t`, so it runs faster. This is so because `robustd.0` codes capture histories in terms of the number of captures for each primary period. Therefore, the length of the response vector for a model fitted with `robustd.t` is $2^{\sum_{i=1}^{I} t_i} - 1$ while it is $\prod_{i=1}^{I}(t_i + 1) - 1$ for a model

fitted with `robustd.0`, where t_i stands for the number of capture occasions at period i. For an experiment such as the one in Section 4.1, with 6 primary periods having each 5 capture occasions, this represents over 1 billion entries in the dependent vector, brought down to 46 655 by the alternative coding of the capture histories.

The function `robustd` uses the data matrix and a vector `vt` containing the numbers of capture occasions for each primary sampling period as input arguments. The closed population models for each period are specified with the arguments `vm`, `vh` and `va` as described in the package documentation. Negative γ parameter estimates in the open population part of the model and negative η parameter estimates for Chao's closed population models are by default set to zero. They can be unconstrained by setting the `neg` function to FALSE, however this also allows survival estimates to be greater than 1 and immigration parameters to be less than 0.

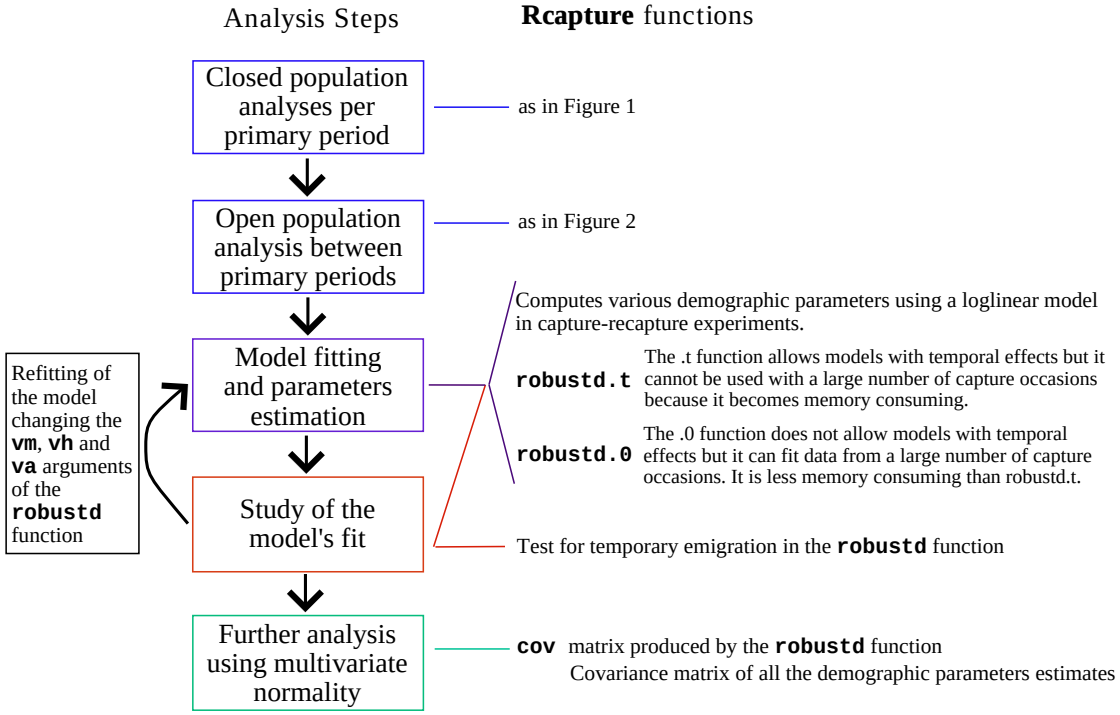

Figure 9: *The steps of a robust design analysis linked to relevant* ***Rcapture*** *functions*

Figure 9 presents the steps in a robust design data analysis. First, closed population analyses should be performed, as described in Section 2, for the short terms study within each period. Then, after converting the data set with the function `periodhist`, an open population analysis is conducted (see Section 3). These preliminary analyses suggest robust design models for consideration, which are now fitted with the function `robustd.t` or `robustd.0`. The selection of closed population models within periods relies on the closed populations analyses. The model's fit is evaluated by testing for the presence of a temporary emigration. This is done by comparing the deviance of the fitted model to the deviance of the same model with temporary emigration, homogenous or not. One can simply use the AIC to do the comparison, or a likelihood ratio test can be performed. If a model with temporary emigration is significantly

better than the model without temporary emigration, then the fitted model might not be appropriate. A bad fit can be associated to a temporary emigration out of the study area if the difference on the logit scale of the between period capture probabilities minus the within period capture probabilities are negative. A bad fit can also be caused by an improper modeling of the within period capture probabilities, especially if the capture probabilities display some heterogeneity. New specifications for the models M_h or M_{th} used in the primary periods might be needed. As in the open population analysis, once a final model is chosen, further analysis can be conducted assuming that the parameter estimates have a multivariate normal distribution.

