ELSEVIER

Available at WWW.MATHEMATICSWEB.ORG POWERED BY SCIENCE DIRECT®

Journal of Statistical Planning and Inference 113 (2003) 699–714

journal of statistical planning and inference

www.elsevier.com/locate/jspi

Population size estimation using local sample coverage for open populations

Richard Huggins[a,*], Hsin-Chou Yang[b], Anne Chao[b], Paul S.F. Yip[c]

[a] *Department of Statistics, School of Statistical Science, La Trobe University, 3083 Bundoora, Vic., Australia*

[b] *Institute of Statistics, National Tsing Hua University, Hsin-Chu 30043, Taiwan*

[c] *Department of Statistics and Actuarial Science, University of Hong Kong, Hong Kong*

Received 8 March 2001; received in revised form 25 September 2001; accepted 26 November 2001

Abstract

An unsolved problem in the analysis of capture–recapture experiments is the estimation of the size of an open population when the capture probabilities are heterogeneous across the population. Here, we extend a kernel smoothing approach of Huggins and Yip (Biometrics 55 (1999) 387) to the martingale estimating functions based on sample coverage of Chao et al. (J. Statist. Plann. Inference 92 (2001) 213) and solve this problem when there are frequent capture occasions. Simulation results are shown to examine the performance of the proposed estimation procedure. A real data set is used for illustration. © 2002 Elsevier Science B.V. All rights reserved.

MSC: primary 62P10; secondary 62G05

Keywords: Capture–recapture; Kernel function; Martingale; Open population; Sample coverage

1. Introduction

The capture–recapture models have been used for estimating parameters in biological populations. There are two types of models: closed and open. A closed model, which is usually valid for data collected in a short time, assumes that there are no additions (birth or immigration) and losses (death or emigration). The population size, which is the main parameter of interest, remains constant during the study period. An open model, which is used to model long-term investigation, allows for additions and losses

* Corresponding author. Tel.: +61-3-9479-1068; fax: +61-3-9479-2466.
E-mail address: r.huggins@latrobe.edu.au (R. Huggins).

0378-3758/02/$ - see front matter © 2002 Elsevier Science B.V. All rights reserved.
PII: S0378-3758(02)00093-9

Nomenclature

s	number of capture occasions
t_j	capture time for capture occasion j, $j = 1,2,\ldots,s$
N_t	population size at time t
X_{ij}	capture indicator, $X_{ij} = 1$, if the ith animal is captured on the jth occasion; $X_{ij} = 0$, otherwise
$k(t)$	the closest capture occasion to time t
K	a pre-selected number of capture occasions such that a locally closed assumption is applied to $2K + 1$ capture occasions
h	bandwidth (or smoothing parameter) of a kernel function. That is, for a fixed $k(t)$, $h = [t_{k(t)+K} - t_{k(t)-K}]/2$
$W(t)$	kernel function at time t
$\omega_j(t)$	$j = 1,2,\ldots,s$, a set of weights such that $\sum_{j=1}^{s} \omega_j(t) = 1$; $\omega_j(t)$ is non-zero if and only if the capture occasions t_j is in the interval $(t_{k(t)-K}, t_{k(t)+K})$
n_j	number of individuals caught on occasion j
$m_j(t)$	number of individuals marked on occasions $k(t)-K,\ldots,j-1$ and are recaptured on occasion j
$u_j(t)$	number of individuals never caught on occasions $k(t) - K,\ldots,j - 1$, and captured for the first time on occasion j
$\tilde{M}_j(t)$	number of distinct individuals caught on occasions $k(t) - K,\ldots,j - 1$ and still in the population on occasion j, with $\tilde{M}_{k(t)-K}(t) = 0$
$f_{ij}(t)$	number of individuals captured exactly i times on occasions $k(t) - K,\ldots,j$

so that population size varies with time in the experiment. The focus of this paper is on the population size estimation in an open model, although the topic of estimating survival probabilities is also of interest in the biological and ecological sciences.

For closed models, there are many approaches to estimate population size under various assumptions. A practical and important class of models is called heterogeneous models in which capture probabilities are allowed to vary among animals. Several authors have proposed estimators covering a wide range of statistical methodologies. See a recent review by Schwarz and Seber (1999) on animal abundance models, in general, and on this topic in particular.

For open models, a commonly used model is the Jolly–Seber model (Seber, 1982; Pollock et al., 1990). An unresolved problem with this model is the effect of unequal capture probabilities on the Jolly–Seber estimators under general conditions. Another problem with the Jolly–Seber model is that all emigration is permanent so that the animals may not leave and rejoin the population. It has been documented that the Jolly–Seber population size estimators are generally biased downwards if there is heterogeneity in the capture probabilities; see Carothers (1979), Hwang and Chao (1995)

and Pledger and Efford (1998). These authors have proposed approaches to take account of heterogeneity under various assumptions. For example, Hwang and Chao (1995) assumed permanent emigration, and multiplicative form of individual heterogeneity and time effect. Pledger and Efford (1998) extended Carothers (1979) simulation method under an "open but stable population" assumption, that is, a population with births and immigration replacing deaths and emigration such that the population size throughout the study is a constant. A basic idea in this paper is to treat the open models as a series of overlapped "locally closed" models (defined later) so that techniques valid for closed models can be applied to an open model. In this framework, some restrictive assumptions can be relaxed.

Huggins and Yip (1999) used kernel smoothing to extend weighted martingale methods developed to estimate the size of a closed population to open populations. They showed that their method performed better than the Jolly–Seber estimator when individuals could leave and re-enter the population. However, they assumed the capture probabilities were homogeneous across the population at each time point. Thus it is to be expected that their method would share the disadvantages of the Jolly–Seber method in the presence of heterogeneous capture probabilities. Chao et al. (2001) developed martingale-based estimating functions to estimate the size of a closed population if the capture probabilities are heterogeneous and Huggins and Chao (2002) have examined the asymptotic properties of these estimators. Here, we apply the kernel smoothing approach of Huggins and Yip (1999) to these latter estimating functions to estimate the size of an open population with heterogeneous capture probabilities under the assumption the population is locally closed and there are regular capture occasions. We show in simulations that this estimator does perform better than the Huggins and Yip (1999) estimator in the presence of heterogeneous capture probabilities.

