

A new method for estimating the size of small populations from genetic mark–recapture data

CRAIG R. MILLER,* PAUL JOYCE† and LISETTE P. WAITS*

*Department of Fish & Wildlife, College of Natural Resources, PO Box 44-1136, University of Idaho, Moscow, ID 83844-1136,

†Department of Mathematics, Division of Statistics, University of Idaho, Moscow, ID 83844-1103

Abstract

The use of non-invasive genetic sampling to estimate population size in elusive or rare species is increasing. The data generated from this sampling differ from traditional mark–recapture data in that individuals may be captured multiple times within a session or there may only be a single sampling event. To accommodate this type of data, we develop a method, named *capwire*, based on a simple urn model containing individuals of two capture probabilities. The method is evaluated using simulations of an urn and of a more biologically realistic system where individuals occupy space, and display heterogeneous movement and DNA deposition patterns. We also analyse a small number of real data sets. The results indicate that when the data contain capture heterogeneity the method provides estimates with small bias and good coverage, along with high accuracy and precision. Performance is not as consistent when capture rates are homogeneous and when dealing with populations substantially larger than 100. For the few real data sets where N is approximately known, *capwire*'s estimates are very good. We compare *capwire*'s performance to commonly used rarefaction methods and to two heterogeneity estimators in program CAPTURE: M_h -Chao and M_h -jackknife. No method works best in all situations. While less precise, the Chao estimator is very robust. We also examine how large samples should be to achieve a given level of accuracy using *capwire*. We conclude that *capwire* provides an improved way to estimate N for some DNA-based data sets. *Capwire* is available at www.cnr.uidaho.edu/lecg/.

Keywords: capture heterogeneity, mark–recapture, microsatellites, non-invasive genetic sampling, population estimation

Received 11 October 2004; revision accepted 8 March 2005

Introduction

Estimating the size of wild populations plays a central role in managing harvested populations and conserving rare and endangered species. One of the most common ways to estimate population size has been to capture, mark, release and later recapture (or redetect) individuals. The advancement of genetic techniques has made it possible to capture an individual's DNA rather than the individual itself (Taberlet *et al.* 1999).

Scat and hair are the most common sources of DNA obtained non-invasively. The technique of hair snaring has been used to study populations of the brown bear (*Ursus arctos*), black bear (*Ursus americanus*) (e.g. Woods *et al.* 1999;

Mowat & Strobeck 2000), the hairy-nosed wombat (*Lasiorhinus krefftii*, Banks *et al.* 2003) and the marten (*Martes americanus*, Mowat & Paetkau 2002). Collecting scat to identify individuals and estimate population numbers was first used on coyotes (*Canis latrans*, Kohn *et al.* 1999) and has since been used to study red wolves (*Canis rufus*, Adams *et al.*, personal communications), grey wolves (*Canis lupus*, Creel *et al.* 2003), forest elephants (*Loxodonta cyclotis*; Eggert *et al.* 2003), the European badger (*Meles meles*, Frantz *et al.* 2003; Wilson *et al.* 2003), brown bears in Europe (Bellemain *et al.* 2005) and the Scandinavian wolverine (*Gulo gulo*, Flagstad *et al.* 2004). Population size has also been estimated in humpback whales (*Megaptera novaengliae*), using sloughed and biopsied skin as the source of DNA (Palsbøll *et al.* 1997). There are a number of potential benefits to genetic tagging over physical tagging (Taberlet *et al.* 1999). These include increasing the number of observations and thereby improving

Correspondence: Craig R. Miller, Fax: 208 885 9080; E-mail: mill8560@uidaho.edu

estimates, reducing stress and mortality, reducing capture bias caused by trap response, and shortening the sampling period to better approximate closure.

In traditional trap based mark-recapture studies, an individual may be captured only once per session. Estimating population size has focused on estimating the probability of capture for each individual in each session. An important difference in the data arising from DNA-based mark-recapture studies is that sampling is approximately done with replacement. That is, since an individual is not physically confined at any time, it may leave multiple hair tufts or scats at multiple locations during a sampling session. One option is to condense all such multiple captures to one capture per session as has sometimes been done (e.g. Banks *et al.* 2003; Frantz *et al.* 2003; Bellemain *et al.* 2005), but this potentially wastes information. At the extreme of this multiple-capture scenario are studies with only a single sampling session (e.g. Kohn *et al.* 1999; Eggert *et al.* 2003). In this case, a different approach is needed.

The purpose of this research is to develop a method for estimating population size when the data may contain multiple observations of an individual within a session. The performance of the proposed method is evaluated by analysing a wide variety of simulated data sets where the true population size is known. Some of these data sets are generated under the same simple model that underlies the method of analysis, but most of the focus is on data arising from more biologically reasonable models in which individuals move about in space, display differing fidelities to a home range, and deposit DNA at different rates. The same data are analysed using several of the available methods including rarefaction and estimators within program CAPTURE (Otis *et al.* 1978). We also present analyses of real genetic and traditional data sets using the proposed method. Based on our results, we make recommendations for current and future research on population estimation using non-invasive genetic sampling.

Methods

Proposed estimator (capwire)

The data are modelled as if they arose from S samples of size one from the population. We assume that all individuals are correctly and uniquely identified from their genotypes (i.e. there are no undetected genotyping errors and no two captured individuals have identical genotypes; Miller *et al.* 2002; McKelvey & Schwartz 2004). All draws are assumed to be independent and identically distributed (IID). Thus being captured does not affect an individual's probability of subsequent capture (e.g. there is no trap response). The resulting data is a multinomial vector of capture counts for each individual, $\vec{c} = c_1, c_2, \dots, c_T$, where T is the number of different individuals sampled. In the simple even capture

probability model (ECM), every individual is equally likely to be captured on each draw with probability one on the population size ($1/N$). The likelihood function with respect to N is the multinomial probability distribution,

$$L(N) = \left(\frac{N!}{T!(N-T)!} \right) \left(\frac{S!}{c_1!c_2! \dots c_T!} \right) \prod_{i=1}^T (1/N)^{c_i}$$

Taking the natural log and ignoring the combinatorial terms which involve only the data (constants) indicates that T and the total number of observations are sufficient statistics for finding the maximum-likelihood estimator (MLE):

$$\ln L(N) \propto \sum_{x=1}^N \ln(x) - \sum_{x=1}^{N-T} \ln(x) + \ln(1/N) \sum_{i=1}^T c_i \quad (\text{eqn 1})$$

It is well known that in real populations, individuals do not display equal capture probability (Burnham & Overton 1979). Several approaches have been proposed in classic mark-recapture modelling for dealing with this heterogeneity (e.g. Burnham & Overton 1979; Chao 1988). One such approach is to view the population as a mixture of individuals with distinct capture probabilities (Norris & Pollock 1996; Pledger 2000). Our method is based on the simplest of these mixture models in which there are two types of individuals (the two innate rates model, or TIRM). Let the relative capture probability of the harder to capture type B individuals be 1 and of the easier to capture type A individuals be α ($\alpha > 1$). Let the number in each class be N_A and N_B and note that $N_A + N_B = N$. For now, suppose that each sampled individuals' type (A vs. B) is observable. Let the number of sampled type A and type B individuals be T_A and T_B with their individual identities indexed by i_A and i_B . Then, the log-likelihood is similar to equation 1 except there are two subpopulations with differing capture probabilities:

$$\begin{aligned} \ln L(N_A, N_B, \alpha) \propto & \sum_{x=1}^{N_A} \ln(x) - \sum_{x=1}^{N_A-T_A} \ln(x) + \ln \left(\frac{\alpha}{\alpha N_A + N_B} \right) \sum_{i_A=1}^{T_A} c_{i_A} \\ & + \sum_{x=1}^{N_B} \ln(x) - \sum_{x=1}^{N_B-T_B} \ln(x) + \ln \left(\frac{1}{\alpha N_A + N_B} \right) \sum_{i_B=1}^{T_B} c_{i_B} \end{aligned} \quad (\text{eqn 2})$$

Of course, the types of sampled individuals are not observable. For computational simplicity and speed, we chose to assign each sampled individual the type that maximizes the overall likelihood.

