

ESTIMATING THE ABUNDANCE OF MOUSE POPULATIONS OF KNOWN SIZE: PROMISES AND PITFALLS OF NEW METHODS

PAUL B. CONN,^{1,4} ANTHONY D. ARTHUR,² LARISSA L. BAILEY,³ AND GRANT R. SINGLETON²

¹Department of Fishery and Wildlife Biology, Colorado State University, Waggar Hall, Fort Collins, Colorado 80523 USA

²CSIRO Sustainable Ecosystems, GPO Box 284, Canberra ACT 2601 Australia

³USGS Cooperative Fish and Wildlife Research Unit, North Carolina State University and Patuxent Wildlife Research Center, 12100 Beech Forest Road, Laurel, Maryland 20708-4017 USA

Abstract. Knowledge of animal abundance is fundamental to many ecological studies. Frequently, researchers cannot determine true abundance, and so must estimate it using a method such as mark–recapture or distance sampling. Recent advances in abundance estimation allow one to model heterogeneity with individual covariates or mixture distributions and to derive multimodel abundance estimators that explicitly address uncertainty about which model parameterization best represents truth. Further, it is possible to borrow information on detection probability across several populations when data are sparse. While promising, these methods have not been evaluated using mark–recapture data from populations of known abundance, and thus far have largely been overlooked by ecologists. In this paper, we explored the utility of newly developed mark–recapture methods for estimating the abundance of 12 captive populations of wild house mice (*Mus musculus*). We found that mark–recapture methods employing individual covariates yielded satisfactory abundance estimates for most populations. In contrast, model sets with heterogeneity formulations consisting solely of mixture distributions did not perform well for several of the populations. We show through simulation that a higher number of trapping occasions would have been necessary to achieve good estimator performance in this case. Finally, we show that simultaneous analysis of data from low abundance populations can yield viable abundance estimates.

Key words: abundance estimation; Huggins-Alho model; MARK; mark–recapture; model averaging; *Mus musculus*; Pledger model.

INTRODUCTION

Knowledge of abundance is fundamental to the study of ecology. For instance, ecologists may use time series of abundance to parameterize predator–prey models, to examine species–habitat relationships, or to evaluate the effects of experimental treatments on populations. In an applied context, wildlife managers often use abundance estimates for determining harvest quotas, for deciding management options for pest populations, and for formulating conservation plans. Reliable methods for estimating abundance are thus critical for the success of a diverse array of ecological applications, and should include both design-based and model-based considerations.

One common approach for estimating abundance is mark–recapture. In these studies, capture data from uniquely marked individuals over two or more trapping occasions provides the information necessary to estimate a probability of detection, p , and an abundance parameter, N . Often, ecologists are confronted with the additional challenge of diagnosing and modeling heterogeneity, behavioral response, and time effects on

the probability of capture (Otis et al. 1978). If not properly addressed, these sources of variation can seriously compromise abundance estimates. Since the late 1970s, many biologists have used program CAPTURE (Otis et al. 1978, White et al. 1982, Rexstad and Burnham 1991) to help in making inferences about detection probability and to calculate abundance estimates (White 2005).

A number of studies have examined the performance of individual closed capture models, such as those available in CAPTURE, either through simulation (e.g., Otis et al. 1978, Menkens and Anderson 1988, McKelvey and Pearson 2001, Conn et al. 2004) or by comparing mark–recapture abundance estimates to “known” population sizes (e.g., Edwards and Eberhardt 1967, Carothers 1973, Manning et al. 1995, Davis et al. 2003, Parmenter et al. 2003). An additional focus in several of these studies was to evaluate the reliability of indexes of abundance such as number of unique individuals captured (M_{t+1}). Conn et al. (2004) and White (2005) showed that unanticipated variation in detection probability can cause biases in inferences made from indexes, particularly when detection probabilities vary systematically with treatment, site condition, or any other feature being compared. In order to use indexes, one must make the assumption that detection probability does not vary in a systematic way. However,

Manuscript received 14 December 2004; revised 19 July 2005; accepted 3 August 2005. Corresponding Editor R. A. Relyea.

⁴E-mail: pconn@cnr.colostate.edu.

the data typically gathered for indexes are often insufficient to address variation in detection probabilities, and thus cannot reliably assess the performance of the index.

Comparisons of index and mark–recapture estimator performance have often been made with respect to a limited number of individual estimators, and have not taken into account recent developments in mark–recapture methodology that allow researchers to perform model selection and averaging, and to combine data to achieve more efficient estimators. These developments allow models where capture probability can be modeled across multiple groups as a function of individual covariates (Huggins 1989, 1991, Alho 1990) or mixture distributions (Pledger 2000), and for a modeling framework that explicitly acknowledges uncertainty about which mark–recapture estimator best represents truth (Stanley and Burnham 1998, Burnham and Anderson 2002). Ecologists can readily implement these developments in program MARK (White and Burnham 1999), which allows users to formulate a diverse array of potential estimators with design matrices (cf. Lebreton et al. 1992).

While these methodological advances are compelling, ecologists have tended to avoid MARK for abundance estimation, in part because of its recent development, and in part because of the lack of published literature on estimator performance (G. C. White, *personal communication*). In this paper, we analyzed trapping data from 12 populations of wild house mice (*Mus musculus*) of known abundance using program MARK. Davis et al. (2003) employed similar data to estimate population size for a number of default estimators, but did not examine model-averaged estimators or approaches that model capture probability as a function of individual covariates. We provide a short description of field trapping methods, and review the modeling framework and the models we used for analysis. In describing our analysis, we explore several common nuances that ecologists are likely to encounter when modeling abundance with covariates, with multiple sparse data sets, and when trap saturation occurs. Finally, we present estimates of abundance, compare these to true population sizes, and discuss the reliability of each modeling approach.