As in Huggins and Yip (1999), we are again motivated by weekly banding data on the bird species *Prinia flaviventris* (yellow bellied prinia) collected at the Mai Po bird sanctuary in Hong Kong for a period of 34 weeks from September 1991 to April 1992, previously examined in Huggins and Yip (1999) and described there. Lin and Yip (1999) analyzed part of the data (January–April, 1992) using relevant covariates. *Prinia flaviventris* is a very common territorial species mainly inhabiting the reed beds in the swamp. The birds were captured in mist nets that were set in the reed beds. The analysis of Huggins and Yip (1999) revealed distinct seasonal behavior with a large population peak in October–mid-December representing the presence of juvenile birds after the breeding season. The population then decreased in late December as many of the juvenile birds moved out of the trapping area or died. Another peak in February–March represented increased activity prior to breeding. During the subsequent breeding season, the females were on their nests and were largely not catchable so that the catchable population again decreased. Thus catchability may be expected to depend on the mobility of the birds which may vary between age classes and the gender. Recall that model M_h assumes that the capture probability for any individual might be different and this probability remains fixed over time. Although age and gender information is available for some captured birds, the information is missing for most of the birds caught in the study period. Therefore, the covariates cannot be utilized for the whole data of 34 weeks. If it could be assumed that the population was closed

over a relatively short time and the capture probabilities were time homogeneous, then model M_h would be most appropriate to estimate the population size over this short time period.

Consider a capture release experiment with captures at times $0 < t_1 < \cdots < t_s$. For notational simplicity we suppose the capture occasions are evenly spaced, although this is not crucial. Let $k(t)$ denote the closest capture occasion to t. Our approach to estimate the population size at t, N_t, treats the capture data in the local time interval $[t_{k(t)-K}, t_{k(t)+K}]$, where K is a pre-selected value, as a closed model with a constant population size N_t.

We say a heterogeneous population is approximately locally closed if:

1. The capture probabilities of individuals arriving into the population are independent random variables from the distribution F with mean $E(p)$ and coefficient of variation (CV), γ, $\gamma = \{E[p - E(p)]^2\}^{1/2}/E(p)$ and these capture probabilities are independent from individual to individual.
2. The capture probabilities do not depend on t.
3. The size of the population at time t is $N_t = [N\eta(t)]$ where $\eta(t) > 0$ is a continuous function and is bounded away from 0 in the study time period, and $[x]$ denotes the closest integer to x. There exists a constant λ, such that $|\eta(t) - \eta(s)| < \lambda|s - t|$.
4. The individuals in the population behave independently of each other. Given any individual, captures at different occasions are independent.
5. In any interval, the number of marked individuals that die or are otherwise removed from the population tends to zero as the width of the interval tends to zero.
6. Marking does not affect the capture probabilities.
7. The probability of removal has the same distribution for individuals captured and released at a given capture occasion as for individuals captured and released before that occasion.

This definition extends the definition of Huggins and Yip (1999) to the case where the capture probabilities are heterogeneous. Assumption 1 concerns the distribution of the capture probabilities and was not required in the homogeneous case considered in Huggins and Yip (1999). This assumption is required for the closed population estimator of Chao et al. (2001) as is Assumption 4. Assumptions 3–6 are as in Huggins and Yip (1999). Assumptions 3 and 5 ensure the population size changes smoothly. Assumption 6 could be relaxed but for simplicity we concentrate on model M_h rather than model M_{bh}, which includes the behavioural response to capture. Similarly, Assumption 2 could be relaxed if one used the appropriate martingale estimating equations of Chao et al. (2001) or assumed the individual capture probabilities were smooth functions of t. However, this latter condition would require showing the existence of a stochastic process with the appropriate marginal distributions and this is beyond the scope of the present work. Assumption 7 allows us to estimate the number of marked animals that are actually in the population. This assumption is implicit in Huggins and Yip (1999) but was not explicitly stated there. It implies that the relationship between capture and removal does not change as a function of time and seems reasonable in practice. For example, it would be implied by the assumptions that the removal probabilities are

independent of the capture probabilities and that marking and release do not affect the removal probabilities.

Section 2 lists further notation. Our proposed estimating procedure is presented in Section 3. The weekly banding data collected in Mai Po bird sanctuary in Hong Kong is discussed and compared with other estimators in Section 4. Simulation results are reported in Section 5.

2. The weighted martingale estimators

The approach to estimating the population size at time t is to consider capture occasions close to t and weight the resulting estimating functions so that occasions closest to t have the greatest weight.

Let $X_{ij}=1$, if the ith individual is captured on the jth occasion and 0, otherwise. For a given t, $0 < t < t_s$, let $\omega_j(t)$, $j=1,2,\ldots,s$, be a set of weights such that $\sum_{j=1}^{s} \omega_j(t)=1$. We suppose that $\omega_j(t)$ is non-zero if and only if the time t_j for capture occasion j is in $(t_{k(t)-K}, t_{k(t)+K})$. In this paper, the weights are determined by a pre-specified kernel function $W(t)$ and a bandwidth h, specifically,

$$\omega_j(t) = W\{(t-t_j)/h\} \Big/ \sum_{m=1}^{s} W\{(t-t_m)/h\} .$$

2.1. Estimating the number of marked individuals

Huggins and Yip (1999) noted that whilst arguments based on locally closed populations may be applied as the time between occasions decreases, in practice there is a non-negligible time interval between captures. Thus the number of individuals that had been marked may overestimate the number of marked individuals in the population. Before constructing our population size estimators we examine the estimation of the number of marked individuals in the population. We have supposed that the probability of removal has the same distribution for individuals captured and released at a given capture occasion as for individuals captured and released before that occasion.

