ACTA THERIOLOGICA
Vol. 29, 15: 187—203, 1984

Empirical Evaluation of Several Population Size Estimates Applied to the Grey Squirrel

Bruce A. C. DON

Don B.A.C., 1984: Empirical evaluation of several population size estimates applied to the grey squirrel. Acta theriol., 29, 15: 187—203 [With 7 Tables]

A trap-mark-recapture study of the grey squirrel in southern England (Shorten & Courtier, 1955) was reanalysed using ten different population estimators. Estimates of population size which were not biased by variation in trappability were derived from shot samples. Population closurė was assumed. Trapping-based estimates were evaluated by comparison with shooting-based estimates for various subpopulations and sampling schemes. Several tests were employed to examine the assumptions underlying the capture-recapture process. The results indicated that assumption tests and consequent model selection procedures may often fail to recommed the best possible estimator. This is considered to be a reflection of the low power of assumption tests and the non-robustness of certain estimators to violation of their assumptions. Robustness is emphasised as a desirable property for population estimators. Of the estimators considered, the Jackknife procedure yielded reasonable population estimates most consistently.

[Animal Ecology Research Group, South Parks Road, Oxford OXI 3PS, U.K.]

1. INTRODUCTION

The size of animal populations is often estimated using mark-recapture methods, and there is now an extensive array of mathematical techniques which attempt to derive unbiased population estimates from such capture data (recently reviewed by Otis, Burnham, White & Anderson, 1978; Cormack, 1979; Seber, 1982). All such estimation techniques depend upon certain assumptions. These assumptions define the conceptual model for the mark-recapture process. Choosing the correct model is of prime importance, since estimates based upon differing assumptions may yield quite disparate results from the same data set (e.g. see Carothers, 1973; Otis *et al.*, 1978). The problem is how to select the best estimator for a given study.

Ideally, the biologist should be able to examine a mark-recapture data set for conformance to a particular model, and, if the tested model is inappropriate, use an alternative model whose assumptions are more compatible with the data. Unfortunately, assumption tests used in mark-recapture studies are typically of low power, i.e., type I errors are often produced (Roff, 1973b; Otis *et al.*, 1978). Also, competing hypotheses concerning which assumptions are appropriate to a particular study may interact, confounding the interpretation of specific tests. Moreover, such tests may suggest a conceptual model for which there is no useful estimation procedure. Some of these problems may be lessened by the use of ad hoc procedures for examining the interactions between assumption tests: This approach is developed by Otis *at al.* (1978).

An alternative, more pragmatic, approach to the problem is to use estimators which perform well when applied to a population of known size. There are three situations in which population size *(N)* or at least a putatively unbiased estimate *(Ñ)* is available; (i) in numerically simulated populations (e.g. Manly, 1970; Roff, 1973a; Otis *et al.*, 1978; Romesburg & Marshall, 1979; Zarnoch, 1979), (ii) captive populations (e.g. Edwards & Eberhardt, 1967; Brady & Pelton, 1976; Mares, Streilein & Willig, 1981; and Carother's 1973 novel study of taxi-cabs) or (iii) by use of two very different sampling techniques such as trap-mark-shoot (e.g. Keith & Meslow, 1968; Nixon, Edwards & Eberhardt, 1967; Edwards & Eberhardt, 1967). Simulation studies are useful for examining the behaviour of various estimators under certain assumption sets, but it follows from the above that we are unlikely to know how those assumptions match the real world. The relatively few vertebrate studies in categories (ii) and (iii) above suffer from rather few recaptures, or they have not been analysed using a comprehensive range of population estimators.

The purpose of this paper is to analyse an extensive set of mark-recapture data, from a grey squirrel *(Sciurus carolinensis)* population, for which an "unbiased" estimate of N is available from trap-mark-shoot data. By considering several models and asumption tests, theoretically "best" (Otis *et al.*, 1978 approach) and empirically "best" (pragmatic approach) estimates can be identified for various sampling schemes and various sub-populations.