In `robustd.t` and `robustd.0`, the parameter values of the closed population models change between periods. Also these functions do not have a `keep` argument for investigating the impact of a particular capture history on the outcome.

4.1. Meadow vole example

This example presents a study of the complete meadow vole data set in Chapter 19 of Williams *et al.* (2002). The third period of this data set has been analysed in Section 2.3. This data set concerns a robust design with 30 capture occasions pertaining to 6 primary periods having 5 capture occasions each. These capture occasions represent five consecutive days of trapping every month from June to December 1981 at Patuxent Wildlife Research Center, Laurel, Maryland. This data set has 10 trap deaths that are ignored in this analysis.

First, a between primary period Jolly-Seber analysis is presented.

```
> data(mvole)
> mvole.op<-periodhist(mvole,vt=rep(5,6))
> op.m1<-openp(mvole.op,dfreq=TRUE)
```

There is one large residual, removing the corresponding capture history from the analysis does not change the results. The model fits well.

```
> keep2<-residuals(op.m1$glm,type="pearson")<4
> op.m2<-openp(mvole.op,dfreq=TRUE,keep=keep2)
> op.m2$model.fit
```

```
             deviance     df       AIC
fitted model  36.0454     47  150.9319
```

To find a suitable model within each primary period, the function `closedp` has been used repeatedly. Heterogeneity has been detected in all periods except the second one where the data collection was perturbed by a racoon (the last capture occasion for the second period does not have any new animal captured and is taken out of the analysis). In a robust design we use M_h models for all primary periods bearing in mind the questionable fit in the second one. Since there is no time effect within primary periods, we use the function `robustd.0` to fit the model.

```
> rd.m1<-robustd.0(mvole[,-10], vt=c(5,4,rep(5,4)),vm="Mh",vh="Chao")
> rd.m1$model.fit
```

```
             deviance     df       AIC
fitted model 627.3967  38847  911.4528
```

```
> rd.m1$emig.fit
```

```
                                          deviance     df       AIC
model with homogeneous temporary emigration 627.2376  38846  913.2937
model with temporary emigration             621.8716  38843  913.9277
```

The test for temporary immigration is not significant (χ^2_4=5.53, pvalue=0.238) meaning that capture probabilities estimated with the Jolly-Seber model are not different from those estimated with the individual closed population models. The differences, on the logit scale, of the Jolly-Seber minus the closed population models capture probabilities can be obtained with

```
> rd.m1$emig.param
```

```
             estimate    stderr
period 2    0.5896720 1.1255156
period 3    0.7590532 0.7494904
period 4    0.7576857 0.9043055
period 5   -1.6842044 0.9378936
homogenous  0.1690180 0.4260088
```

Even in period 2 where the closed population model does not fit well, the difference on the logit scale is non significant (estimate=0.59, s.e.=1.12). Using Darroch's model to handle heterogeneity yields

```
> rd.m2<-robustd.0(mvole[,-10], vt=c(5,4,rep(5,4)),vm="Mh",vh="Darroch")
> rd.m2$model.fit
```

```
             deviance     df       AIC
fitted model  640.769  38857  904.8251
```

```
> rd.m2$emig.fit
```

```
                                          deviance     df       AIC
model with homogeneous temporary emigration 635.2724  38856  901.3285
model with temporary emigration             630.1530  38853  902.2091
```

```
> rd.m2$emig.param
```

```
             estimate    stderr
period 2    0.3922484 1.2188324
period 3    1.4190404 0.7481082
period 4    2.1327701 0.9108267
period 5   -0.4681385 0.8769895
homogenous  1.0120808 0.4397377
```

Now the deviance difference of 10.64 on 4 degrees of freedom for temporary immigration has a pvalue of 3.1%. With Darroch's model, the closed population estimates of the capture probabilities are significantly smaller than those obtained from the Jolly-Seber model. This cannot be interpreted as indicating a temporary emigration. This suggests that Darroch's model is not appropriate within primary periods.