Let $I[A]$ be the indicator function of the event A. For $k(t)-K \leqslant j \leqslant k(t)+K$, let

$$r_j(t) = \sum_{i=1}^{n_j} I\left[\sum_{\ell=j+1}^{k(t)+K} X_{i\ell} > 0\right]$$

denote the number of animals captured on the jth occasion and released immediately, and caught again at least once on occasions $j+1,\ldots,k(t)+K$. Let

$$z_j(t) = \sum_{i=1}^{\tilde{M}_j(t)-m_j(t)} I\left[\sum_{\ell=j+1}^{k(t)+K} X_{i\ell} > 0\right]$$

denote the number of animals captured at least once on occasions $k(t)-K,\ldots,j-1$, not caught on occasion j, and caught again at least once on occasions $j+1,\ldots,k(t)+K$.

Due to our Assumption 7, the conditional expectations of $r_j(t)$ and $z_j(t)$ for any j can be derived as follows:

$$E[r_j(t) \mid n_j] = n_j E[1 - (1 - p)^{k(t)+K-j}]$$

and

$$E[z_j(t) \mid \tilde{M}_j(t), m_j(t)] = [\tilde{M}_j(t) - m_j(t)]E[1 - (1 - p)^{k(t)+K-j}],$$

where the expectation is taken with respect to the distribution F. A natural estimating equation can be constructed as

$$E\{z_j(t)n_j - r_j(t)[\tilde{M}_j(t) - m_j(t)] \mid \tilde{M}_j(t), m_j(t), n_j\} = 0$$

and results in the estimator

$$\bar{M}_j(t) = \frac{\sum_{\ell=k(t)-K}^{k(t)+K} \omega_\ell(t_{k(t)-K+j})[z_\ell(t)n_\ell + r_\ell(t)m_\ell(t)]}{\sum_{\ell=k(t)-K}^{k(t)+K} \omega_\ell(t_{k(t)-K+j})r_\ell(t)} \tag{1}$$

of the number of individuals in $k(t) - K, \ldots, j - 1$ that are still in the population on occasion j.

2.2. Local sample coverage and the CV

In a capture–recapture experiment on a closed population of size N with capture occasions $1, 2, \ldots, \tau$ and individual capture probability p_i for the ith animal, the sample coverage for individuals captured in the first j capture occasions was defined by Chao et al. (2001) as

$$C_j = \sum_{i=1}^{N} p_i I\left[\sum_{k=1}^{j} X_{ik} > 0\right] \Big/ \sum_{i=1}^{N} p_i, \quad j = 1, \ldots, \tau.$$

The concept of sample coverage was originally proposed by Good (1953). We first extend this definition to neighborhoods of a time point for open populations. For $k(t) - K \leqslant j \leqslant k(t) + K$, we define the local sample coverage at time t conditional on the capture probabilities $\{p_1, p_2, \ldots, p_{N_t}\}$ as

$$C_j(t) = \sum_{i=1}^{N_t} p_i I\left[\sum_{\ell=k(t)-K}^{j} X_{i\ell} > 0\right] \Big/ \sum_{i=1}^{N_t} p_i, \tag{2}$$

where both summations for i are over the individuals in the population at time t. This is just the sample coverage of Chao et al. (2001) for a capture–recapture conducted on a closed population on the capture occasions in $[t_{k(t)-K}, t_{k(t)+K}]$. To estimate the sample coverage, let $f_{ij}(t)$ denote the number of individuals captured exactly i times on occasions $k(t) - K, \ldots, j$. A simple estimator of sample coverage is

$$\hat{C}_{j-1}(t) = 1 - f_{1j}(t) \Big/ \sum_{\ell=k(t)-K}^{j} n_\ell \,. \tag{3}$$

This type of estimator is shown to be working well even in heterogeneous closed populations; see Esty (1986).

Using a similar derivation to that given in Lee and Chao (1994) in closed population, we obtain the following estimator for the square of coefficient of variation:

$$\hat{\gamma}^2(t) = \max\left\{ \frac{\hat{N}_0(t)(2K+1)\sum_{j=1}^{2K+1} j(j-1) f_{j,k(t)+K}(t)}{2K[\sum_{j=1}^{2K+1} j f_{j,k(t)+K}(t)]^2} - 1, 0 \right\}, \tag{4}$$

where

$$\hat{N}_0(t) = \frac{\sum_{i=1}^{N_t} I[\sum_{j=k(t)-K}^{k(t)+K} X_{ij} > 0][1 + x/(2K+1)]}{1 - f_{1,k(t)+K}(t)/\sum_{j=k(t)-K}^{k(t)+K} n_j}$$

is a simple estimator of population size under homogenous closed population assumption, where x is the number of "missing" occasions due to an edge effect. For example, if $k(t) = 2$ and $K = 3$, we only have five occasions and two occasions are "missing", so $x = 2$ and the term $[1 + x/(2K+1)]$ is an adjustment factor. Only in the relatively short time interval in the beginning and finishing time, the value of x is non-zero.