With some data sets a more serious problem occurs when maximizing the parameters in equation 2 because the likelihood surface plateaus as α and N_B simultaneously increase. This specifically plagues 'sparse' data sets where many individuals are observed only once. For example, in a tiny data set with two individuals each captured once and two individuals captured five times each ($\vec{c} = 1,1,5,5$),

a good likelihood score is observed at the reasonable parameter values of $\alpha = 9$, $N_A = 2$, $N_B = 4$ ($N = 6$), while a slightly better score is obtained at the somewhat ridiculous values, $\alpha = 415$, $N_A = 2$, $N_B = 167$ ($N = 169$). While this problem can be circumvented by requiring that there are few single captures in the data, this would render the method inappropriate for many real data sets.

Instead, we address the problem by assuming that the sample is large enough that the capture count disparity between the seldom-captured and the often-captured individuals provides good information about the value of α . With α restricted, we can use equation 2 to estimate N . In practice, we accomplish this by initially finding the MLE of α under the assumption that the all individuals in the population have been sampled (i.e. $N_A = T_A$, $N_B = T_B$, $N = N_A + N_B$). Fixing this as the value of α , the MLEs of N_A and N_B are obtained. A single bias correction on α is then made from the fact that the expected total number of observations of type A individuals is equal to $[\alpha N_A S / (\alpha N_A + N_B)]$. Solving for α yields

$$\hat{\alpha} = \left(\hat{N}_B \sum_{i_A=1}^{T_A} c_{i_A} \right) / \left(\hat{N}_A S - \hat{N}_A \sum_{i_A=1}^{T_A} c_{i_A} \right) \quad (\text{eqn 3})$$

Finally, N_A and N_B are remaximized using this bias adjusted value of α . Confidence intervals are estimated using the parametric bootstrap. We refer to this proposed estimator as *capwire* because it is premised on *capture with replacement*. The mechanics of the estimator, including how sampled individuals are assigned a capture type, are provided in Appendix I.

Simulation study

Urn simulations. Simulations were conducted to check the performance of *capwire* when the model assumptions are met (i.e. when samples are IID). This was done using an urn model (Appendix I, step 10). N and S were both set at 25 throughout, and α was varied from 1 to 11 in increments of 2. For $\alpha > 1$, we considered the scenarios of $N_A = 12$, $N_B = 13$ and $N_A = 3$, $N_B = 22$. For each scenario we simulated 100 replicates. Method performance was based on four criteria. Letting r index the individual replicate and \hat{N} the estimate of N averaged over 100 replicates, these are (i) relative bias = $(\hat{N} - N)/N$, (ii) mean relative error = $1/[100(\sum_{r=1}^{100} |\hat{N}_r - N|/N)]$ (iii) coverage = proportion of 100 replicates where 95% confidence intervals (CIs) cover N , and (iv) median 95% CI width.

In the urn phase of the study we considered the use of a likelihood-ratio test (LRT) to determine if the ECM should be employed for a given data set instead of the TIRM (Pledger 2000). The likelihood was maximized under each model and the ratio of their likelihoods, Λ_{obs} calculated. The distribution of Λ was obtained by simulation: using

the MLE of N under the ECM, we generated 100 data sets, repeated the maximization process and calculated Λ_{sim} . The P value was the proportion of these 100 where $\Lambda_{\text{sim}} > \Lambda_{\text{obs}}$. When the P value was less than 0.1 we employed TIRM rather than ECM.

Grid simulations. In real mark-recapture studies individuals usually occupy distinct space, display heterogeneous behaviours, and samples are drawn from different locations. Consequently, samples are likely to be neither independent nor identically distributed. For *capwire* to be useful, it must be robust to these violations. Simulations were therefore conducted to explore how *capwire* performs on biologically reasonable data.

Simulations were designed to emulate a scat-based mark-recapture study. A square grid of specified size was populated with a specified number of individuals that were initially evenly spaced. Individuals then moved randomly about the grid depositing scats. Individuals were of two movement types (sedentary and transient) and three deposition rates (seldom, moderate and often) to yield six total types. Combinations of these six types were used to define 12 types of populations (Table 1). Population compositions were designed to introduce increasing amounts variance in movement pattern and deposition rate to the population. The details of how grid simulations were conducted are given in Appendix II. All simulations were of the same duration (4000 steps). After movement and deposition, parallel transects were established at regular, specified,

Table 1 Population composition distributions used in grid simulation study and corresponding code. Values in the table are probabilities that an individual is assigned as this type. Individual types are abbreviated as follows: first letter represents movement with S , sedentary and T , transient; last two letters represent deposition rate with SD , seldom depositor; MD , moderate depositor and OD , often depositor. Hence 'SSD' represents sedentary, seldom depositor. See Appendix IIB for details

Code	Type of Individual					
	SSD	TSD	SMD	TMD	SOD	TOD
G1			1.0			
G2				1.0		
G3			0.5	0.5		
G4			0.8	0.2		
G5	0.5				0.5	
G6	0.33		0.34		0.33	
G7		0.5				0.5
G8		0.34		0.33		0.33
G9	0.5					0.5
G10	0.8					0.2
G11	0.25	0.25			0.25	0.25
G12	0.167	0.166	0.167	0.166	0.167	0.167

intervals along a cardinal direction. All scats encountered were collected and genotyped correctly and uniquely to the individual. If the number collected was less than the specified sample size, the simulation was restarted. If more were collected, observations were removed from the data at random. The data were summarized as the number of observations of each sampled individual and analysed.

An initial set of simulations was conducted for all 12 compositions (Table 1) with both N and $S = 25$. Subsequent simulations focused on four of these (G1, G4, G10, and G12), providing a gradient from zero (G1) to very high (G12) heterogeneity. These subsequent simulations were conducted at $N = 16, 49, 100$ and 196 . Grid size was always scaled to maintain overall density at 1 individual/10 000 steps². Populations were sampled at round numbers approximating $1/2N$, N , and $2N$. Transects were spaced as follows so that the desired sample size was obtained in most simulations: G1 = 40 steps; G4 = 30; G10 = 20; G12 = 30. For each set of conditions 100 replicate simulations were performed. Data were analysed with *capwire* using 95% confidence intervals and only the TIRM (i.e. no LRT for model selection was conducted, see Results).

A subset of the simulated data was also analysed with several other methods that have been applied to genetic mark–recapture data. At $N = 16, 49$, and 100 we analysed the grid simulated data using rarefaction. In this technique the data are plotted as the number of distinct individuals observed (y) as a function of sample size (x) and then fit to a curve. The asymptote of the curve provides an estimate of the population size. We employed both the hyperbolic function proposed by Kohn *et al.* [1999; $y = \alpha x / (\beta + x)$] and the exponential equation of Eggert *et al.* [2003; $y = \alpha(1 - e^{-\beta x})$]. In both equations α represents the asymptote and β determines the rate of increase. For point estimates the ‘observed’ curve was obtained by finding the number of distinct individuals expected if the sample was subsampled at size x for $x = 1$ to S (Comps *et al.* 2001). To obtain confidence bounds we randomized the order of observations 100 times and fit the equations to each accumulation pattern. After sorting, the 3rd and 98th estimates of α were used as bounds. Equations were fit using least squares regression.

For the same subset of simulations, we also examined the performance of the two closed-population heterogeneity estimators available in program CAPTURE: M_h -jackknife (Burnham & Overton 1979) and M_h -Chao (Chao 1988). These estimators assume sampling sessions during which an individual is either caught once, or not at all. However, the simulations contained no sampling sessions. The data were therefore rigged to fit the estimators by defining the number of sampling sessions for a given data set equal to the largest number of times any individual in the data set was captured (i.e. the most caught individual was captured every imaginary session). The rarefaction and CAPTURE estimators

were evaluated using the four criteria described in the urn simulation section above.

Analysis of real data. The most important test of an estimator is its performance on real data, but true population size is rarely known in mark–recapture studies. Among DNA-based studies, N was approximately known in only two cases: a population of European badgers (Frantz *et al.* 2003; Wilson *et al.* 2003) and a population of red wolves (Adams *et al.*, personal communication). We analysed both data sets using *capwire*. We also analysed DNA-based count data from a population of forest elephants (Eggert *et al.* 2003) and a population of northern hairy-nosed wombats (Banks *et al.* 2003). In the wombat study (Banks *et al.* 2003), there were nightly capture sessions. In their analysis, Banks *et al.* reduced multiple within-night captures to one in order to meet CAPTURE assumptions. To explore the effect that such data reduction has on estimates and confidence intervals in *capwire*, we analysed the data both with multiple nightly captures included and excluded.