We note that it is possible not to specify any model for the second period. It would be done with the following command.

```
> rd.m3<-robustd.0(mvole[,-10], vt=c(5,4,rep(5,4)),vm=c("Mh","none","Mh","Mh","Mh","Mh"),v
```

We have tried many models, and the smallest AIC is obtained with the Poisson model, with parameter a=1.5 within sessions.

```
> rd.m4<-robustd.0(mvole[,-10], vt=c(5,4,rep(5,4)),vm="Mh",vh="Poisson",vtheta=1.5)
```

As can be seen in the comparative tables printed below, the estimators of the demographic parameters obtained with the robust design are similar to those obtained with the Jolly-Seber model applied to the between primary period data.

```
> survivals<-data.frame(op.m1$survivals,rep("|",5),rd.m4$survivals)
> N<-data.frame(op.m1$N,rep("|",6),rd.m4$N)
> birth<-data.frame(op.m1$birth,rep("|",5),rd.m4$birth)
> Ntot<-data.frame(op.m1$Ntot,c("|"),rd.m4$Ntot)
> name<-c("estimate.open","stderr.open","|","estimate.robust","stderr.robust")
> colnames(survivals)<-colnames(N)<-colnames(birth)<-colnames(Ntot)<- name
> survivals
```

```
             estimate.open stderr.open | estimate.robust stderr.robust
period 1 -> 2     0.8195489  0.05653036 |       0.8228273    0.05516940
period 2 -> 3     0.5605845  0.06475147 |       0.5687647    0.06528882
period 3 -> 4     0.7011614  0.07268043 |       0.7261012    0.07697910
period 4 -> 5     0.5787844  0.06847395 |       0.5542232    0.06763400
period 5 -> 6            NA          NA |       0.9989390    0.09125118
```

```
> N
```

```
         estimate.open stderr.open | estimate.robust stderr.robust
period 1            NA          NA |        63.17356      4.639529
period 2      75.10048    2.591686 |        75.53404      2.024295
period 3      59.69964    3.789494 |        61.41466      3.861024
period 4      62.64063    3.320978 |        67.22767      4.227080
period 5      55.60073    3.229606 |        53.70764      2.044445
period 6            NA          NA |        92.35621      7.352186
```

```
> birth
```

```
              estimate.open stderr.open | estimate.robust stderr.robust
period 1 -> 2            NA          NA |        23.55311      6.422676
period 2 -> 3      17.59948    5.025697 |        18.45357      5.181694
period 3 -> 4      20.78154    5.391299 |        22.63441      5.972292
period 4 -> 5      19.34531    5.349862 |        16.44851      4.822338
period 5 -> 6            NA          NA |        38.70555      7.491438

> Ntot

            estimate.open stderr.open | estimate.robust stderr.robust
all periods       174.001     2.06165 |        182.9687      4.626267
```

5. Limitations and comparison to other softwares

One limitation of **Rcapture** is that it does not handle trap deaths. This occurs if some captured animals are not released in the population after their capture. Animals cannot be recaptured after a trap death so that their capture histories will have zeros for the remaining capture occasions. In closed population models, trap deaths can be considered as a subpopulation with a known size. The analysis can focus on the estimation of the size of the non trap death population, using the data on the animals that did not experience a trap death. In open population models, the goodness of fit statistics of the `openp` functions are valid in the presence of trap deaths. Their demographic parameter estimates are however biased; alternative formulas for converting loglinear parameters into demographic parameters need to be develop to account for trap deaths. A robust design analysis is also sensitive to the occurrence of trap deaths; they might bias its conclusions. Methods to deal with death traps with this software are under investigation.

The robust design functions highlight another limitation of the Poisson regression for modeling capture recapture data. In large experiments, the number of observable capture histories can be very large. Most of them have a zero frequency; still, all these zero frequencies must appear in the dependent vector for the Poisson regression. This makes the number of cases in the Poisson regression unnecessarily large. An alternative fitting strategy discussed in Barker and White (2004) is to model the capture histories of the released animals at each capture occasion using multinomial distributions. Then, only capture histories with a positive frequency contribute to the likelihood. For an open population model, the likelihood to maximize can be written in terms of the m-array matrix for the experiment. This fitting strategy is implemented in **Mark**, see White and Burnham (1999); White (2005), which is the main software for analyzing capture recapture data. Since the likelihood is written in terms of aggregated data, testing the fit of the model is not straightforward under this approach. The simple tests for trap dependence and the goodness of fit diagnostics based on residuals presented in Section 3 are not available anymore.