2.3. Weighted estimating functions

For $k(t) - K \leqslant j \leqslant k(t) + K$, define $M_j^*(t) = N_t C_{j-1}(t)$ and $\bar{p}(t) = \sum_{i=1}^{N_t} p_i/N_t$. If there is no heterogeneity among capture probabilities, then $M_j^*(t) = \tilde{M}_j(t)$. Following Chao et al. (2001) and Huggins and Chao (2002) noted that $u_j(t) - [N_t - M_j^*(t)]\bar{p}(t)$ and $n_j - N_t\bar{p}(t)$ form a sequence of approximate martingale differences with respect to F_{j-1}, the σ-field generated by the capture histories up to occasion $j-1$. The weighted estimating functions are of the form

$$\sum_{j=k(t)-K}^{k(t)+K} \omega_j(t)[1 - C_{j-1}(t)]^{-1}\{u_j(t) - [N_t - M_j^*(t)]\bar{p}(t)\} = 0 \tag{5}$$

and

$$\sum_{j=k(t)-K}^{k(t)+K} \omega_j(t)[n_j - N_t\bar{p}(t)] = 0. \tag{6}$$

The latter equation yields an estimator

$$\hat{\bar{p}}(t) = \sum_{j=k(t)-K}^{k(t)+K} \omega_j(t) n_j/N_t. \tag{7}$$

We follow Chao et al. (2001) and Lee and Chao (1994) and estimate $M_j^*(t)$ by $\hat{M}_j^*(t) = \bar{M}_j(t) + \hat{\gamma}^2(t) f_{1,j-1}(t)$, where $\bar{M}_j(t)$ is the smoothing estimator given in Eq. (1). Hence, substituting (7) into (5) and replacing the unknown quantities by

their estimates, we have the estimator

$$\bar{N}_t = \frac{\sum_{j=k(t)-K}^{k(t)+K} \omega_j(t)\hat{M}_j^*(t)/[1-\hat{C}_{j-1}(t)]}{\sum_{j=k(t)-K}^{k(t)+K} \omega_j(t)[1-\eta_t^{-1}u_j(t)]/[1-\hat{C}_{j-1}(t)]}, \tag{8}$$

where $\eta_t = \sum_{j=k(t)-K}^{k(t)+K} \omega_j(t)n_j$. The weight function $\omega_j(t)$ is determined by the pre-specified kernel function.

However, the estimator derived from the smoothing estimating equations may not be smooth. Similar to the treatment in Huggins and Yip (1999), a second smoothing procedure has been applied to improve smoothness of estimates. Therefore, a double-smoothed estimator can be written as

$$\hat{N}_t = \sum_{j=k(t)-K^*}^{k(t)+K^*} \omega_j(t)\bar{N}_t, \tag{9}$$

where K^* is a pre-specified number of occasions.

Following a similar derivation to that in Huggins and Yip (1999) for model M_t and that for model M_h in Huggins and Chao (2002), we obtain an estimator of the variance for population size estimator $\bar{N}_t$ as

$$\hat{s}_t^2 = \frac{\sum_{j=k(t)-K}^{k(t)+K} \omega_j(t)^2\hat{b}_j^2(t)\hat{N}_t\{\hat{\sigma}_j^2(t) + [\hat{c}_j(t)+\hat{g}_j(t)+\hat{\rho}_j(t)]^2\,\hat{\bar{p}}(t)[1-\hat{\bar{p}}(t)]\}\frac{\hat{N}_j}{\hat{N}_t}}{[\sum_{j=k(t)-K}^{k(t)+K} \omega_i(t)\hat{b}_j(t)\hat{d}_j(t)\frac{\hat{N}_j}{\hat{N}_t}]^2} \tag{10}$$

where $\hat{b}_j(t) = [1-\hat{C}_{j-1}(t)]^{-1}$, $\hat{d}_j(t) = u_j(t)/\hat{N}_j - \hat{\bar{p}}(t)$, $\hat{c}_j(t) = \hat{M}_j^*(t)(\hat{N}_t^{-1} - \hat{N}_j^{-1})$, $\hat{\sigma}_j^2(t) = \hat{\bar{p}}(t)[1-\hat{C}_{j-1}(t)]\{1-[1+\hat{\gamma}(t)^2]\hat{\bar{p}}(t)\}$, $\hat{\rho}_j(t) = \hat{\gamma}^2(t)f_{j,k(t)+K}/\hat{N}_j$ and

$$\hat{g}_j(t) = \frac{\sum_{\ell=k(t)-K}^{k(t)+K} \omega_\ell(t)r_\ell(t)\cdot[\hat{M}_\ell^* - \hat{M}_j^*]/\hat{N}_\ell}{\sum_{\ell=k(t)-K}^{k(t)+K} \omega_\ell(t)r_\ell(t)}.$$

Then a second smoothing procedure is applied to the variance formula (10) to obtain a final variance estimator for the estimator $\hat{N}_t$.

3. Example

As indicated in the Introduction section, we were motivated by the capture data set of *Prinia flaviventris* collected from the Mai Po bird sanctuary for a period of 34 weeks from September 1991 to April 1992. In several weeks no banding was conducted because of the weather condition. A total of 216 birds were captured in this period of which 163 were only captured once, 45 were captured twice, 6 captured 3 times, 1 captured four times and one captured 6 times. Detailed trapping data are given in Huggins and Yip (1999).

Huggins and Yip (1999) analyzed these data by adopting a model M_t, that is, all birds are assumed to have the same capture probabilities. Lin and Yip (1999) focused only on the data with covariates and found that weight has little effect but gender has

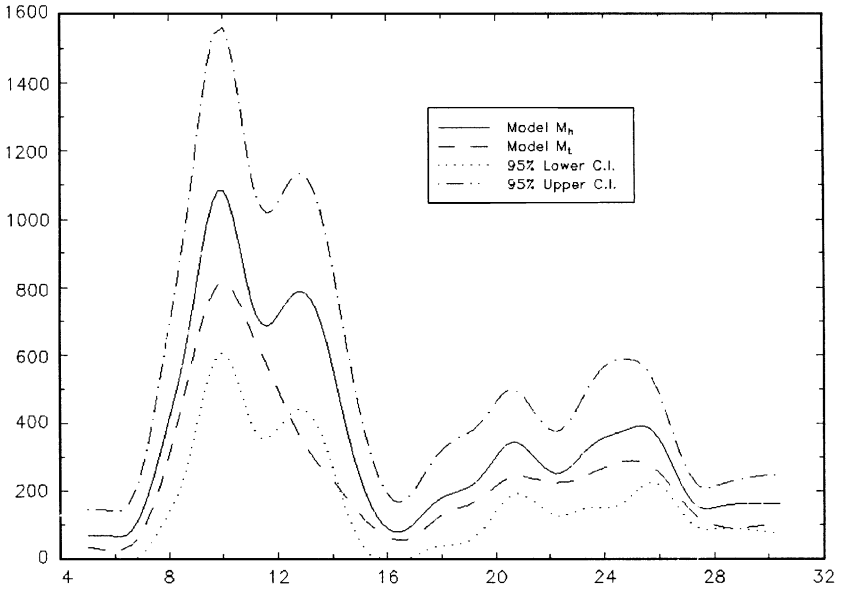

Fig. 1. Behavior of population size point and interval estimates for quartic kernel, $K = h = 4$.