Traditional mark–recapture studies do not sample with replacement and therefore violate a basic assumption of *capwire*. However, if the number of sessions conducted is not small, this violation may become unimportant. To analyse these data sets using *capwire*, all session information was ignored, and the data were simply defined as the total number of times each individual was caught in the study. A handful of traditional mark–recapture data sets exist for which the true N is known. These include two populations of cottontail rabbits (Edwards & Eberhardt 1967), a population of eastern chipmunks (Mares *et al.* 1981), two populations of striped skunks (Greenwood *et al.* 1985) and a population of taxicabs in Edinburgh, Scotland (Carothers 1973). For all data sets in which N was known (traditional and DNA-based), the four performance criteria described in the urn simulation section were used.

Results

Urn simulations

The urn simulations indicate that when samples are independent and all individuals have equal capture probability ($\alpha = 1$), using the LRT in *capwire* yields approximately unbiased estimates with good coverage (Table 2). This is because the correct ECM is selected most of the time. Unfortunately, when the population has capture heterogeneity ($\alpha > 1$), ECM is rejected in favour of the correct TIRM model only one-fourth to one-third of the time at this sample size. As with ECMs in traditional mark–recapture methods (e.g. Lincoln–Peterson), using ECM in *capwire* when heterogeneity exists yields underestimates of N (see α values ≥ 5). Thus, the low power of small samples can lead to incorrect model selection and consequent underestimation. This is a serious

Table 2 Performance of *capwire* on urn simulated data with $N = 25$, $S = 25$, 100 replicates per scenario. When $\alpha > 1$, $N_A = 13$ and $N_B = 12$; when $\alpha = 1$, capture classes do not exist

Mod sel*	α	\tilde{N}	Rel bias†	CI Width‡	Coverage of 95% CIs§	MRE¶	Cor mod**
LRT	1	26.2	0.05	26	0.96 (0.03, 0.01)	0.23	0.89
	3	25.2	0.01	25	0.92 (0.08, 0)	0.27	0.35
	5	21.0	-0.16	15.5	0.75 (0.25, 0)	0.25	0.25
	7	19.2	-0.23	14.5	0.65 (0.35, 0)	0.28	0.26
	9	18.0	-0.28	11.5	0.55 (0.45, 0)	0.30	0.22
	11	17.4	-0.31	8.5	0.45 (0.55, 0)	0.33	0.25
TIRM	1	31.1	0.24	30	0.98 (0, 0.02)	0.31	n/a
	3	28.6	0.14	29	0.99 (0, 0.01)	0.26	n/a
	5	24.5	-0.02	24	0.96 (0.04, 0)	0.21	n/a
	7	22.4	-0.10	22	0.95 (0.05, 0)	0.20	n/a
	9	21.1	-0.16	19	0.84 (0.16, 0)	0.23	n/a
	11	19.9	-0.20	18	0.81 (0.19, 0)	0.24	n/a

*Model selection process: LRT, use ECM unless rejected using LRT at the 0.1 level, and then use TIRM; TIRM, skip LRT and use TIRM regardless of likelihood under ECM.

†Relative bias; ‡median width of 95% confidence interval; §numbers in parentheses are proportions of replicates where CI falls below and above N , respectively; ¶mean relative error.

**Correct Model: proportion of replicates where correct model was selected (ECM when $\alpha = 1$, TIRM when $\alpha > 1$). Not applicable (n/a) when LRT not used to select a model.

concern since most real data sets will contain capture heterogeneity. We were therefore compelled to investigate the performance of *capwire* when no LRT is conducted and TIRM is used irrespective of the consistency of the data with ECM.

Using TIRM when capture probability is even across individuals yields overestimates of N (Table 2). However, it does not reduce coverage when $\alpha = 1$; it substantially improves coverage for $\alpha > 1$; it reduces mean relative error for $\alpha > 1$; and it reduces bias for $\alpha > 3$. As α becomes large, N is still underestimated. Similar patterns are observed in simulations with $N_A = 3$ and $N_B = 22$ (data not shown). Because we view it as more important to improve estimates in the common event that capture heterogeneity exists and accept poorer estimates in the rare event that capture probability is even than the converse, we used TIRM in all subsequent analyses with *capwire*.

Grid simulations

In the grid simulations, composition distributions G1 to G4 represent cases where individual deposition rates are held constant and individuals may or may not differ in their use of space. The results from these scenarios where $N = 25$ and $S = 25$ are consistent with the urn simulations: coverages are above 95% and estimates are, on average, about 20% above the true N (Table 3). In compositions G5 through G8, individuals differ in their DNA deposition rates but not their use of space. *Capwire* performs well here with estimates within 10% of N and coverages very near 95%. Compositions G9 to G12 present more complex scenarios where individuals differ both in their use of space and their deposition rates.

Table 3 Results of grid simulated data based on $N = 25$ and $S = 25$ analysed using *capwire*. See Table 2 for descriptions of column headers

Comp Dist*	\tilde{N}	Rel bias	CI width	Coverage of 95% CIs	MRE
G1	29.7	0.19	30	0.97 (0.02, 0.01)	0.26
G2	30.1	0.20	34	0.97 (0.02, 0.01)	0.28
G3	31.4	0.25	34	0.96 (0, 0.04)	0.33
G4	30.7	0.23	35	0.98 (0, 0.02)	0.32
G5	24.1	-0.04	23	0.93 (0.07, 0)	0.20
G6	27.3	0.09	28	0.97 (0.02, 0.01)	0.24
G7	25.2	0.01	24	0.95 (0.05, 0)	0.20
G8	26.0	0.04	25	0.94 (0.05, 0.01)	0.24
G9	24.5	-0.02	23	0.91 (0.09, 0)	0.21
G10	25.2	0.01	25	0.93 (0.07, 0)	0.24
G11	25.2	0.01	24	0.99 (0.01, 0)	0.19
G12	25.6	0.03	25	0.95 (0.05, 0)	0.19

*Composition distribution. See Table 1.

The results are similar to the uneven deposition simulations G5 to G8 with estimates of N being approximately unbiased and coverages in the 90th percentile.

To explore the performance of *capwire* across a range of N and S and to compare to other methods, we limited further simulations to compositions G1, G4, G10 and G12. Several important trends emerge from the results (Table 4). For composition G1, estimates of N are biased somewhat high (10–35%), with bias declining as sample size increases. Coverage is generally near 95%. Performance on composition G4 is similar. For both G1 and G4, coverage begins to

Table 4 Analysis of grid simulated data sets for a range of population and sample sizes using *capwire*. See Tables 2 and 3 for descriptions of column headers