2. METHODS

2.1. Fieldwork

The analyses presented in this paper are based upon data collected from a grey squirrel population by Shorten & Courtier (1955). Full details of methodology for the fieldwork are given in Shorten & Courtier (1955: 497—499). Trapping was carried out in 19 ha of relatively isolated deciduous woodland in southern England, May 1954. Twenty multi-catch and eleven single catch traps were set for a total of nine days. Traps were initially prebaited for three days and were checked each morning and each evening when set (eighteen traprounds). After the first four days of trapping (session A), all traps were repositioned and set, following four days of prebaiting, for a further five days (session B). As soon as trapping ceased the area was systematically searched for five days, and all squirrels found were shot. Trapped squirrels were individually toe-clipped or ear-tagged and all shot squirrels were examined for marks. The time between first capture and last death was eighteen days.

This data set is particularly useful for the evaluation of population estimators since: (i) we can assume that the population was more or less closed (*sensu* Otis *et al.*, 1978) because the duration of the sampling programme was fairly short, and not at a time of year when very high immigration/emigration is expected (Don, 1981), (ii) trap positions were changed once, thus making the probability of capture more "random" over space than would be the case with a fixed grid, (iii) large numbers of individuals were marked, with a good recapture rate, and the shooting similarly yielded high numbers with a considerable proportion being marked.

2.2. Analysis

The classification of models and notation used by Otis *et al.* (1978) has been adopted here. Their paper should be consulted for full details of the models, their estimators, estimated variances and assumption tests. The models differ only in their assumptions concerning capture probabilities (p). The null model (M_o) assumes p is equal for all individuals, on all trapping occasions, irrespective of previous capture history. M_h assumes p varies between individuals, irrespective of time or capture history. M_b assumes p varies according to a behavioural response following first capture (i.e. animals become "trap-happy" or "trap-shy" (but there is a constant initial probability of capture which does not vary between trapping occasions. M_t assumes that p is equal for all individuals, irrespective of capture history but varies between trapping occasions. Models M_{tb}, M_{th}, M_{bh} and M_{tbh} are the logical combinations of these basic models. Not all models have an appropriate estimator, whilst for certain models several estimators are available.

Table 1
Summary of each model and estimator used for the present analysis of trap-mark-recapture data.

Model	Causes of variation in p	Estimator	Reference for formulae
M_o	None	Binomial	1
		Poisson	1
		Null	2
M_h	Individual heterogeneity	Geometric	1
		Negative Binomial	1
		Jackknife	2
M_b	Behavioural trap response	Zippin	2
M_{bh}	Individual heterogeneity & behavioural trap response	Removal	2
M_t	Time	Darroch	2
		Schumacher-Eschmeyer	1

References: 1 — Seber (1982), 2 — Otis *et al.* (1978)

The five models for which Otis *et al.* (1978) provide estimators are M_o, M_h, M_b, M_t and M_{bh}. Their estimators will be referred to as Null, Jackknife, Zippin, Darroch and Removal respectively. In order to extend the analysis to other estimators which have been used in the wildlife literature, and for grey squirrels in particular (Flyger, 1959; Nixon *et al.*, 1967; Mosby, 1969; Bouffard & Hein, 1978) I have employed five further methods. These can be classified according to the models above, as follows: zero-truncated Binomial and Poisson (M_o) zero-truncated Geometric and Negative Binomial (M_h), Schumacher-Eschmeyer (M_t). The formulations of these estimators may be found in Seber (1982). Note that no new models are added to those of Otis *et al.* (1978) but the estimators, especially for M_h, may differ greatly in their approach to the solution. The classification of all models and estimators used is summarized in Table 1. Variance estimates for the Schumacher-Eschmeyer method were calculated according to Seber (1982). The iterative procedure of Hartley (1958) was used to calculate variances for the Poisson and Binomial methods. The Geometric and Negative Binomial variance estimates were calculated from methods developed by J. Franklin (pers. comm. — details in prep.).