METHODS

Field trapping

Laboratory-reared house mice were released into six 15×15 m enclosures during each of two fertility control experiments (see Davis et al. 2003 for details). Food and water were provided in each enclosure, and initial population sizes of ~ 20 individuals were allowed to grow. Trapping was conducted during four primary trapping sessions over an 18-wk period for the first experiment, and for seven primary trapping sessions over a 34-wk period for the second experiment. During each primary session, trapping was conducted for four

consecutive nights on a 6×6 trapping grid, consisting of 36 baited Longworth live-capture traps (Longworth Scientific, Abingdon, UK) per enclosure. Researchers tagged new individuals with individually numbered Hauptner brass ear tags (Sieper, Sydney, Australia) and recorded sex, weight, length, and reproductive status. After each experiment, true abundance was measured by completely removing animals from each enclosure, by live trapping animals until none were caught on two consecutive nights, and then by snap trapping for an additional week to ensure that none remained in the enclosure. The total time for removal varied from 8 to 23 days, with the vast majority of individuals caught within the first week of trapping. Most founders were still alive after the 18- and 34-wk experiments, indicating a high weekly survival rate. Thus we anticipated a minimal negative bias in our true abundance values. Very small mice (< 55 mm) generally did not appear in traps until complete removal; thus, we needed to make some assumptions about which mice were part of the catchable population during the final trapping session. We established a threshold cutoff value for length of 55 mm for unmarked mice, but recognized that growth of mice between the final trapping session and removal might result in recruitment of mice into the catchable population. To address this problem, we fitted a growth model to data from both experiments (Fig. 1), and applied an estimated growth rate of 0.75 mm/d for 55-mm mice to back calculate which mice would have had lengths > 55 mm during the final trapping session. Since some mortality did occur throughout the course of the study, we only compiled capture histories for the first and last primary trapping session, to conform to the assumption of population closure and to permit comparison between mark–recapture estimators and true abundances. These adjustments caused our data to differ from that of Davis et al. (2003), who only considered mice with body length > 70 mm.

Analyses

We performed three analyses on each final trapping session data set using program MARK. First, we conducted an analysis that simultaneously employed mixture models (Pledger 2000) and individual covariates (length, weight, sex, reproductive status) to model heterogeneity in the data, using a Huggins-Alho type model (Huggins 1989, 1991, Alho 1990). We refer to these as “Pledger-Huggins-Alho” analyses throughout the paper. Second, we conducted an analysis using the full likelihood models of Pledger (2000). In this approach, heterogeneity is modeled only in terms of mixture distributions. We refer to these as “Pledger” analyses. Finally, we performed a simultaneous analysis of data from several enclosures, allowing multiple data sets to inform the estimation of capture probability (MacKenzie et al. 2005, White 2005). We also performed this type of “simultaneous” analysis on capture histories compiled from the first trapping session in order to examine

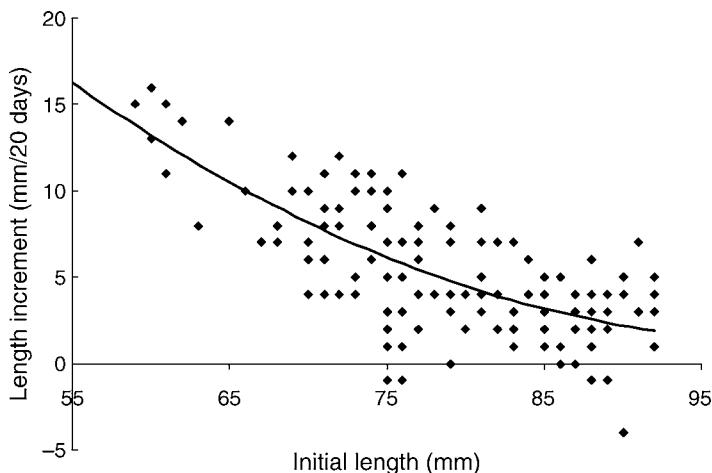


FIG. 1. Estimated length-dependent growth rate over a 20-day period for captive *Mus musculus*. The average daily growth rate for 55-mm mice is ~0.75 mm/d. Negative values reflect measurement error.

estimator performance in a low abundance setting. In each case, we describe models with $\Delta AIC_c < 2.0$ as “top models,” in the sense that these models are highly supported by the data (Burnham and Anderson 2002).

Pledger-Huggins-Alho analysis.—Under this approach, we modeled capture probabilities for trapping occasion i and animal j , p_{ij} , as a function of different combinations of individual covariates, mixture parameters, and time and behavioral effects. For instance, as one possibility we considered an additive model for detection probability that included a behavioral effect and a heterogeneity formulation involving both a length covariate and a mixture parameter (i.e., assuming two mixture groups). We assumed that the capture probability of the first mixture could be expressed as

$$\text{logit}(p_{ij}) = \beta_0 + \beta_1 I_{ij} + \beta_2 l_j$$

with

$$\text{logit}(p_{ij}) = \beta_0 + \beta_1 I_{ij} + \beta_2 l_j + \beta_3$$

for the second mixture. Here, $I_{ij} = 1$ if animal j has been caught prior to time i and 0 otherwise, l_j represents length for animal j , and β_3 designates a mixture effect. Symbolically, we could express this model as $(b + h + l)$ as in Lebreton et al. (1992), to denote behavior (b), length (l), and extra heterogeneity induced by a mixture distribution (h). Once we formulated and fitted individual models for detection probability, we compute derived abundance estimators for each model as $\hat{N} = \sum_j 1/p_j^*$, where p_j^* is defined as the probability that individual j is caught at least once during the session:

$$p_j^* = 1 - \prod_i (1 - p_{ij}).$$

For this analysis, we fitted a total of 57 a priori models to each house mouse data set, allowing the presence or

absence of different factors that we thought might influence detection probability. Time, trapping response, and several types of heterogeneity were assumed to be additive effects when present. We allowed models to include effects for weight or length, but not both at the same time. We included fully interactive models for weight and sex when reproductive status (pregnancy) was not also included because we anticipated pregnancy would complicate the relationship between weight and capture probability. We only allowed mixture parameters on models with less than three other effects so as to limit the chance of over-parameterized models.

After we fitted all models to the data, we examined the standard errors of estimates as well as parameter counts from a singular value decomposition in program MARK (Cooch and White 2005) to determine which models were over-parameterized, and removed these models from analysis. As Pollock (2002) noted, even when all parameters are estimable, the Huggins-Alho model may produce unstable estimates when capture probabilities for some animals are estimated to be close to zero. This problem occurred in several of our data sets, and we dealt with it in two ways. First, we reran all of our models in a post hoc analysis using discrete categories for weight and length, as suggested by Converse (2005). We used values of length < 72 mm to designate juveniles, and > 72 mm to designate adults (Singleton 1989). These assignments were made independent of sex. Second, we deleted all models with estimated standard errors on abundance > 600 . We selected this value post hoc, because several top models had standard errors in the 200–500 range, and we felt uncomfortable removing these models from analysis.

Next, we conducted model averaging on the derived abundance parameters (Burnham and Anderson 2004) in program MARK. Model averaging produces a final abundance estimate conditional on the results from all fitted models. Abundance estimates from highly parsimonious

monious models (in terms of low ΔAIC_c) contribute heavily to the final abundance estimate. The standard error of the model-averaged estimate is a function of the standard error from each model as well as the degree of congruence between model-specific abundance estimates. Thus if several highly supported models yield abundances that differ greatly, the standard error of the model-averaged estimate will be high to reflect uncertainty about which model best represents truth (Burnham and Anderson 2002).

Pledger full likelihood analysis.—In order to contrast estimator performance from models with individual covariates to those without, we conducted an analysis of final trapping session data using the full likelihood approach of Pledger (2000). Since this approach includes abundance as a parameter in the likelihood function, it is more efficient than Huggins-Alho estimators if covariates are unavailable (White 2002). Pledger's models incorporate heterogeneity in the form of a mixture distribution on the probability of capture, but individual covariates cannot be included. We considered a total of eight additive models, which combined hypotheses of behavioral response to trapping, a time effect, and heterogeneity in the form of two mixtures. Once we had fitted these models to the data, we removed those that were over-parameterized, as well as those with estimated standard errors on abundance >600 . We employed model averaging (Burnham and Anderson 2002) to derive final abundance estimates.

Poor estimator performance from the Pledger approach in several of the pens made us question whether four trapping occasions per trapping session were sufficient to adequately model heterogeneity using mixtures. We thus conducted a small post hoc simulation study to address this question. As a simple but reasonable biological scenario, we assumed that capture probability was solely a function of length (Fig. 2B), where

$$\text{logit}(p_j) = -1.22 + 0.66 \times \text{length}_j. \quad (1)$$

This relationship was estimated by the top model in the Pledger-Huggins-Alho analysis of Experiment 2, Pen 3. We then fitted a normal distribution to length data from all individuals captured in Pen 3 at the end of Experiment 2 (232 individuals). Next, we used the simulation capability in MARK to simulate capture histories for the 232 individuals, assuming that lengths were normally distributed and that captures were independent Bernoulli trials with success probability determined according to Eq. 1. In total we generated 1000 data sets for each level of 4, 6, 8, or 10 occasions per trapping session. After the eight models in the full-likelihood model set were fitted to these simulated data sets in MARK, we imported abundance estimates and associated statistics into SAS/IML (SAS 2004) to conduct model averaging. We applied the same estimability and standard error criterion as in the other analyses, and estimated percent relative bias, standard

error, and mean squared error (MSE) for each number of trapping occasions.

Simultaneous analysis.—We next conducted a MARK analysis that simultaneously estimated capture probabilities across all pens in Experiment 1. By combining likelihoods for all pens, we anticipated a reduction in the estimability issues typically associated with sparse data (White 2005). We used all 57 models considered in the Pledger-Huggins-Alho model set, but also included an extra set of 57 models that included an additive pen effect on the probability of capture. Each model was fitted simultaneously to the capture data from Experiment 1, allowing us to select models that were parsimonious across all pens. Since we considered covariates in this analysis, we used the Pledger-Huggins-Alho abundance estimator. An analysis of data from the first trapping session allowed us to examine the utility of applying this approach to low abundance populations, where abundance may be too small to successfully apply mark-recapture methods to single populations. In contrast, simultaneous analysis of the final trapping session permitted comparison to single pen analyses. In this case, we expected an increase in precision if there were similar capture dynamics between pens. For analysis of data from the first trapping session, we deleted all models with standard errors >40 (twice the true population size), as well as those with a larger number of estimated parameters than expected. In practice, we would not know true population size, and so the standard error cut off would instead be based on empirical data. For the last trapping session, we applied previously defined estimability and standard error criterion for inclusion of models in the final model set. We did not attempt to simultaneously analyze data from Experiment 2 because of time constraints.