Comp Dist	<i>N</i>	<i>S</i>	\bar{N}	Rel bias	CI width	Coverage of 95% CIs	MRE
G1	16	15	19.9	0.24	25	0.96 (0.03, 0.01)	0.40
		25	18.7	0.17	17	0.99 (0, 0.01)	0.23
		35	17.3	0.08	10	1.00 (0, 0)	0.13
	25	15	33.8	0.35	74	0.93 (0.04, 0.03)	0.51
		25	29.7	0.19	30	0.97 (0.02, 0.01)	0.26
		50	27.9	0.11	17	0.97 (0.01, 0.02)	0.15
	49	25	60.1	0.23	92	0.97 (0, 0.03)	0.34
		50	58.3	0.19	46	0.97 (0, 0.03)	0.22
		100	55.3	0.13	25	0.83 (0, 0.17)	0.13
	100	50	118.9	0.19	105	0.97 (0.01, 0.02)	0.24
		100	123.5	0.24	67	0.93 (0, 0.07)	0.24
		200	114.5	0.15	38	0.40 (0, 0.6)	0.15
	196	100	234.4	0.20	144	0.98 (0.01, 0.01)	0.21
		200	240.9	0.23	91	0.67 (0, 0.33)	0.23
		400	223.4	0.14	52	0.07 (0, 0.93)	0.14
G4	16	15	19.3	0.21	28	0.96 (0.04, 0)	0.37
		25	18.1	0.13	16	0.99 (0.01, 0)	0.25
		35	17.7	0.11	11	0.97 (0.02, 0.01)	0.16
	25	15	33.3	0.33	59	0.98 (0.02, 0)	0.47
		25	30.7	0.23	35	0.98 (0, 0.02)	0.32
		50	27.7	0.11	17	0.97 (0.01, 0.02)	0.14
	49	25	57.6	0.18	73	0.99 (0, 0.01)	0.31
		50	58.5	0.19	44	0.97 (0, 0.03)	0.21
		100	55.4	0.13	25	0.81 (0, 0.19)	0.14
	100	50	121.5	0.21	106	0.97 (0.01, 0.02)	0.28
		100	125.4	0.25	68	0.88 (0, 0.12)	0.26
		200	114.9	0.15	38	0.36 (0, 0.64)	0.15
	196	100	223.1	0.14	140	1.00 (0, 0)	0.17
		200	241.6	0.23	90	0.67 (0, 0.33)	0.23
		400	223.4	0.14	51	0.02 (0, 0.98)	0.14
G10	16	15	16.0	0	18	0.83 (0.17, 0)	0.33
		25	16.1	0.01	14	0.95 (0.05, 0)	0.22
		35	16.2	0.01	10	0.89 (0.11, 0)	0.17
	25	100	25.2	0.01	31	0.87 (0.13, 0)	0.39
		25	25.2	0.01	25	0.93 (0.07, 0)	0.24
		50	24.2	-0.03	14	0.77 (0.23, 0)	0.19
	49	25	44.6	-0.09	53	0.90 (0.1, 0)	0.25
		50	52.1	0.06	39	0.98 (0.02, 0)	0.17
		100	51.5	0.05	26	0.96 (0.02, 0.02)	0.11
	100	50	92.6	-0.07	75	0.84 (0.16, 0)	0.18
		100	103.2	0.03	56	0.97 (0.03, 0)	0.11
		200	106.2	0.06	38	0.95 (0, 0.05)	0.08
	196	100	171	-0.13	93	0.64 (0.36, 0)	0.16
		200	201.1	0.03	79	0.96 (0.04, 0)	0.08
		400	208.6	0.06	52	0.86 (0, 0.14)	0.07
G12	16	15	16.1	0	18	0.92 (0.08, 0)	0.28
		25	17.1	0.07	16	0.95 (0.05, 0)	0.23
		35	16.0	0	10	0.94 (0.06, 0)	0.15
	25	15	25.0	0	42	0.93 (0.06, 0.01)	0.30
		25	25.6	0.03	25	0.95 (0.05, 0)	0.19
		50	24.6	-0.01	14	0.94 (0.06, 0)	0.13
	49	25	47.0	-0.04	55	0.89 (0.1, 0.01)	0.26
		50	51.1	0.04	37	0.98 (0.01, 0.01)	0.19
		100	50.8	0.04	24	1.00 (0, 0)	0.08
	100	50	104.9	0.05	89	0.95 (0.04, 0.01)	0.19
		100	108.1	0.08	58	1.00 (0, 0)	0.12
		200	104.2	0.04	35	0.93 (0.01, 0.06)	0.07
	196	100	191.9	-0.02	110	0.91 (0.09, 0)	0.11
		200	211	0.08	85	0.98 (0.01, 0.01)	0.10
		400	209.1	0.07	48	0.81 (0, 0.19)	0.07

fall well below 95% in larger populations sampled at moderate to high intensity ($N = 100, S = 2N$ and $N = 196, S = N$ and $2N$). Among the four compositions, G10 is most problematic for *capwire* in terms of coverage. Estimates are generally unbiased but coverage was sometimes below 90%. For composition G12, bias is generally not greater than 10% and coverage is near the desired 95%. Although bias and coverage do not always improve with increasing sample size, the mean relative error (MRE) virtually always does.

The hyperbolic rarefaction curve ($y = \alpha x / (\beta + x)$; Kohn *et al.* 1999) did not perform well. Summarizing over compositions and sampling intensities, in simulations where $N = 49$ the method yielded estimates that were, on average, 43% above the true N , MRE values of 46% and an average coverage of only 45% (results not shown). The exponential curve of Eggert *et al.* [$y = \alpha(1 - e^{-\beta x})$; 2003], and the two CAPTURE heterogeneity estimators (Chao and jackknife) performed much better.

We compared the performance (bias, coverage, CI width, and MRE) of the latter three methods to *capwire* across a range of N , compositions, and sampling intensity (Fig. 1). The Chao estimator is the least biased of the four, but it also produces the broadest CIs and the largest MRE. Its coverage is generally 90% or better. The jackknife estimator performs poorly at low sampling intensities with large negative bias, low coverage and large MRE. The Eggert rarefaction method shows point estimates with excellent MREs, but its coverage declines severely for $N = 100$. *Capwire* is the most balanced of the four estimators. Its coverage is comparable to the Chao method and its CIs are considerably more narrow at small sample sizes. Its MRE is very good, especially at the smaller N and lower sampling intensities. Simulations at $N = 16$ and 25 showed that *capwire* is increasingly superior to the other estimators in terms of coverage, CI width and MRE for $N < 50$ (data not shown). *Capwire* does have a positive bias when capture rates are even as they are in G1 and G4. In the more realistic scenarios where capture rates vary (G10 and G12), it shows the same small bias that the Chao method does. The performance of *capwire* becomes more sporadic as N grows substantially above 100 and sampling intensity increases. In such cases the Chao estimator is best overall.

Real data

The *capwire* estimator performs very well on the real data sets (Table 5). For the red wolf data, the CIs cover N in all three sessions and there is a dramatic decrease in CI width as sample effort increases. For the European badger data (Frantz *et al.* 2003), *capwire*'s estimate of 29 (20–43) is at the midpoint of the range that N is known to lie within (24–34) and is similar to the published estimates of the Chao and jackknife methods of 26 (22–45) and 26 (22–40), respectively. For the forest elephant data (Eggert *et al.* 2003), the *capwire*

Table 5 Analysis of real mark-recapture data sets using *capwire*. See Tables 2 and 3 for descriptions of column headers

Study type	Study	Species	Population/sample	N	\hat{N}	CI	Coverage	MRE	Sample size
DNA	Adams*	Red wolf	Session 1	18	19	13, 22	1	0.06	48
			Session 2	15	22	12, 31	1	0.47	22
			Session 3	15	13	7, 46	1	0.13	10
Traditional	Frantz <i>et al.</i> (2003)	European badger	3 groups combined	24–34	29	20, 43	1	0†	47
			Only one	unk	214	143, 246	unk	unk	114
	Eggert <i>et al.</i> (2003)	Forest elephant	Multiple obs/night‡	unk	88	84, 99	unk	unk	398
			1 obs/night§	unk	98	86, 110	unk	unk	263
	Banks <i>et al.</i> (2003)	N. hairy-nosed wombat	4-ha area	130	123	87, 148	1	0.05	128
			40-acre pen	135	143	99, 164	1	0.06	142
	Edwards & Eberhardt (1967)	Eastern cottontail rabbit	Only one	82	81	76, 93	1	0.01	232
			1977	23	21	19, 28	1	0.09	53
	Mares <i>et al.</i> (1981)	Eastern chipmunk	1978	28	32	17, 44	1	0.14	29
			Greenwood <i>et al.</i> (1985)	420	462	307, 551	0.95	0.13	194
Carothers (1973)¶	Taxi cabs	Taxi cabs	52 samples	420	462	307, 551	0.95	0.13	194

*Personal communication. †Based on the midpoint (29) of range of possible true N (24–34). ‡Full data set where multiple observations of individual in one night are included. §Data set reduced to allow only one observation per night as required in traditional mark-recapture. ¶All 52 data sets presented in paper were analysed. \hat{N} , MRE, and sample size are means while confidence bounds are median values. Coverage is proportion of 52 where CIs covered 420. Note that not all 52 data sets are independent.