a significant effect on the capture probability. Since relevant covariates were missing for most captures, we are interested in analyzing the whole series of weekly data by adopting a heterogeneous model but avoiding the use of covariates. Moreover, the data also show strong evidence of heterogeneity, as will be discussed below.

We selected two local sizes ($K = h = 4$ and 5) and three kernel functions (quartic, Gaussian and Epanechnikov). For $K = h = 4, 130$ evenly spaced grid points were chosen and 120 grid points were used for $K = h = 5$. The estimated CV given in Eq. (4) and population size given in Eq. (8) for each grid point were then obtained. For the case of $K = h = 4$, the CV estimates are in the range of 0.00–1.00 with an average value of 0.49 and standard error 0.29. For the case of $K = h = 5$, the CV estimates are in the range of 0.00–1.06 with an average value of 0.51 and standard error 0.32. These relatively large magnitudes of CV estimates have suggested that the heterogeneity is significant and cannot be ignored. Therefore, we feel a heterogeneous model is a more proper model for these data. A second smoothing bandwidth $h^* = 1.5$ was adopted.

Three types of kernel functions resulted in similar estimates, but the bandwidth $K = h = 4$ produced higher estimates than those using $K = h = 5$. In Figs. 1 and 2, we plot the second-smoothed population size estimates for model M_h over time for quartic kernel for $K = h = 4$ (Fig. 1, solid curve) and $K = h = 5$ (Fig. 2, solid curve). In addition, using the same smoothing technique and the formula in Eq. (10), we also plot in each figure the corresponding 95% confidence interval for model M_h over time (dotted curve for lower limit and dot-dash curve for upper limit) to show the precision of the proposed estimator.

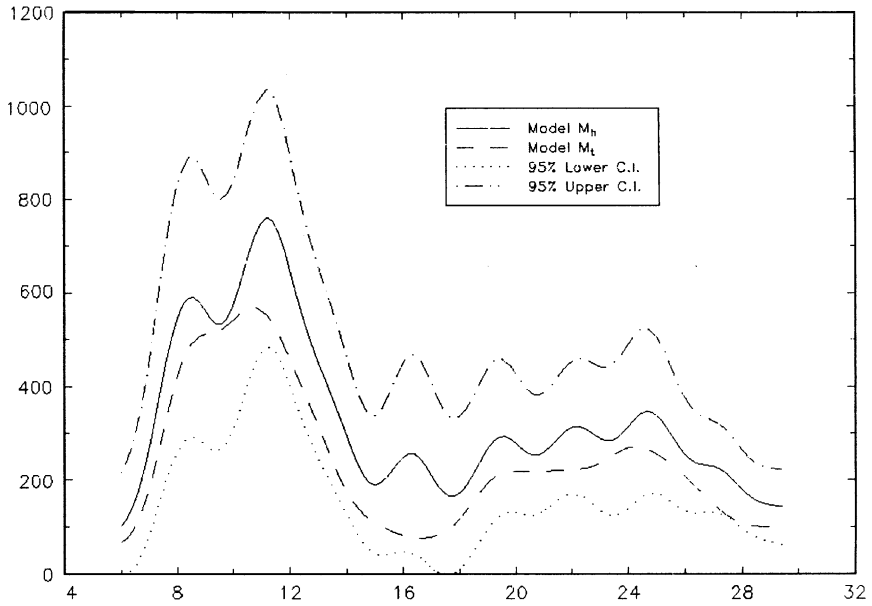

Fig. 2. Behavior of population size point and interval estimates for quartic kernel, $K = h = 5$.

For comparison, the plot of the second smoothed estimates under model M_t is also shown in Figs. 1 and 2 (dash curve) based on the following form (Huggins and Yip, 1999):

$$\tilde{N}_t = \frac{\sum_{j=k(t)-K}^{k(t)+K} \omega_j(t)\bar{M}_j(t)n_j}{\sum_{j=k(t)-K}^{k(t)+K} \omega_j(t)[n_j - u_j(t)]}. \tag{11}$$

All the estimates obtained from model M_h are higher than those based on model M_t. This is consistent with all previous finding that estimators without incorporating individual heterogeneity are negatively biased by the variation of capture probability (Yip et al., 1999). The general trend of the population estimates under the two models remains consistent. There exist two peaks during the whole experiment period. As stated in the Introduction, the first peak occurs in October–December since the presence of juvenile birds after the breeding season. And then, the second peak appears in February–March when the birds increase activity prior to breeding season. Lin and Yip (1999) considered the second half period as a closed population and obtained a rough estimate of 454 based on 58 captures with gender information. Their finding is close to our results as shown in Fig. 1.