RESULTS

Pledger-Huggins-Alho analysis

Estimated abundance from single pen Pledger-Huggins-Alho analyses generally tracked true abundance, although higher population sizes were consistently underestimated (Table 1). Continuous covariate models sometimes estimated capture probabilities close to zero for individuals with small length or weight covariates, or for one group of a mixture distribution (Fig. 2A). In these cases, abundance estimators were unstable, with unrealistically high abundance values and large standard errors. Removing models with estimated standard errors >600 was not always sufficient to prevent model-averaged estimates with high standard errors or estimates differing substantially from true abundances. Although average percent relative bias was close to zero, a nominal 95% confidence interval procedure actually produced coverage of 83% (10 out of 12 pens). With the exception of Pen 3 in Experiment 1, length or weight always appeared in top models, sometimes accompanied by extra heterogeneity from a mixture distribution (see

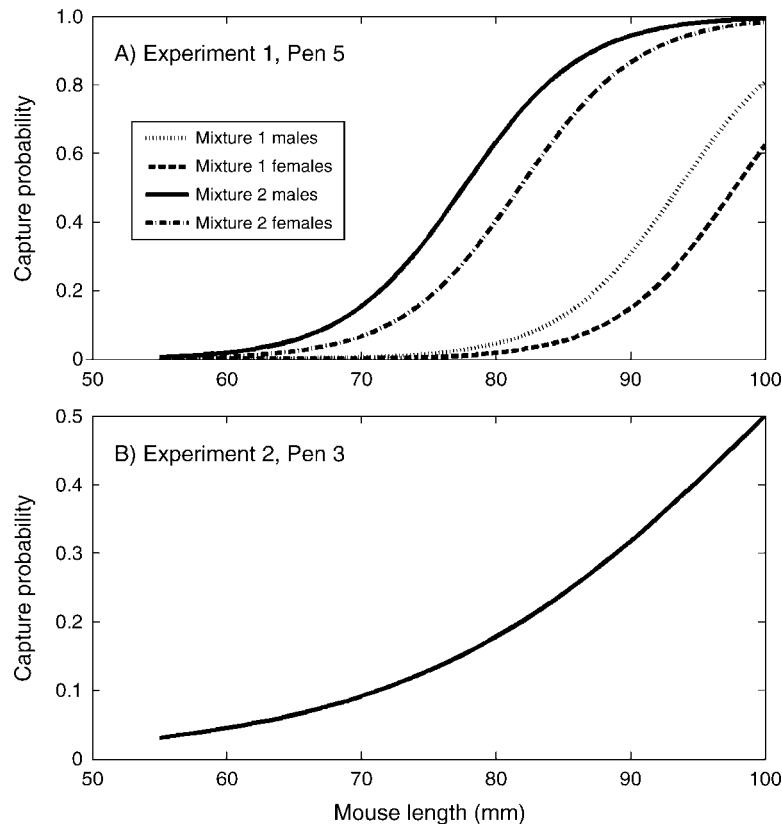


FIG. 2. (A) The estimated relationship between capture probability and individual covariates as obtained from the top-ranked model from the Pledger-Huggins-Alho analysis of Experiment 1, Pen 5. According to this relationship, individuals with length <60 mm have capture probabilities approaching zero, which leads to a large standard error on the abundance estimator. The top model included a two-point finite mixture (i.e., a high and a low group) to describe individual heterogeneity in capture probability. (B) The estimated relationship between length and capture probability, as determined by the top-ranked model in the Pledger-Huggins-Alho analysis of Experiment 2, Pen 3. In this case, individuals with length <60 mm have small detection probabilities but are theoretically still detectable. Thus the estimated standard error for abundance is lower.

Appendix). Sex and behavioral effects were often present in top models as well, as was pregnancy for several of the pens. The ability to detect an effect of pregnancy was likely related to the sample size of pregnant females in the marked sample, which varied from pen to pen.

When we reran the Pledger-Huggins-Alho analysis with a two-category length covariate, fewer models produced standard errors >600 . In addition, fewer models with large standard errors contributed to model-averaged abundance estimates. Although absolute percent relative bias was larger, the standard errors of estimates were smaller, resulting in a smaller overall mean squared error (Table 1). Abundance estimator confidence interval coverage for the 12 pens was 75%, with a negative percent relative bias. The binary length covariate often appeared in top models, with models including mixture-based heterogeneity also receiving high support (see Appendix). We detected similar sex, behavior, and pregnancy effects as with the continuous analysis.

Pledger full likelihood analysis

In the Pledger full likelihood analysis, few models had large enough standard errors (>600) to meet our protocol for deletion. However, a number of pens had models with a standard error of abundance >200 when the capture probability for one portion of a heterogeneity mixture was estimated close to zero, contributing to large standard errors on final abundance estimators (Table 1). Confidence interval coverage for the 12 pens was 75%, with a negative percent relative bias. Heterogeneity was the most consistently present effect in top models, indicating strong support for heterogeneous detection probabilities (see Appendix).

For several of the pens, abundance estimates differed substantially from true population sizes (Table 1). Our simulation results indicated that the Pledger model set may perform poorly with four trapping occasions. Estimated expected standard error on abundance for four trapping occasions was approximately one-third of absolute abundance (Table 2). Thus it would not be abnormal to find a point estimate for abundance that

TABLE 1. Comparison of true abundance (N) with model-averaged estimates of abundance (\hat{N}) from four analyses.