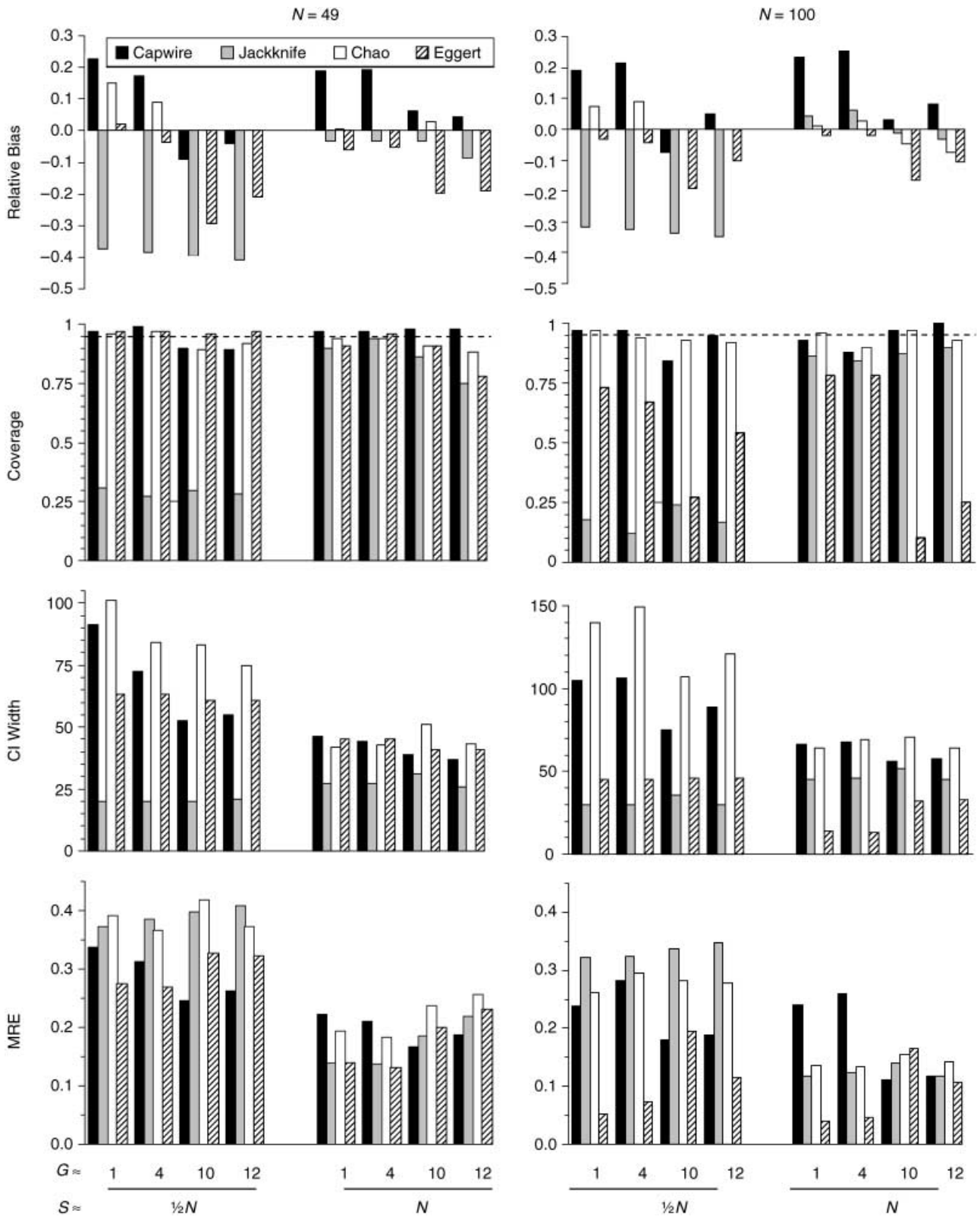


Fig. 1 Relative performance of *capwire* (black bars), jackknife (grey), Chao (white), and Eggert rarefaction (hatched) estimators on grid simulated data. Two major columns correspond to populations sizes labelled at top. Within each N , each subcolumn corresponds to sampling intensity (S) labelled at bottom. Within each subcolumn, each set of four bars corresponds to population composition (G) labelled at bottom above S . In the coverage figures, the dashed line represents the 95% objective.

estimate of 214 (143–246) is similar, though more precise, compared with the published estimate based on the jackknife method of 225 (173–308) and an estimate based on dung counts of 228 (158–337). For the northern hair-nosed wombat, the *capwire* estimate based on all observations of 88 (84–98) is smaller and narrower than the published jackknife estimate of 113 (96–150). When only a single observation is allowed per night, the *capwire* estimate increases to 98 (86–110). Among the five wildlife populations studied with traditional mark–recapture techniques where N was known, estimates are always within 15% of the true N and in three cases within 6%. In all cases the CIs cover the true N . In analysing all 52 data sets published by Carothers (1973) on a taxicab population in Edinburgh, *capwire* overestimated N by an average of 13% and produced CIs covering the true N 50 of 52 times ($\approx 95\%$).

Discussion

Capwire and differing types of populations

The *capwire* model treats the population as an urn in which individuals are continuously mixing. Animals that maintain cohesive groups such as ungulate herds, primate troupes, canid packs, and marine mammal pods may approximate a mixing urn. When deposition rates are even in such cases, it would be better to use ECM compared to TIRM. Whereas TIRM overestimates N , ECM provides unbiased estimates of N , produces narrower confidence intervals, appropriate coverage and smaller MRE (Table 2). In fact, in the urn simulations with $\alpha = 1$, the small bias observed in the LRT model selection process comes from the 11% of the time the incorrect TIRM is selected (Table 2). If ECM is used exclusively, the mean population estimate is 24.9. If a researcher has good biological reason to believe that deposition rates are approximately even, the LRT does not suggest otherwise, and the urn model is reasonable, it may be preferable to use ECM. On the other hand, if the urn model is reasonable but assuming even deposition rates is not, researchers should impose TIRM. Under TIRM, coverage is good and bias small so long as $\alpha \leq 7$, which should cover most field applications. This problem of model choice illustrates a reoccurring observation in the mark–recapture literature: it is best to know enough about the species and sampling design to have some view of the appropriate model and method of analysis a priori (Otis *et al.* 1978; Dorazio & Royle 2003; Link 2003; Boulanger *et al.* 2004).

More common than a mixing urn-like scenario, individuals will occupy semidiscrete areas. Ursids are a good example of a group that has been extensively studied using genetic methods where individuals occupy home ranges (e.g. Woods *et al.* 1999; Bellemain *et al.* 2005). Grid simulations were run to determine how robust *capwire*'s urn model is to

spatial segregation. In the event that DNA deposition rates are approximately even, *capwire* tends to overestimate N . The results from grid simulations with N and $S = 25$ (Table 3) change very little with varying proportions of sedentary and transient individuals (G1–G4) and they are very similar to the analysis of urn data with $\alpha = 1$ (i.e. relative bias $\approx 20\%$, coverage $\approx 97\%$, MRE $\approx 30\%$; Table 2). This suggests the spatial dimension of real data is having a minimal effect on *capwire*.

Most often researchers will deal with populations where individuals neither mix frequently nor deposit DNA at the same rate. For example, among the seven real DNA data sets analysed (Table 5), ECM was rejected for all seven using the LRT at the 0.05 level (results not shown). The grid simulations indicate that *capwire* generally performs better with capture heterogeneity in the data (G5–G12) than without it (G1–G4; Tables 3 and 4 and Fig. 1). For a given N and sample size, the presence of capture heterogeneity reduces bias, narrows CIs, and lowers MRE. Coverage, however, becomes problematic for *capwire* as N and sample intensity increase. Without capture heterogeneity, the drop in coverage to low levels (consistently $< 80\%$) appears somewhere in the $N = 100$ – 200 range (Table 4). With heterogeneity present, the decline is pushed out to greater N s. Grid simulations at $N = 400$ indicate that the decline is between 200 and 400 (data not shown).