Over the years several softwares have been written for the analysis of data from capture recapture experiments. Several of these softwares are now available within **Mark**, see White and Burnham (1999); White (2005). For instance it contains the package **Popan** of Schwarz and Arnason (1996) for the modeling of abundance in open population models, see also http://www.cs.umanitoba.ca/~popan/. Package **Care** for closed populations data, with

an emphasis on epidemiological applications, is discussed in Chao, Tsay, Lin, Shau, and Chao (2001). Package **M-Surge**, see Choquet, Reboulet, Pradel, Gimenez, and Lebreton (2004), is also available to model multistate recapture data in the Cormack-Jolly-Seber setting. Bayesian methods can also be used to fit capture-recapture model, see Madigan and York (1997) and Brooks, Catchpole, and Morgan (2000). Durban and Elston (2005) suggest a Bayesian approach to M_{h}. Gibbs sampling and Markov chain Monte Carlo techniques are implemented for fitting complex models, see Gimenez, Crainiceanu, Barbraud, Jenouvrier, and Morgan (2006).

The package **Rcapture** covers the basic statistical models for capture-recapture experiments. It is the only package that focuses on the use of loglinear models for the analysis of closed and open population data. As illustrated in Section 3, the fit of complex models can be investigated with the maximum likelihood estimates and their asymptotic variances obtained from **Rcapture**. This package provides diagnostic tools and several alternatives for fitting model $\mathrm{M_h}$ and $\mathrm{M_{th}}$ to closed population data. In view of the lack of identifiability of such models pointed out by Link (2003), this flexibility is welcomed when confronting heterogeneity. **Rcapture** tries to emphasize that there is more to data analysis than model fitting by providing probability and residual plots to guide the analysis. It takes advantage of the flexible R programming environment which allows users to build their own R function by using multipurpose minimization functions such as `optim`. For instance a function `closedp.Mtb` for fitting closed population model $\mathrm{M_{tb}}$ which does not have a loglinear form is provided with the package as an illustration of the application of the R programming language to the building of models for capture-recapture data.

6. Acknowledgements

We are grateful to R. M. Cormack for providing copies of his most recent papers and to the Natural Sciences and Engineering Research Council of Canada for his support. We also want to acknowledge the contribution of Emmanuelle Manouvelle who wrote the first versions of many of the functions in this package.

References

Abeni D, Brancato G, Perucci C (1994). "Capture-Recapture to Estimate the Size of the Population with Human Immunodeficiency Virus type 1 Infection." *Epidemiology*, **5**, 410–414.

Agresti A (1994). "Simple Capture-Recapture Models Permitting Unequal Catchability and Variable Sampling Effort." *Biometrics*, **50**, 494–500.

Barker R, White G (2004). "Towards the Mother-of-all-Models: Customised Construction of the Mark-Recapture Likelihood Function." *Animal Biodiversity and Conservation*, **27.1**, 177–185.

Briand LC, Emam KE, Freimut BG, Laitenberger O (2000). "A Comprehensive Evaluation of Capture-Recapture Models for Estimating Software Defect Content." *IEEE Trans. Softw. Eng.*, **26**(6), 518–540. doi:http://dx.doi.org/10.1109/32.852741.

Brooks SP, Catchpole EA, Morgan BJT (2000). "Bayesian Animal Survival Estimation." *Statistical Science*, **15**(4), 357–376.

Chao A (1987). "Estimating the Population Size for Capture-Recapture Data with Unequal Catchability." *Biometrics*, **43**(4), 783–791.

Chao A, Tsay PK, Lin SH, Shau WY, Chao DY (2001). "Tutorial in Biostatistics: The Applications of Capture-Recapture Models to Epidemiological Data." *Statist. Med.*, **20**(3), 3123–3157.

Choquet R, Reboulet AM, Pradel R, Gimenez O, Lebreton JD (2004). "M-SURGE: New Software Specifically Designed for Multistate Capture-Recapture Models." *Animal Biodiversity and Conservation*, **27.1**, 207–215.

Cormack RM (1985). "Example of the Use of GLIM to Analyze Capture-Recapture Studies." In B Morgan, P North (eds.), "Lecture Notes in Statistics 29: Statistics in Ornithology," pp. 242–274. Springer-Verlag, New York.

Cormack RM (1989). "Loglinear Models for Capture-Recapture." *Biometrics*, **45**, 395–413.

Cormack RM (1992). "Interval Estimation for Mark-Recapture Studies of Closed Populations (Ack: V49 P315; Ref: 91StatMed V10 P717-721)." *Biometrics*, **48**(2), 567–576.

Cormack RM (1993a). "The Flexibility of GLIM Analyses of Multiple Recapture or Resighting Data." In JD Lebreton, P North (eds.), "Marked Individuals in the Study of Bird Population," pp. 39–49. Birkhaüser Verlag, Basel, Switzerland.