Experiment	Pen	N	M_{t+1}	Continuous H-A		Discrete H-A		Simultaneous		Pledger	
				\hat{N}	$\widehat{SE}(\hat{N})$	\hat{N}	$\widehat{SE}(\hat{N})$	\hat{N}	$\widehat{SE}(\hat{N})$	\hat{N}	$\widehat{SE}(\hat{N})$
1	1	69	38	46	8	46	8	53	8	50	26
1	2	244	92	159	34	198	43	368	130	216†	276
1	3	108	46	124†	43	126	43	122	38	72	16
1	4	89	49	88†	26	74†	19	86	15	68	20
1	5	136	58	267†‡	257	131†	33	144	39	230‡	367
1	6	195	80	159	58	139	49	254	82	125	59
2	1	287	91	191	59	217	106	202‡	155
2	2	233	88	166†	52	197	62	214‡	227
2	3	291	95	219†‡	145	167	49	145	27
2	4	179	74	143	90	119	19	160‡	238
2	5	226	98	208	65	203	55	195	72
2	6	399	107	294†‡	193	254†‡	191	166†	32
$\widehat{\% \text{ bias}} \text{ (SE)}$			-59 (8)	-9 (37)		-21 (16)		12 (26)		-19 (32)	
MSE			21 612	17 329		9837		...		37 995	
Coverage			...	10/12		9/12		6/6		9/12	
r^2			0.935	0.745		0.913		0.994		0.565	

Notes: The analyses are Pledger-Huggins-Alho with continuous covariates (Continuous H-A), Pledger-Huggins-Alho with a binary length covariate (Discrete H-A), Pledger-Huggins-Alho with data jointly analyzed from multiple enclosures (Simultaneous), and Pledger models (Pledger). Also reported is the number of unique individuals captured (M_{t+1}), average percent relative bias ($\widehat{\% \text{ bias}}$), mean square error (MSE), 95% confidence interval coverage (Coverage), and Spearman rank correlation coefficient (r^2). We did not attempt to simultaneously analyze data from Experiment 2.

† One or more models had $\widehat{SE}(\hat{N}) > 600$, and were thus deleted from analysis.

‡ Models with SE > 200 accumulated between 0.10 and 0.92 of the model weight, contributing to the model-averaged estimate.

was half of its true value. Standard errors decreased substantially when increasing the number of trapping occasions (Table 2). Estimated percent relative bias for each estimator was also slightly negative (Table 2), but approached zero with increasing trapping occasions if only models with heterogeneity were considered.

Simultaneous analysis

When we fitted Pledger-Huggins-Alho models simultaneously to data from the first trapping session in Experiment 1, the model-averaged abundance estimator had a slightly lower percent relative bias, and had a slightly larger mean squared error (MSE) than the count statistic M_{t+1} (Table 3). However, none of the top models had a pen effect on detection probability (Appendix), providing evidence that detection probabilities did not vary systematically between pens (insofar

TABLE 2. Estimated percentage relative bias ($\widehat{\% \text{ bias}}$), expected standard error ($E[\widehat{SE}(\hat{N})]$), and mean square error (MSE) associated with model-averaged abundance estimators for different numbers of trapping occasions.

Trapping occasions	$\widehat{\% \text{ bias}}$	$E[\widehat{SE}(\hat{N})]$	MSE
4	-0.079	79.01	17 274
6	-0.074	35.98	4096
8	-0.077	13.68	736
10	-0.070	6.45	311

Notes: In all cases, encounter history files were simulated by assuming that capture probability was a function of length, and abundance was estimated by performing model averaging on a model set that included Pledger (2000) mixture models for heterogeneity. We analyzed 1000 simulated data sets for each number of trapping occasions.

as we were able to detect with small sample sizes). In contrast, when we analyzed data from the last trapping session in Experiment 1, the top model without a pen effect had $\Delta AIC_c = 24.6$, providing strong evidence that capture probabilities differed between pens (Burnham and Anderson 2002). Thus, M_{t+1} could be ruled out as a viable abundance index for making absolute comparisons of population size between pens for these data (Skalski and Robson 1992). Detection probabilities for

TABLE 3. Comparison of true abundance (N) with model-averaged estimates of abundance (\hat{N}) from the Pledger-Huggins-Alho analysis of Experiment 1 first trapping session data.

Pen	N	M_{t+1}	\hat{N}	$\widehat{SE}(\hat{N})$
1	20	14	17	5
2	20	13	16	5
3	20	14	16	4
4	20†	16	19	5
5	20†	13	16	7
6	20	14	17	7
$\widehat{\% \text{ bias}} \text{ (SE)}$			-31 (50)	-18 (58)
MSE			39	41
Coverage			...	6/6

Notes: Data were jointly analyzed across multiple pens. Also reported is minimum number known alive (M_{t+1}), average percent relative bias ($\widehat{\% \text{ bias}}$), mean square error (MSE), and 95% confidence interval coverage (Coverage). Spearman rank correlation coefficient (r^2) could not be computed due to uniformity in N . We did not attempt to jointly analyze data from Experiment 2.