Sample size considerations

To this point we have presented sampling intensity as a proportion of N (e.g. $S = 1/2N$). Because N is not known, this measure of sampling intensity is not particularly useful for study design. One possibility is to use an upper bound for what N might reasonably be in setting a sample size goal for a study. An alternative view of sampling effort is to quantify it as the average number of observations per sampled individual (obs/ind). Across the range of sampling intensity studied here, there is a clear decrease in MRE as the number of obs/ind increases (Fig. 2a). In small populations ($N \leq 25$), an average of *c.* 2.5 obs/ind are necessary to obtain estimates which are *c.* 15% from N and close to 3.0 obs/ind are needed to be *c.* 10% from N . For N in the range of 49–100, the news is better: 2.0 obs/ind will provide average estimates *c.* 15% off of N and 2.5 obs/ind, *c.* 10% from N . Precision shows a similar improvement with an increasing number of obs/ind (Fig. 2b). For smaller populations, 2.5 obs/ind will tend to yield interval widths of $3/4N$ and increasing to near 3.0 obs/ind will tend to reduce interval width to $1/2N$. Precision increases substantially as N gets larger. With $N = 100$, 2.0 and 2.5 obs/ind will tend to produce CIs of width $1/2$ and $1/3N$, respectively. These figures may provide some guidance to researchers in designing or adjusting studies such that the desired level of accuracy can be achieved.

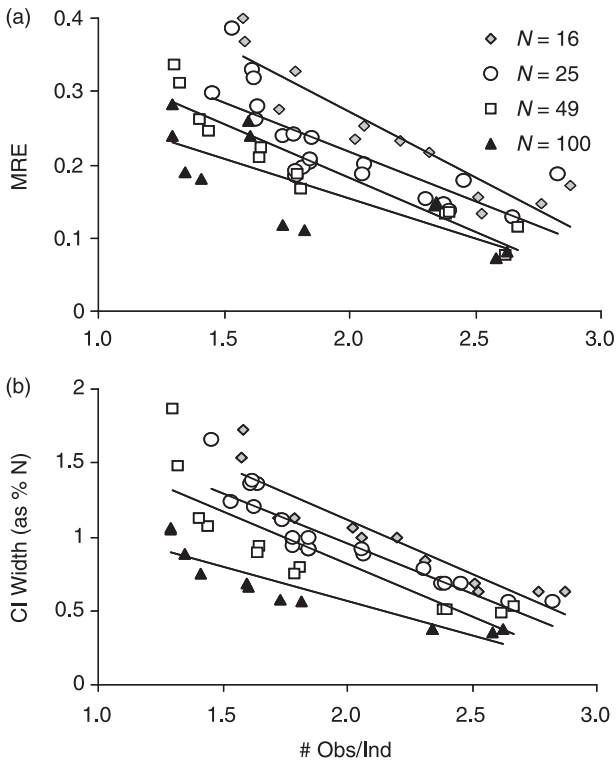


Fig. 2 Effect of number of observations/sampled individual on mean relative error (a) and precision (b) in grid simulated populations of differing size (indicated in legend). Data from populations of different compositions (G1, G4, G10 and G12) are combined.

Comparison to other methods

Rarefaction curve-fitting has been used in a number of DNA-based studies to estimate population size (e.g. Kohn *et al.* 1999; Eggert *et al.* 2003; Wilson *et al.* 2003; Bellemain *et al.* 2005). Rarefaction has an intuitive appeal in that it assumes replacement sampling and has no temporal dimension. In grid simulations the hyperbolic curve ($y = \alpha x / (\beta + x)$; Kohn *et al.* 1999) overestimates population size nearly 50% on average, shows coverage of less than 50%, and large MREs. The exponential curve of Eggert *et al.*, $y = \alpha(1 - e^{-\beta x})$ (2003), however, performs comparably well (Fig. 1). It produces point estimates with low MREs. While biased low when capture rates are heterogeneous (G10 and G12), it is approximately unbiased when deposition rates are even (G1 and G4). For moderate populations (e.g. $N = 25-49$), coverage is generally near 90% or better, but in small and large populations (i.e. $16 \leq N \leq 100$), it shows much poorer coverage (data for small N not shown). These results suggest it might be fruitful to explore alternate ways of calculating CIs for this rarefaction method.

Two of the most commonly used methods for short-duration, mark-recapture studies have been the Chao and

jackknife estimators within program CAPTURE (e.g. Kohn *et al.* 1999; Banks *et al.* 2003; Eggert *et al.* 2003; Frantz *et al.* 2003). Like *capwire*, these estimators assume a closed population and allow capture heterogeneity. We were interested in whether they can be used to analyse DNA-based mark-recapture studies that lack capture sessions. The grid simulations indicate that the Chao estimator compares well with *capwire* (Fig. 1). It generally shows low bias and good coverage, though its CIs and MREs tend to be larger. The jackknife estimator generally suffers from high bias and poor coverage when sampling intensity is low. Even at higher sampling intensities, its coverage drops into the 20–80% range in larger populations ($N = 196$ and 400 ; data not shown).

Which estimator is best depends on the situation and the priorities of the researcher. The strengths of *capwire* are that it displays good coverage, relatively narrow CIs, and it handles capture heterogeneity well. In simulations with heterogeneity (G10 and G12), it is consistently as good or better than all other methods in all four measures of performance (Fig. 1). Its drawbacks include a tendency to overestimate N when DNA deposition rates are approximately even and a drop in coverage in larger populations sampled at high intensity. Alternatively, if an accurate point estimate is of paramount concern and a researcher has good reason to believe that deposition rates are approximately even, then the exponential rarefaction method of Eggert *et al.* (2003) is a good option. Using *capwire* and imposing ECM will also provide good estimates in such cases. Unfortunately, coverage for the Eggert method is quite poor unless N is moderate. The strength of the Chao estimator is that it is robust across a broad range of conditions. It shows low bias, coverage is generally above 90%, and although it lacks precision (wide CIs) and accuracy (large MRE) at low sample intensity, it displays the desirable property of providing consistently better estimates as sample size increases. In small populations, it is generally outperformed by other methods (particularly *capwire*), but as populations get larger it is increasingly superior. This was confirmed in simulations on $N = 196$ and 400 (data not shown).

One potential criticism of this analysis is that both the Chao and jackknife estimators have been used in a way that violates a basic model assumption. They assume that the data come from multiple sessions with no more than one capture per individual per session. In fact, the simulated data have no sessions. Recall that we arbitrarily created the number of sessions for these estimators by defining it as the largest number of times any individual was captured. This is not a trivial concern because these methods estimate N by estimating the per session capture probability, which depends directly on the number of capture sessions. This begs the question: would these methods perform better if sampling really were conducted in temporally distinct sessions with multiple within-session observations reduced

Table 6 Comparison of Chao estimator performance on grid simulated data sets with and without sampling divided into temporal capture sessions. For each of the four data sets (G1, G4, G10 and G12), analysis was conducted in two ways: no. of sampling sessions = 4 (and multiple within-session observations of an individual removed) and no. sample sessions = maximum no. of captures of any individual (with the four real sessions ignored and no observations removed). For all simulations $N = 49$ and $S = 100$

Comp Dist	No. of samp sessions	\hat{N}	Rel bias	CI width	Coverage	MRE
G1	4	52.2	0.07	24	0.86 (0.14, 0)	0.06
	max no. of captures	50.3	0.03	25	0.91 (0.08, 0.01)	0.03
G4	4	53.0	0.08	26	0.82 (0.17, 0.01)	0.11
	max no. of captures	51.1	0.02	24	0.89 (0.10, 0.01)	0.10
G10	4	52.4	0.07	32	0.91 (0.00, 0)	0.15
	max no. of captures	50.3	0.03	31	0.95 (0.03, 0.02)	0.13
G12	4	49.5	0.01	27	0.95 (0.05, 0)	0.12
	max no. of captures	48.8	0.00	30	0.94 (0.04, 0.02)	0.12

to a single observation? When possible, should DNA-based population estimation studies be designed this way? We conducted a cursory exploration of this issue by running a small set of simulations ($N = 49$, $S = 100$, compositions G1, G4, G10 and G12) where sampling was divided into four sessions. Hence in the 4000 simulated steps, sampling was conducted along transects as before every 1000 time-steps. The data were then analysed with the Chao estimator in two ways: (i) in the appropriate manner with the number of sampling sessions equal to four and multiple within-session observations of an individual reduced to one and (ii) as it has been done in this study with the session data ignored, no observations removed, and the number of sessions arbitrarily defined as the maximum number of captures.