4. Simulation results

A limited simulation study was performed to examine the performance of the estimation procedure under open heterogeneous populations. We considered death-only

Table 1
Simulation comparison for death-only models with beta type of heterogeneity (500 simulation trials)

Kernel function		Quartic kernel					
Bandwidth		$K = h = 3$		$K = h = 4$		$K = h = 5$	
Models		M_t	M_h	M_t	M_h	M_t	M_h
Beta $(10, 10)$	Mean error	−24.3	−11.8	−20.0	−7.0	−16.1	−2.5
$E(p) = 0.500$	Mean s.e.	16.6	10.3	15.2	9.1	14.3	8.5
CV = 0.218	Mean RMSE	29.6	16.1	25.4	12.4	21.9	10.3
Beta $(10, 20)$	Mean error	−28.9	−12.4	−25.9	−7.9	−22.3	−2.4
$E(p) = 0.333$	Mean s.e.	20.1	12.4	18.4	10.9	17.3	9.8
CV = 0.254	Mean RMSE	35.4	18.3	32.0	14.6	28.4	11.7
Beta $(5, 5)$	Mean error	−27.9	−12.9	−23.9	−8.3	−19.6	−2.9
$E(p) = 0.500$	Mean s.e.	18.9	11.0	17.2	9.4	16.3	8.6
CV = 0.302	Mean RMSE	33.9	17.3	29.7	13.4	25.7	10.3
Beta $(5, 8)$	Mean error	−32.5	−14.7	−29.1	−9.4	−25.0	−3.4
$E(p) = 0.385$	Mean s.e.	22.0	12.5	20.2	10.6	19.0	9.4
CV = 0.338	Mean RMSE	39.4	19.8	35.6	15.0	31.6	11.4
Beta $(4, 8)$	Mean error	−37.7	−17.5	−33.2	−10.5	−30.5	−5.0
$E(p) = 0.333$	Mean s.e.	25.1	14.4	23.1	11.5	22.1	10.4
CV = 0.392	Mean RMSE	45.4	23.0	40.5	16.2	37.8	12.8
Beta $(3, 5)$	Mean error	−39.0	−18.8	−34.6	−11.7	−30.7	−5.4
$E(p) = 0.375$	Mean s.e.	26.0	14.6	23.8	11.6	22.7	10.2
CV = 0.430	Mean RMSE	47.0	24.1	42.1	17.0	38.3	12.7
Beta $(3, 10)$	Mean error	−47.1	−22.8	−43.4	−15.7	−41.1	−10.0
$E(p) = 0.231$	Mean s.e.	31.4	18.6	29.4	15.4	28.3	13.4
CV = 0.489	Mean RMSE	56.7	29.8	52.5	22.7	50.0	17.8
Beta $(1, 1)$	Mean error	−47.2	−24.7	−43.3	−19.6	−38.7	−13.3
$E(p) = 0.500$	Mean s.e.	30.8	17.3	28.8	14.8	27.7	13.3
CV = 0.577	Mean RMSE	56.4	30.3	52.0	24.8	47.7	19.3
Beta $(2, 10)$	Mean error	−57.0	−30.6	−55.9	−25.9	−53.2	−18.1
$E(p) = 0.167$	Mean s.e.	38.0	24.2	37.0	21.1	35.7	18.0
CV = 0.600	Mean RMSE	68.7	39.5	67.2	33.9	64.2	26.3
Beta $(0.5, 0.5)$	Mean error	−58.2	−34.1	−53.7	−29.3	−49.6	−23.4
$E(p) = 0.500$	Mean s.e.	37.4	22.7	35.4	20.5	34.1	18.6
CV = 0.707	Mean RMSE	69.2	41.0	64.4	35.9	60.3	30.1

models. The initial population size was fixed to be 400 and the capture probabilities for these 400 birds were generated from beta distributions. We selected ten beta distributions as given in Table 1. The capture data were generated by using 20 evenly spaced trapping occasions and capture times were $1, 2, \ldots, 20$. Each bird has a survival rate

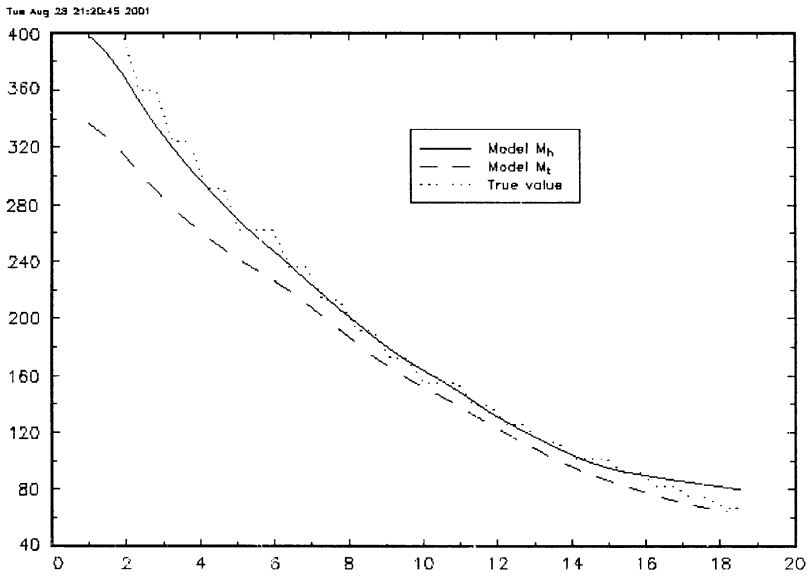

Fig. 3. Comparison of estimates for beta $(10, 10)$ ($K = h = 5$, quartic kernel, 500 simulated trials).

0.9 between any two occasions. Hence, the true population size is $N_t = 400 \times 0.9^{[t]-1}$, where $[A]$ denotes the largest positive integer less than A.

In the capture time interval $[1, 20]$, we divided it into 40 grid points and used these 40 grid points to compute population size estimates. To compare our results with those for model M_t, we also calculated the estimator given in Eq. (11) proposed in Huggins and Yip (1999).