† True abundance for Pens 4 and 5 could not be determined definitively, but was between 18 and 20. Thus we did not use these pens in $\widehat{\% \text{ bias}}$ or MSE calculations.

the first trapping session averaged to 0.56 when covariate effects were not considered, while detection probabilities for the last trapping session ranged from 0.29 to 0.63 (mean 0.39) depending on the pen, and were negatively correlated with true abundance ($r^2 = -0.84$). The top model fitted to the data from the last trapping session indicated length, behavior, and pen effects on capture probability, with extra heterogeneity induced by a mixture distribution. This model acquired 92% of the model weight, thus contributing heavily to the model-averaged abundance estimate (Appendix). Estimates of several detection probabilities were near zero, producing abundance estimates substantially larger than for other models, with standard errors on abundance two to four times as large (Table 1).

DISCUSSION

This study represents the first attempt at validating several new types of mark-recapture abundance estimators with real data. We considered model averaging within a likelihood framework and the use of covariates and mixture distributions to help model heterogeneity in detection probabilities. All MARK analyses yielded abundance estimators with percent relative bias statistically indistinguishable from zero at the $\alpha = 0.10$ level, albeit with low precision. Approaches that accounted for heterogeneity through individual covariates (i.e., with Pledger-Huggins-Alho models) were particularly successful, and estimation issues with small detection probabilities were reduced by employing a binary classification for length data that was based on an age-length relationship.

In a similar data set, Davis et al. (2003) was able to detect heterogeneity using a discriminant function in program CAPTURE (Rexstad and Burnham 1991). However, CAPTURE does not provide a unified framework for multimodel selection and inference, as MARK does through use of ΔAIC_c . If a researcher detects heterogeneity using CAPTURE, they must then decide which of a number of heterogeneity estimators to use. Davis et al. (2003) reported satisfactory performance when using one of Chao's (1988) sample coverage estimators to estimate abundance under heterogeneity. However, it is unclear how their results can be extrapolated to other research studies. A key problem is that it is difficult to compare the performance of different moment-based estimators such as the estimators attributed to Chao (1988) when a specific study is not replicated. In contrast, comparing the relative parsimony of likelihood-based estimators is straightforward.

Model sets that solely used mixture distributions to characterize heterogeneity produced unacceptable estimates for several of the pens. This is likely a function of the number of trapping occasions used in the study. Our simulation results indicated that six or more trapping occasions may be necessary to adequately model

heterogeneity using mixture distributions in a full likelihood approach, at least with the range of capture probabilities explored here. Link (2003) argued against the use of mixture models for population abundance, showing that abundance is a nonidentifiable parameter when there is individual heterogeneity in capture probabilities. Nevertheless, our simulation results indicated that mixture models may perform reasonably well when there are a large number of encounter occasions (e.g., six or more). Indeed, intensive sampling programs should often be employed when heterogeneity is suspected no matter which estimation strategy is employed. In addition to the increased power for detecting differences between individual capture probabilities, increasing sample effort is essential for adequate performance of abundance estimators incorporating heterogeneity (Chao 1989, Boulanger et al. 2004).

Simultaneous analysis of data from multiple populations may lead to abundance estimators with good performance, provided that sources of variation in capture probability are similar between populations. This approach works by combining information on detection probability across several populations, which is sometimes necessary if populations are small or data sparse. This approach worked well for multiple low abundance populations ($N = 20$), but was problematic for some populations with higher densities. One potential reason for this discrepancy is some sort of density dependence in capture probabilities, such that individuals in populations of different densities respond differently to trapping. However, differences in size, sex, or age structure between high density pens could have produced bias as well. This effect might be eliminated by increasing model complexity, specifically with regard to interaction terms.

While modeling approaches using MARK often produced acceptable estimates, we sometimes had to employ relatively arbitrary judgments about which models to retain for model averaging purposes in order to obtain coherent abundance estimates. These judgments were necessary because certain models estimated capture probabilities close to zero for a segment of the population in some pens, and abundance estimation is notoriously unstable under these circumstances (Pollock 2002). Inspection of the data indicated that substantial trap saturation occurred over the course of these experiments, which may have been partially responsible for these problems. Supporting this contention, capture probabilities were negatively correlated with population density ($r^2 = -0.84$), and the bias of our abundance estimators was often larger for high abundance pens. Four nights of trapping coupled with 36 traps per pen permitted a maximum of 144 captures per pen in each experiment. This relatively low number of traps per individual likely exacerbated the heterogeneity in probability of capture, resulting in a population segment with capture probabilities close to zero (e.g., Fig. 2). The

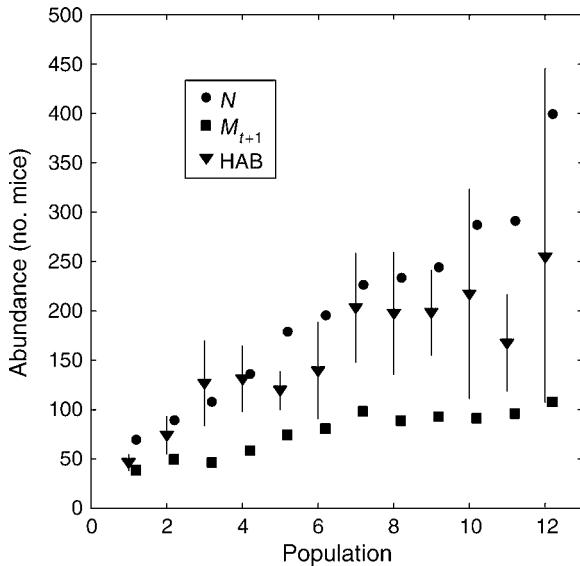


FIG. 3. True abundance (N), M_{t+1} , and the binary length Pledger-Huggins-Alho abundance estimator (HAB) when mouse populations during final trapping sessions are arranged in order of abundance. Vertical lines represent \pm SE for the HAB estimator.

captive mouse populations used in these experiments obtained higher densities than normally would be expected in wild populations. Thus, we anticipate that the approaches considered in this article will be even more robust in ecologically relevant situations, particularly when populations do not fluctuate drastically over the course of a study. In cases of trap saturation or when behavioral responses, such as territoriality, limit the availability of traps (cf. Davis et al. 2003), inferences should be tempered and exploratory data analysis should be performed to determine which portions of the population are actually observable.