The model testing procedure of Otis *et al.* (1978) has been used, with additional goodness-of-fit tests pertaining to the Binomial, Poisson, Geometric and Negative Binomial distributions. All methods used assume population closure, and this was tested using the closure statistic of Otis *et al.* (1978).

The "unbiased" population estimates and associated confidence limits are based upon marked: unmarked ratios in the shot sample, employing Chapman's (1951) modified Petersen estimator (Seber, 1982).

3. RESULTS

3.1. Treatment of Data

Trapped and shot squirrels have been classified as adult male, adult female and juvenile (spring-born young), the age separation being based on body weight and pelage (Shorten & Courtier, 1955: p. 507). I have followed the original authors in assuming that squirrels shot in the drey weighing less than 150 g were not old enough to have been captured. (Juveniles rarely leave the nest before eight weeks of age, at a weight of 150—200 g, Shorten 1951). These juveniles along with four animals which died in traps have been excluded from further analysis. Comparison of the remaining juveniles in the shot sample showed no significant difference in mean body weight between marked ($\bar{x}=316.7$ g) and unmarked ($\bar{x}=317.8$ g), thus it is probably not the case that most unmarked juveniles were untrappable because of their small size.

I have stratified the trapping data and their analysis by animal category (adult male, adult female, juvenile) and by time (session A only, session B only, evening traprounds only). When considering session B only or evenings only, captures made in session A or mornings (respectively) were ignored. Hence an animal marked in session A and recaptured in session B would be treated as unmarked on its first session B capture, for analysis of session B only. The stratification adopted with corresponding totals, yields thirteen data sets, of which six (each category (3) by each session (2)) are truly independent. Morning and evening traprounds are considered as separate sampling occasions.

3.2. Trap-mark-shoot Estimates

Table 2 summarizes the information from which the "unbiased" population estimates are derived. It should be noted that the total population estimate is not independent of the subgroup estimates, and this is true for all subsequent analyses. Chapman (1951) estimates *($\hat{N}_{CH}$)* and their associated 95% confidence limits, presented in Table 2, are the values to which all subsequent estimates will be compared. Each population estimate in Table 2 has relatively narrow estimated confidence limits, a reflection of the good sample sizes obtained during both trapping and shooting (although see discussion of coefficients of variation below).

Table 2

„Unbiased" population estimates and associated confidence limits calculated from shooting data.

	n_1	n_2	m_2	$\hat{N}_{CH}$	95% confidence limits of $\hat{N}_{CH}$	$\frac{100 \cdot CI}{\hat{N}_{CH}}$
Adult male	52	49	39	65.3	60.8— 69.8	±6.9
Adult female	66	57	38	98.6	87.2—110.0	±11.6
Juvenile	42	44	26	70.7	60.5— 80.9	±14.4
Total	160	150	103	232.8	218.0—247.6	±6.4

$\hat{N}_{CH}$: Chapman's (1951) population estimator, n_1: number of individuals marked at end of trapping. n_2: number of individuals shot, m_2: number of n_2 bearing marks, CI: 95% confidence interval of $\hat{N}_{CH}$.

The validity of the Chapman estimates rests critically on the assumption that squirrels were shot randomly with respect to whether they were marked or not. We cannot test this directly, but the proportion of squirrels marked was independent of the location at which they were shot (dreys, trapsites or elm trees) — $\chi^2=3.61$, 2 df, NS. This suggests that there was a thorough mixing of marked and unmarked animals prior to the shooting.

3.3. Trapping Data

Table 3 summarizes the number of captures (n_j), number of marked animals in the population (M_j) and number of unmarked captures (u_j) for each trapping occasion (j), for all thirteen data sets. (This is an

Table 4

Frequencies of capture (f_j) for each subgroup and each time period analysed.