As stated in Section 4, the effect of kernel function on estimates is limited, but the selection of bandwidth may be influential. We considered three bandwidths ($h = K = 3$–5) and the quartic kernel in the simulation. The bandwidth for the second smoothing was selected as $h^* = 1.5$. To avoid the impact of the edge effects, we only investigated the performance of the grid points 6–40. The average of the 35 biases yielded the "Mean error". Similarly, "Mean s.e." and "Mean RMSE" (root mean squared error) were also calculated by averaging over the 35 occasions. We generated 500 trials and the averages over these 500 trials are given in Table 1.

Table 1 shows that the estimates based on a model M_t have severe negative bias when the capture probabilities are generated from beta distributions. Our proposed estimates that incorporate the heterogeneity of capture probabilities are also biased downwards. The RMSE decreases when the bandwidth is increased. The simulation results indicate that a selection of bandwidth of $h = 5$ is preferable to other values.

In Figs. 3–6, we plot the estimators $\hat{N}_t$ (for model M_h) and the second smoothed estimator based on $\tilde{N}_t$ (for M_t) along with the true population size curve for all grid points for beta $(10, 10)$, beta $(5, 5)$, beta $(3, 5)$ and beta $(0.5, 0.5)$. As expected, the estimator $\tilde{N}_t$ that ignores the heterogeneity between animals severely underestimates the

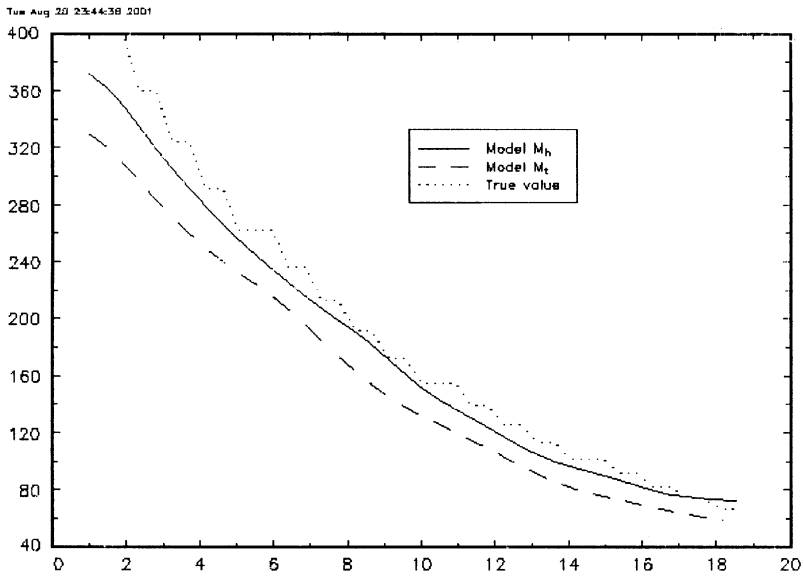

Fig. 4. Comparison of estimates for beta $(5,5)$ ($K = h = 5$, quartic kernel, 500 simulated trials).

true population size. In the beta type of heterogeneity, it is clear that our proposed estimator that takes heterogeneity into account performs better than the estimator without considering heterogeneity.

The simulations also reveal limitations on our method. Fig. 6 shows that for beta $(0.5, 0.5)$, substantial negative bias still exists for our estimator. A similar problem is found for beta $(1,1)$ distribution. See related discussion below.

5. Discussion

The simulation study reveals that our approach has advantages over the Huggins and Yip (1999) method in the presence of heterogeneous capture probabilities. However, if there is non-negligible fraction of un-catchable individuals in the population, then the estimator is biased. This is consistent with the analytic closed population results of Huggins and Chao (2002) for the closed population estimator upon which our approach is based, and the more general results of Huggins (2001). In the two beta cases, beta$(0.5, 0.5)$ and beta$(1,1)$, examined that resulted in a significant bias, the density does not vanish near zero, which implies a portion of the population is un-catchable. It seems no method can work well when there are un-catchable individuals and conditions such as those of Theorem 2 of Huggins (2001) or Norris and Pollock (1996) are required. Another limitation was that we require sufficient data to allow stable estimation of the CV. If there is not enough capture information, the resulting estimates would have low precision. This is also the case with any existing methods allowing for heterogeneity.

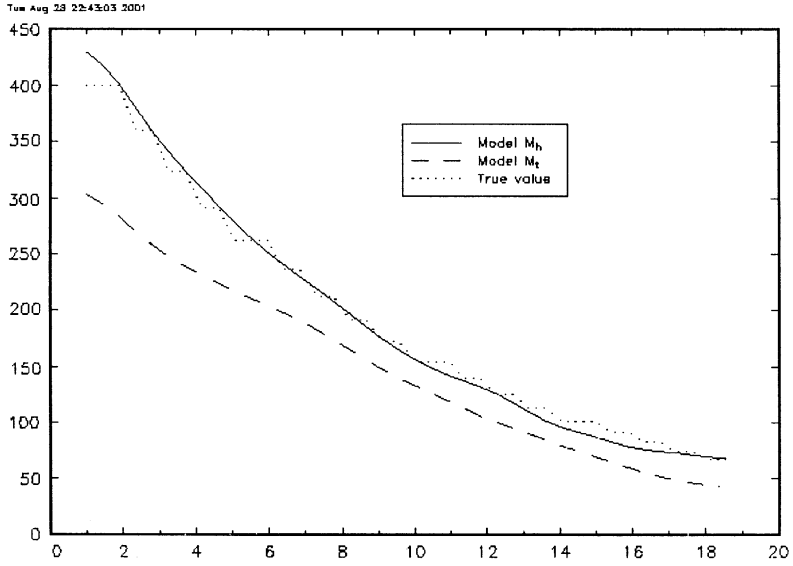

Fig. 5. Comparison of estimates for beta $(3,5)$ ($K = h = 5$, quartic kernel, 500 simulated trials).