Despite mark-recapture estimators performing sub-optimally in some cases, we found a larger potential for erroneous inference when one ignores capture probabilities and looks only at M_{t+1} (Fig. 3). Despite having a higher r^2 than the estimation methods employed on data from this experiment, one cannot properly make inferences about the magnitude of abundance differences between pens by looking at M_{t+1} alone. In the last trapping session, true abundance differs substantially between pens, but M_{t+1} does not provide a clear picture of these differences because it cannot account for the negative relationship between detection probability and abundance. In contrast, the Pledger-Huggins-Alho (HAB) estimators tended to track true abundances. The unusual degree of trap saturation found in this experiment may have limited the performance of these estimators at higher population sizes due to decreases in capture probability and increases in heterogeneity, thus reducing the “catchable” portion of the population (White et al. 1982). Nevertheless, the HAB estimator

exhibited better performance in terms of mean squared error (MSE). For our low abundance analysis, M_{t+1} performed better, with a lower MSE than the simultaneous mark-recapture analysis. This may often be the case when M_{t+1} approaches true abundance. The appropriateness of using M_{t+1} as an index for the populations in question was reflected in Δ AIC_c values of models containing pen effects relative to those not containing pen effects. In general, we might expect that if simultaneous analysis models with low Δ AIC_c frequently include population effects, then M_{t+1} could be dismissed as a potentially biased index. Conversely, M_{t+1} may be an appropriate tool for inference if models with low Δ AIC_c do not include population effects. In either case, we suggest that ecologists use mark-recapture models to explore the validity of assumptions required for M_{t+1} as a first step in any analysis.

In some situations, scientists and managers are more interested in density than absolute abundances, especially because there is often ambiguity regarding the sampled area (the “effective trapping area”). Furthermore, the spatial arrangement of traps may actually induce heterogeneous capture probabilities based on the proximity of an animal to traps. Simply stated, an individual is more likely to be caught during a study if one or more traps are located well within its home range vs. traps located near the edge of its home range (e.g., Boulanger et al. 2004). A method that accommodates this trap-induced spatial heterogeneity has recently been developed by Efford and colleagues (Pledger and Efford 1998, Efford 2004). While spatial heterogeneity did not seem to be an important consideration for our data because of the small spatial scale of the enclosed populations, this approach seems promising for populations in the wild, and thus is worthy of more investigation.

We conclude by emphasizing that heterogeneity is often the most important factor impeding reliable abundance estimation. Researchers should thus pay due attention to design-based sampling protocols that help to reduce the degree of heterogeneity in capture probabilities (cf. Otis et al. 1978, White et al. 1982, Boulanger et al. 2004). We recommend that the investigator (1) place multiple traps at each point on the trapping grid to help reduce trap saturation in high density populations, (2) record individual covariates such as length, sex, and possibly weight to help model heterogeneity, (3) space traps so that an animal is likely to encounter more than one trap during the course of the study, and (4) use as many trapping occasions as possible without violating the population closure assumption, especially when heterogeneity is anticipated. The last two suggestions require some knowledge of the target population. For house mice in Australian wheat fields, intensive studies of home range size indicate that they have home ranges of \sim 0.04 ha (20 \times 20 m) during the breeding season and they roam widely during the on-breeding season (Krebs et al. 1995,

Chambers et al. 2000). These findings suggest 10 m spacing between traps would adequately meet the third recommendation. Similarly, literature on daily survival can be used to guide selection of an appropriate number of trapping occasions.

ACKNOWLEDGMENTS

We acknowledge Lisa Chambers for access to the enclosure data. Discussions with Gary White, Ken Burnham, Paul Doherty, Sarah Converse, and Murray Efford greatly improved the content of this manuscript. Funding for the first author was provided by the National Science Foundation under Grant DGE-0221595003, administered by the PRIMES program at Colorado State University.