Cormack RM (1993b). "Variances of Mark-Recapture Estimates." *Biometrics*, **49**, 1188–1193.

Cormack RM, Jupp PE (1991). "Inference for Poisson and Multinomial Models for Capture-Recapture Experiments." *Biometrika*, **78**(4), 911–916.

Darroch JN, Fienberg SE, Glonek GFV, Junker BW (1993). "A Three-Sample Multiple-Recapture Approach to Census Population Estimation with Heterogeneous Catchability." *J. Am. Stat. Assoc.*, **88**, 1137–1148.

Durban JW, Elston DA (2005). "Mark-Recapture with Occasion and Individual Effects: Abundance Estimation Through Bayesian Model Selection in a Fixed Dimensional Parameter Space." *J. Agric. Biol. Environ. Stat.*, **10**(3), 291–305.

Ebrahimi NB (1997). "On the Statistical Analysis of the Number of Errors Remaining in Software Design Document After Inspection." *IEEE Trans. Softw. Eng.*, **23**, 529–532.

Gimenez O, Crainiceanu C, Barbraud C, Jenouvrier S, Morgan BJT (2006). "Semiparametric Regression in Capture-Recapture Modelling." *Biometrics*, **62**, 691–698.

Lindsay BG (1986). "Exponential Family Mixture Models (With Least-Squares Estimators)." *Ann. Statist.*, **14**, 124–137.

Link WA (2003). "Nonidentifiability of Population Size from Capture-Recapture Data with Heterogeneous Detection Probabilities." *Biometrics*, **59**(4), 1123–1130.

Madigan D, York JC (1997). "Bayesian Methods for Estimation of the Size of a Closed Population." *Biometrika*, **84**, 19–31.

Otis DL, Burnham KP, White GC, Anderson DR (1978). *Statistical Inference from Capture Data on Closed Animal Populations*, volume 62 of *Wildlife Monographs*. Wildlife Society.

Pollock KH (1982). "A Capture-Recapture Design Robust to Unequal Probability of Capture." *J. Wildl. Manage.*, **46**, 752–757.

Pollock KH, Nichols JD, Brownie C, Hines JE (1990). *Statistical Inference for Capture-Recapture Experiments*, volume 107 of *Wildlife Monographs*. Wildlife Society.

Rivest LP (2007). "Why a Time Effect Has a Limited Impact on Capture-Recapture Estimates in Closed Populations." *Canad. J. Statist.* Under revision.

Rivest LP, Baillargeon S (2007). "Applications and Extensions of Chao's Moment Estimator for the Size of a Closed Population." *Biometrics*, **to be published**.

Rivest LP, Daigle G (2004). "Loglinear Models for the Robust Design in Mark-Recapture Experiments." *Biometrics*, **60**(1), 100–107.

Rivest LP, Lévesque T (2001). "Improved Log-linear Model Estimators of Abundance in Capture-Recapture Experiments." *Canad. J. Statist.*, **29**(4), 555–572.

Sandland RL, Cormack RM (1984). "Statistical Inference for Poisson and Multinomial Models for Capture-Recapture Experiments." *Biometrika*, **71**(1), 27–33.

Schwarz CJ, Arnason AN (1996). "A General Methodology for the Analysis of Capture-Recapture Experiments in Open Populations." *Biometrics*, **52**, 860–873.

Schwarz CJ, Stobo WT (1997). "Estimating Temporary Migration Using the Robust Design." *Biometrics*, **53**, 178–194.

Seber GAF (1982). *The Estimation of Animal Abundance and Related Parameters*. Macmillan, New York, 2nd edition.

White G (2005). "Software Developed by Department of Fishery and Wildlife Biology and Colorado Coop. Fish and Wildlife Unit, Colorado State University." WWW. URL http://www.warnercnr.colostate.edu/~gwhite/software.html.

White G, Burnham KP (1999). "Program MARK: Survival Estimation from Populations of Marked Animals." *Bird Study*, **46 (Supplement)**, 120–138.

Williams BK, Nichols J, Conroy MJ (2002). *Analysis and Management of Animal Populations*. Academic Press, San Diego.

Wohlin C, Runeson P, Brantestam B (1995). "An Experimental Evaluation of Capture-Recapture in Software Inspection." *Softw. Test. Verif. Reliab.*, **5**, 213–232.

Affiliation:

Sophie Baillargeon & Louis-Paul Rivest
Département de mathématiques et de statistique
Université Laval
Québec (Québec), G1K 7P4, Canada
E-mail: sbaillar@mat.ulaval.ca
E-mail: lpr@mat.ulaval.ca
URL: http://www.mat.ulaval.ca/pages/lpr/