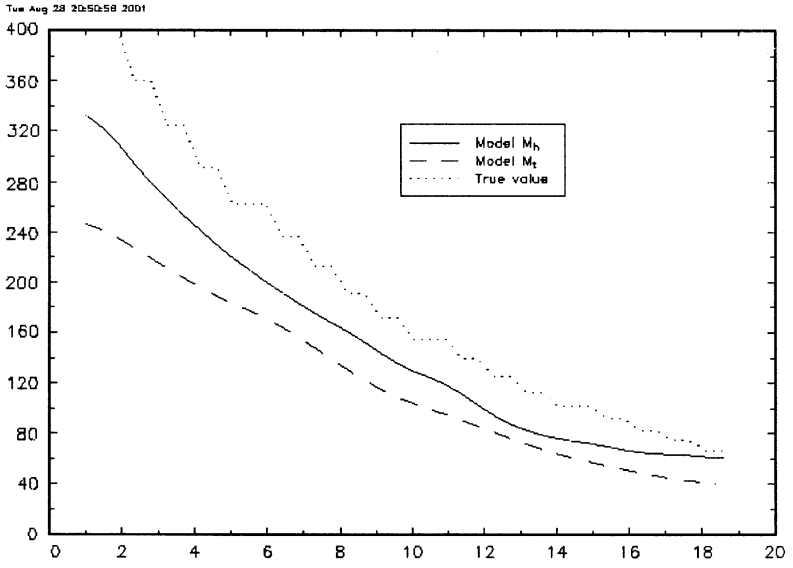

Fig. 6. Comparison of estimates for beta $(0.5, 0.5)$ ($K = h = 5$, quartic kernel, 500 simulated trials).

Pollock (1982) proposed a design robust to unequal probability of capture by using a combination of the Jolly–Seber estimators with closed population methods. In a robust design, there are a series of primary long-term sampling periods (for example, years) and several short-term samples (for example, consecutive days) under each primary period. Pollock's approach is to analyze the short-term data as a closed model in which heterogeneity may be allowed. The short-term data are then pooled to carry out a Jolly–Seber analysis to obtain survival estimates. However, the kernel smoothing methods generally cannot be applied to such a design because the "locally closed" assumption is usually not valid for samples across long-term periods. In our approach, locally closed populations might cover overlapped samples whereas in a robust design, only disjoint short-term samples are modeled as closed populations.

Our simulation studies suggest that the choice of kernel function used to determine the weights has little effect on the estimates. However, the choice of bandwidth is influential as is evident in comparing Figs. 1 and 2. Preliminary modeling using bandwidths ranging from 3 to 8 suggested that $h=4$ or $h=5$ was the most appropriate in our case. The main biological interest is in long-term trends in the population size and this can be revealed by estimates arising from a variety of bandwidths. We have deferred an analytic examination of the estimator, the objective determination of suitable bandwidths, and the extension of the method to model M_{bh}, which includes a behavioral response to capture, and the other models of Chao et al. (2001). However, the approach here combined with that of Huggins and Yip (1999) reveals that kernel smoothing methods have an important role to play in the analysis of capture–recapture experiments when there are frequent capture occasions and that further work, particularly into methods for determining suitable bandwidths will be worthwhile.

Acknowledgements

This research was completed when the first and the last authors visited National Tsing Hua University, Hsin-Chu, Taiwan. The visit for the first author was supported by an exchange program between Australian Academy of Science and Taiwan National Science Council.

References

Carothers, A.D., 1979. Quantifying unequal catchability and its effects on survival estimates in an actual population. J. Anim. Ecol. 48, 863–869.

Chao, A., Yip, P.S.F., Lee, S.-M., Chu, W., 2001. Population size estimation based on estimating functions for closed capture–recapture models. J. Statist. Plann. Inference 92, 213–232.

Esty, W.W., 1986. The efficiency of Good's nonparametric estimator. Ann. Statist. 14, 1257–1260.

Good, I.J., 1953. The population frequencies of species and the estimation of population parameters. Biometrika 40, 237–264.

Huggins, R.M., 2001. A note on the difficulties associated with the analysis of capture–recapture experiments. Statist. Probab. Lett. 54, 147–152.

Huggins, R.M., Chao, A., 2002. Asymptotic properties of an optimal estimating function approach to the analysis of mark recapture data. Comm. Stat. 31, 575–597.

Huggins, R.M., Yip, P.S.F., 1999. Estimation of the size of an open population from capture–recapture data using weighted martingale methods. Biometrics 55, 387–395.

Hwang, W.-D., Chao, A., 1995. Quantifying the effects of unequal catchabilities on Jolly–Seber estimators via sample coverage. Biometrics 51, 128–141.

Lee, S.-M., Chao, A., 1994. Estimating population size via sample coverage for closed capture–recapture models. Biometrics 50, 88–97.

Lin, D.Y., Yip, P.S.F., 1999. Parametric regression models for continuous time removal and recapture studies. J. Roy. Statist. Soc. Ser. B 61, 401–411.

Norris, J.L., Pollock, K.H., 1996. Nonparametric MLE under two closed capture–recapture models with heterogeneity. Biometrics 52, 639–649.

Pledger, S., Efford, M., 1998. Correction of bias due to heterogeneous capture probability in capture–recapture studies of open populations. Biometrics 54, 888–898.

Pollock, K.H., 1982. A capture–recapture design robust to unequal probability of capture. J. Wildl. Manage. 46, 752–757.

Pollock, K.H., Nichols, J.D., Brownie, C., Hines, J.E., 1990. Statistical inference for capture–recapture experiments. Wildl. Monogr. 107, 97pp.

Schwarz, C.J., Seber, G.A.F., 1999. A review of estimating animal abundance III. Statist. Sci. 14, 427–456.

Seber, G.A.F., 1982. The Estimation of Animal Abundance and Related Parameters, 2nd Edition. Griffin, London.

Yip, P.S.F., Zhou, Y., Lin, D.Y., Fang, X.Z., 1999. Estimation of population size based on additive hazards models for continuous time recapture experiments. Biometrics 55, 904–908.