In this case the results are clear: bias, MRE, coverage and CI width are all good and usually better when the data are pooled into a single sample and not divided into sessions (Table 6). While this result needs to be tested across a broader range of conditions, it suggests that researchers are throwing away valuable information for estimating N by condensing multiple captures into one. In fact, the ability to capture an individual multiple times without expending greatly more sampling effort is one of the potential strengths of DNA-based mark-recapture. Designing sampling so that it is intense but of short duration is exactly what is needed to minimize violation of the closed-population assumption. Of course, it is important that multiple observations are not pseudoreplicates from the same time and the same place. For example, a bear investigating a sent lure is likely to leave several hair snares at the same trap on the same occasion (entering and leaving). Clearly these multiple observations should be treated as a single observation and are quite distinct from capturing the individual at two separate locations. *Capwire* was designed to make use of multiple observations. This type of DNA-based data can also be accommodated by the traditional Chao estimator.

Capwire and traditional mark-recapture data

We were additionally interested in whether *capwire* can be used to analyse traditional mark-recapture data sets. As an exploratory step we analysed five wildlife data sets from the literature where N was independently known (Table 5). The performance of *capwire* is remarkably good. In all cases the true N is contained within the CI and in four of five cases the MRE is less than 10%. The method also performs well on a population of taxicabs. These data sets have been analysed and reanalysed many times in the literature using different estimators (e.g. Otis *et al.* 1978; Chao 1988; Tardella 2002). Considered across data sets, *capwire's* performance is as good as or better than any of the other estimators. For example, an analysis of the five wildlife data sets in Table 6 using the Chao estimator give very similar bias, MRE values, but substantially wider CIs in four of five cases (results not shown).

Why should *capwire* perform well on data that violate its sample with replacement model assumption? For *capwire* to perform well, the number of observations of each individual must be proportional to the fraction of the population that individual represents. Even when data are collected as binary events over multiple sessions this should still tend to occur, especially as the number of sample sessions increases. This also suggests that *capwire* may be robust to time heterogeneity in capturability. In the session-based methods, one complexity is that capturability may vary between sessions due to factors such as weather and season (Otis *et al.* 1978). Parameters accounting for these changes in capturability must be estimated. But so long as every individual is affected in the same way (i.e. there are no individual by time interactions), then temporal changes will not distort the proportional representation of individuals in the sample. For this same reason, the lack of temporal information in many DNA-based mark-recapture studies (e.g. single sweep scat collection studies) should

not pose a problem. This indication that *capwire* can be effectively used on traditional mark–recapture data sets is promising, but needs to be substantiated by further study.

Improving capwire

Many traditional mark–recapture models view an animal's capture probability as being drawn from a distribution. Examples include point mixture (or latent class) models with two or more types (e.g. Norris & Pollock 1996; Pledger 2000), beta-binomial mixture models (e.g. Dorazio & Royle 2003), logistical and log-linear models (e.g. Coull & Agresti 1999). It may be possible to improve the performance of *capwire* by considering a wider range of underlying capture distributions. *Capwire* might also be improved by accounting for uncertainty in class membership by employing, for example, the EM algorithm (Coull & Agresti 1999). It also seems reasonable to adjust the urn model so that samples are not independent of one another. In reality, if two samples come from nearby geographical locations, they will tend to come from the same individual more often than randomly drawn samples. Samples from disparate locations will tend to be the same individual less often than random. This geographical component of the data is currently ignored, but it seems that such information could be utilized by adding a spatial dimension to the urn model.

Conclusions

The ability to sample with replacement in DNA-based mark–recapture studies yields a different type of data than that obtained in traditional trap-based studies. Removing multiple observations from disparate locations within sessions wastes valuable information. In other DNA-based studies, there are no clearly defined sessions. The proposed *capwire* method is suited to these types of raw data sets. The simulation study conducted here suggests that *capwire* does a particularly good job when dealing with smaller populations ($N \leq 100$) and substantial capture heterogeneity. An analysis of a number of real genetic mark–recapture data sets demonstrates that capture heterogeneity will be commonly encountered.

Acknowledgements

We thank Jen Adams, Andrea Taylor and Sam Banks for generously providing raw data for analysis. Funding was provided by National Science Foundation Experimental Program to Stimulate Competitive Research Grant 9720634 and National Science Foundation Grants 0080935 and 9871024. Paul Joyce's research is partially sponsored by the Initiative in Bioinformatics and Evolutionary Studies (IBEST) at the University of Idaho; funding was provided by NSF EPSCoR, EPS-0132626 and NIH NCRR grant NIH NCRR-1P20RR016448-01, and NSF DEB-0089756.

References

- Banks SC, Hoyle SD, Horsup A, Sunnucks P, Taylor AC (2003) Demographic monitoring of an entire species (the northern hairy-nosed wombat, *Lasiorhinus krefftii*) by genetic analysis of non-invasively collected material. *Animal Conservation*, **6**, 101–107.
- Bellemain E, Swenson JE, Tallmon D, Brunberg S, Taberlet P (2005) Estimating population size of elusive animals with DNA from hunter-collected feces: four methods for brown bears. *Conservation Biology*, **19**, 150–161.
- Boulanger J, McLellan BN, Woods JG, Proctor MF, Strobeck C (2004) Sampling design and bias in DNA-based capture–mark–recapture population and density estimates of grizzly bears. *Journal of Wildlife Management*, **68**, 457–469.
- Burnham KP, Overton WS (1979) Robust estimation of population size when capture probabilities vary among animals. *Ecology*, **60**, 927–936.
- Carothers AD (1973) Capture–recapture methods applied to a population with known parameters. *Journal of Animal Ecology*, **42**, 125–146.
- Chao A (1988) Estimating animal abundance with capture frequency data. *Journal of Wildlife Management*, **52**, 295–300.
- Comps B, Gömöry D, Letouzey J, Thiébaud B, Petit RJ (2001) Diverging trends between heterozygosity and allelic richness during postglacial colonization in the European beech. *Genetics*, **157**, 389–397.
- Coull BA, Agresti A (1999) The use of mixed logit models to reflect heterogeneity in capture–recapture studies. *Biometrics*, **55**, 294–301.
- Creel S, Spong G, Sands JL *et al.* (2003) Population size estimation in Yellowstone wolves with error-prone noninvasive microsatellite genotypes. *Molecular Ecology*, **12**, 2003–2009.
- Dorazio RM, Royle JA (2003) Mixture models for estimating the size of a closed population when capture rates vary among individuals. *Biometrics*, **59**, 351–364.
- Edwards WR, Eberhardt L (1967) Estimating cottontail abundance from live-trapping data. *Journal of Wildlife Management*, **31**, 87–96.
- Eggert LS, Eggert JA, Woodruff DS (2003) Estimating population sizes for elusive animals: the forest elephants of Kakum National Park, Ghana. *Molecular Ecology*, **12**, 1389–1402.
- Flagstad Ø, Hedmark E, Landa A *et al.* (2004) Colonization history and noninvasive monitoring of a reestablished wolverine population. *Conservation Biology*, **18**, 676–688.
- Frantz AC, Pope LC, Carpenter PJ *et al.* (2003) Reliable microsatellite genotyping of the Eurasian badger (*Meles meles*) using faecal DNA. *Molecular Ecology*, **12**, 649–1661.
- Greenwood RJ, Sargeant AB, Johnson DH (1985) Evaluation of mark–recapture for estimating striped skunk abundance. *Journal of Wildlife Management*, **49**, 332–340.
- Kohn MH, York EC, Kamradt DA, Haught G, Sauvajot RM, Wayne RK (1999) Estimating population size by genotyping faeces. *Proceedings of the Royal Society of London. Series B, Biological Sciences*, **266**, 657–663.
- Link WA (2003) Nonidentifiability of population size from capture–recapture data with heterogeneous detection probabilities. *Biometrics*, **59**, 1123–1130.
- Mares MA, Streilein KE, Willig MR (1981) Experimental assessment of several population estimation techniques on an introduced population of eastern chipmunks. *Journal of Mammalogy*, **62**, 315–328.