LITERATURE CITED

- Alho, J. M. 1990. Logistic regression in capture–recapture models. *Biometrics* **46**:623–635.
- Boulanger, J., G. Stenhouse, and R. Munro. 2004. Sources of heterogeneity bias when DNA mark–recapture sampling methods are applied to grizzly bear (*Ursus arctos*) populations. *Journal of Mammalogy* **85**:618–624.
- Burnham, K. P., and D. R. Anderson. 2002. Model selection and multimodel inference: a practical information–theoretic approach. Second edition. Springer-Verlag, New York, New York, USA.
- Burnham, K. P., and D. R. Anderson. 2004. Model selection: understanding AIC and multimodel inference, with contrasts to BIC. *Sociological Methods and Research* **33**:261–304.
- Carothers, A. D. 1973. Capture–recapture methods applied to a population with known parameters. *Journal of Animal Ecology* **42**:125–146.
- Chambers, L. K., G. R. Singleton, and C. J. Krebs. 2000. Movements and social organisation of wild house mice (*Mus domesticus*) in the mallee wheatlands of Victoria, Australia. *Journal of Mammalogy* **81**:59–69.
- Chao, A. 1988. Estimating animal abundance with capture frequency data. *Journal of Wildlife Management* **52**:295–300.
- Chao, A. 1989. Estimating population size for sparse data in capture–recapture experiments. *Biometrics* **45**:427–438.
- Corn, P. B., L. L. Bailey, and J. R. Sauer. 2004. Indexes as surrogates to abundance for low abundance species. Pages 59–74 in W. L. Thompson, editor. *Sampling rare and elusive species: concepts, designs, and techniques for estimating population parameters*. Island Press, Washington, D.C., USA.
- Converse, S. J. 2005. Small mammal responses to forest restoration and fuel reduction. Dissertation. Colorado State University, Fort Collins, Colorado, USA.
- Cooch, E. G., and G. C. White. 2005. Program MARK: a gentle introduction. (<http://www.phidot.org/software/mark/docs/book/>).
- Davis, S. A., L. K. Akison, L. N. Farroway, G. R. Singleton, and K. E. Leslie. 2003. Abundance estimators and truth: accounting for individual heterogeneity in wild house mice. *Journal of Wildlife Management* **67**:634–645.
- Edwards, W. R., and L. Eberhardt. 1967. Estimating cottontail abundance from livetrapping data. *Journal of Wildlife Management* **31**:87–96.
- Efford, M. G. 2004. Density estimation in live-trapping studies. *Oikos* **106**:598–610.
- Huggins, R. M. 1989. On the statistical analysis of capture–recapture experiments. *Biometrika* **76**:133–140.
- Huggins, R. M. 1991. Some practical aspects of a conditional likelihood approach to capture experiments. *Biometrics* **47**:725–732.
- Krebs, C. J., A. J. Kenney, and G. R. Singleton. 1995. Movements of feral house mice in agricultural landscapes. *Australian Journal of Zoology* **43**:293–302.
- Lebreton, J.-D., K. P. Burnham, J. Clobert, and D. R. Anderson. 1992. Modeling survival and testing biological hypotheses using marked animals: a unified approach with case studies. *Ecological Monographs* **62**:67–118.
- Link, W. A. 2003. Nonidentifiability of population size from capture–recapture data with heterogeneous detection probabilities. *Biometrics* **59**:1123–1130.
- MacKenzie, D. I., J. D. Nichols, N. Sutton, K. Kawanishi, and L. L. Bailey. 2005. Improving inferences in population studies of rare species that are detected imperfectly. *Ecology* **86**:1101–1113.
- Manning, T., W. D. Edge, and J. O. Wolfe. 1995. Evaluating population size estimators: an empirical approach. *Journal of Mammalogy* **76**:1149–1158.
- McKelvey, K. S., and D. E. Pearson. 2001. Population estimation with sparse data: the role of estimators versus indices revisited. *Canadian Journal of Zoology* **79**:1754–1765.
- Menkins, G. E., and S. H. Anderson. 1988. Estimation of small-mammal population size. *Ecology* **69**:1952–1959.
- Otis, D. L., K. P. Burnham, G. C. White, and D. R. Anderson. 1978. Statistical inference from capture data on closed populations. *Wildlife Monographs* **62**:1–135.
- Parmenter, R. R. et al. 2003. Small-mammal density estimation: a field comparison of grid-based vs. web-based density estimators. *Ecological Monographs* **73**:1–26.
- Pledger, S. 2000. Unified maximum likelihood estimates for closed capture–recapture models using mixtures. *Biometrics* **56**:434–442.
- Pledger, S., and M. Efford. 1998. Correction of bias due to heterogeneous capture probability in capture–recapture studies of open populations. *Biometrics* **54**:888–898.
- Pollock, K. H. 2002. The use of auxiliary variables in capture–recapture modeling: an overview. *Journal of Applied Statistics* **29**(1–4):85–102.
- Rexstad, E., and K. P. Burnham. 1991. User's guide for interactive program CAPTURE: abundance estimation of closed animal populations. Colorado Cooperative Fish and Wildlife Research Unit, Colorado State University, Fort Collins, Colorado, USA.
- SAS. 2004. SAS/IML, Interactive Matrix Programming. SAS Institute, Cary, North Carolina, USA.
- Singleton, G. 1989. Population dynamics of an outbreak of house mice (*Mus domesticus*) in the mallee wheatlands of Australia—hypothesis of plague formation. *Journal of Zoology* **219**:495–515.
- Skalski, J. R., and D. S. Robson. 1992. Techniques for wildlife investigations. Academic Press, San Diego, California, USA.
- Stanley, T. R., and K. P. Burnham. 1998. Information-theoretic model selection and model averaging for closed population capture–recapture studies. *Biometrical Journal* **40**:475–494.
- White, G. C. 2002. Discussion comments on: the use of auxiliary variables in capture–recapture modeling. An overview. *Journal of Applied Statistics* **29**:103–106.
- White, G. C. 2005. Correcting wildlife counts using detection probabilities. *Wildlife Research* **32**:211–216.
- White, G. C., D. R. Anderson, K. P. Burnham, and D. L. Otis. 1982. Capture–recapture and removal methods for sampling closed populations. Report LA-8787-NERP, Los Alamos National Laboratory, Los Alamos, New Mexico, USA.
- White, G. C., and K. P. Burnham. 1999. Program MARK: survival estimation from populations of marked animals. *Bird Study* **46**(supplement):120–138.

APPENDIX

A summary of model selection and model averaging procedures for mouse abundance analysis (*Ecological Archives* A016-034-A1).