- McKelvey KS, Schwartz MK (2004) Genetic errors associated with population estimation using non-invasive molecular tagging: problems and new solutions. *Journal of Wildlife Management*, **68**, 439–448.
- Miller CR, Joyce P, Waits LP (2002) Assessing allelic dropout and genotype reliability using maximum likelihood. *Genetics*, **160**, 357–366.
- Mowat G, Paetkau D (2002) Estimating marten *Martes americanus* population size using hair capture and genetic tagging. *Wildlife Biology*, **8**, 201–208.
- Mowat G, Strobeck C (2000) Estimating population size of grizzly bears using hair capture, DNA profiling, and mark-recapture analysis. *Journal of Wildlife Management*, **64**, 183–193.
- Norris III JL, Pollock KH (1996) Nonparametric MLE under two closed capture-recapture models with heterogeneity. *Biometrics*, **52**, 639–646.
- Otis DL, Burnham KP, White GC, Anderson DR (1978) statistical inference from capture data on closed animal populations. *Wildlife Monographs*, **62**.
- Palsbøll PJ, Allen J, Bérubé M *et al.* (1997) Genetic tagging of humpback whales. *Nature*, **388**, 767–769.
- Pledger S (2000) Unified maximum likelihood estimation for closed capture-recapture models using mixtures. *Biometrics*, **56**, 434–442.
- Taberlet P, Waits LP, Luikart G (1999) Noninvasive genetic sampling: look before you leap. *Trends in Ecology & Evolution*, **14**, 323–327.
- Tardella L (2002) A new Bayesian method for nonparametric capture-recapture models in presence of heterogeneity. *Biometrika*, **89**, 807–817.
- Wilson GJ, Frantz AC, Pope LC *et al.* (2003) Estimation of badger abundance using faecal DNA typing. *Journal of Applied Ecology*, **40**, 658–666.
- Woods JG, Paetkau D, Lewis D, McLellan BN, Proctor M, Strobeck C (1999) Genetic tagging of free-ranging black and brown bears. *Wildlife Society Bulletin*, **27**, 616–627.

Craig Miller is a post-doc at University of Idaho working on statistical problems in conservation genetics. Paul Joyce is a Professor of Mathematics, Statistics and Bioinformatics at the University of Idaho. His interdisciplinary work involves mathematical modeling and statistical theory in population genetics, experimental evolution and Systematic Biology. Lisette Waits is the co-director of the Laboratory for Ecological and Conservation Genetics and the Center for Research on invasive species and small populations. Her research program focuses on the conservation genetics of a variety of wildlife species.

Appendix I

Capwire algorithm used to estimate population size

1. *Initialize α* : Calculate mean no. of captures/sampled individual. Calculate mean no. of captures for individuals captured more than this average (average above) and less than this average (average below). Define initial $\hat{\alpha} = (\text{average above})/(\text{average below})$.
 2. *Find expected capture counts*: Assuming $N = T$, $N_A = N/2$, $N_B = N/2$, and $\alpha = \hat{\alpha}$, calculate the expected number of captures an individual of each type: $E(c_A) = S[\alpha/(\alpha N_A + N_B)]$ and $E(c_B) = S[1/(\alpha N_A + N_B)]$.
 3. *Assign capture classes*: Assign each sampled individual a capture class based on the following rules. If it is observed one time, assign it to capture class B. Otherwise calculate the absolute difference between the observed no. of captures and that expected for each class: $|c_i - E(c_A)|$ and $|c_i - E(c_B)|$. Assign the individual to whichever absolute difference is smaller. This defines T_A and T_B as well.
 4. *MLE estimation of N*: Given this vector of capture class assignments, T_A , T_B , and α , find the MLE of N_A and N_B . Do this by initially calculating likelihood (equation 2) assuming $N_A = T_A$, $N_B = T_B$. Then begin incrementing N_B up by one and calculating the likelihood. Continue until the likelihood begins to decline; the largest likelihood defines the MLE of N_B . MLE of N_A is T_A . (N_A is not incremented because adding the easier to capture type A individuals yields a smaller likelihood than adding type B individuals).
 5. *Bias adjust α* using equation 3 to obtain $\hat{\alpha}_{\text{adjusted}}$.
 6. *Repeat MLE estimation*: Repeat step 4 using $\hat{\alpha}_{\text{adjusted}}$ to update estimates of N_A and N_B .
 7. *Repeat capture class assignment*: Repeat steps 2 and 3 based on $\hat{\alpha}_{\text{adjusted}}$, and update estimates of T_A and T_B .
 8. *Check for convergence*: If any of the capture class assignments in the sample changed in step 7, return to step 4. If capture assignments do not change, go to step 9.
 9. *Estimate N*: $\hat{N} = \hat{N}_A + \hat{N}_B$.
 10. *Bootstrap to obtain confidence intervals*: Generate many data sets by drawing with replacement from an urn in which there are \hat{N}_A type A balls, \hat{N}_B type B balls, and the balls have weights of 1 and α , respectively. For each conduct parameter estimation (steps 1–9). Let α_{ts} be the test size specified by the user. Define the lower and upper confidence bounds as the $\alpha_{\text{ts}}/2$ and $1 - \alpha_{\text{ts}}/2$ quantiles of the bootstrap estimates.
-

Appendix II

Grid simulations were conducted using the algorithm given in section A. The types of individuals composing populations are listed and parameterized in section B

A. Grid Simulation algorithm. Note that individuals are independent; their movements and depositions have no effect on each other.

- 1 Specify size of square grid and population size, N .
- 2 Subdivide grid into N non-overlapping squares and place an individual in centre of each square. Designate this as an individual's home region centre (HRC).
- 3 Randomly assign each individual a type according to specified composition distribution (Table 1). This defines movement type (sedentary or transient) and deposition rate (seldom, moderate, or often) for each individual as given in Appendix IIB.
- 4 Determine deposition schedule for each individual. For each individual draw number of steps until 1st depositing a scat from an exponential distribution with rate λ_{deposit} (Appendix IIB). Determine number of additional steps until 2nd scat by drawing again. Continue until sum of steps exceeds 4000 (the duration of the simulation). While simulating movement of each individual (steps 5–10), deposit scat at that grid location according to this deposition schedule.
- 5 For each individual chose one of the Cartesian directions to move with equal probabilities.
- 6 For each individual, determine number of steps to move before potentially changing direction by drawing from an exponential distribution with rate λ_{turn} (Appendix IIB).
- 7 For each individual, move this number of steps. Deposit scats as specified by deposition schedule (step 4).
- 8 For each individual, determine new direction to move. With specified probability $P(H)$, the individual moves in the direction that takes it most directly toward its HRC and in one of other three directions with probability $1 - P(H)/3$. Hence $P(H)$ describes fidelity to home with $P(H) > 0.25$ causing sedentary behaviour and $P(H) < 0.25$ causing transient behaviour. When two directions tie for directness to home, they are each assigned probability $P(H)/2$ and the other two directions are assigned probability $(1 - P(H)/2)/2$. When individuals encountered boundaries, the same rules were employed except that there are only three (at sides) or two (at corners) potential directions to choose from.
- 9 For each individual repeat steps 6, 7, and 8 until total number of steps is 4000.

B. Parameterization of six types of individuals in grid simulations and resulting deposition count and home region size. Based on simulation of 4000 step duration. Deposition rate, draw to home, and turning rate defined in Appendix IIA

Ind Type code	Type description	Deposition rate (λ_{deposit})	Draw to home = $P(H)$	Turning rate (λ_{deposit})	Exp. no. of scats on grid	90% home region*
SOD	sedentary, often depositor	1/40	0.35	5	100	66 (± 21)
SMD	sedentary, moderate depositor	1/80	0.35	5	50	66 (± 21)
SSD	sedentary, seldom depositor	1/160	0.35	5	25	66 (± 21)
TOD	transient, often depositor	1/40	0.15	5	100	300 (± 96)
TMD	transient, moderate depositor	1/80	0.15	5	50	300 (± 96)
TSD	transient, seldom depositor	1/160	0.15	5	25	300 (± 96)

*Distance to home region centre average individual spent 90% of time within (\pm SD) based on 100 simulations on grid of size 500×500 steps.