Number of captures (j)		1	2	3	4	5	6	7	Σf_j	$\Sigma j \cdot f_j$
Session A+B	Adult male	17	17	10	6	2			52	115
	Adult female	20	24	11	4	4	3		66	155
	Juvenile	22	11	6	0	0	1		42	82
	Total	59	52	27	10	6	4	2	160	352
Session A	Adult male	22	10						32	42
	Adult female	26	16	1					43	61
	Juvenile	14	1	1	1				17	23
	Total	62	27	2	1				92	126
Session B	Adult male	23	14	6	1				44	73
	Adult female	28	16	6	4				54	94
	Juvenile	18	9	6	0	1			34	59
	Total	69	39	18	5	1			132	226
Session A+B (evening only)	Total	78	32	14	3	1			128	201

Table 3

Number of individuals (n), cumulative number of individuals marked (M) and number of unmarked individuals (u) for each trapping occasion (j), presented for each subgroup and time period analysed. Occasions 1, 3,5... were morning traprounds; 2, 4, 6... were evening traprounds.

Occasion	(j)	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
									Session A+B											
Adult male	n	3	7	4	3	5	9	3	8	14	12	1	6	5	7	7	9	4	8	
	M	—	3	10	14	17	20	24	25	32	39	44	44	44	46	47	47	49	49	52
	u	3	7	4	3	3	4	1	7	7	5	0	0	2	1	0	2	0	3	
Adult female	n	9	6	4	10	4	9	12	7	13	11	6	11	10	10	9	5	9	10	
	M	—	9	15	19	26	28	36	42	43	47	52	54	57	59	61	61	62	63	66
	u	9	6	4	7	2	8	6	1	4	5	2	2	2	2	0	1	1	3	
Juvenile	n	1	2	0	5	3	5	3	5	7	7	1	8	4	6	4	10	7	5	
	M	—	1	3	3	6	8	12	13	17	22	26	26	31	32	36	38	40	40	42
	u	1	2	0	3	2	4	1	4	5	4	0	5	1	4	2	2	0	2	
Total	n	13	15	8	18	12	23	17	20	34	30	8	25	19	23	20	24	20	23	
	M	—	13	28	36	49	56	72	80	92	108	122	124	132	137	144	146	151	152	160
	u	13	15	8	13	7	16	8	12	16	14	2	8	5	7	2	5	1	8	
					Session A only										Session B only					
Adult male	n									14	12	1	6	5	7	7	9	4	8	
	M									—	14	26	27	29	31	34	36	40	40	44
	u									14	12	1	2	2	3	2	4	0	4	
Adult female	n									13	11	6	11	10	10	9	5	9	10	
	M									—	13	24	30	36	40	44	46	48	49	54
	u	(As for occasions 1—8, above)								13	11	6	6	4	4	2	2	1	5	
Juvenile	n									7	7	1	8	4	6	4	10	7	5	
	M									—	7	13	14	20	22	26	28	31	32	34
	u									7	6	1	6	2	4	2	3	1	2	
Total	n									34	30	8	25	19	23	20	24	20	23	
	M									—	34	63	71	85	93	104	110	119	121	132
	u									34	29	8	14	8	11	6	9	2	11	
							Session A+B, evening traprounds only													
Total	n		15		18		23		20		30		25		23		24		23	
	M		—		15		30		48		64		81		94		105		114	128
	u		15		1!		18		16		17		13		11		9		14	

extension of Table 4 in Shorten & Courtier, 1955). Note that this information for session A only is simply that for the first eight traprounds of A+B. Table 4 summarizes the frequencies of capture *(f_j)* for each data stratum. Tables 3 and 4 provide the necessary information for calculating all the estimates which follow.

Analysing animal categories or time periods separately potentially provides a means of reducing variation in capture probabilities, at the cost of reducing sample sizes. This will have the effect of reducing one's confidence in an estimate. There is inevitably a "trade-off" between bias and precision in such stratification schemes.

3.4. Assumption Testing

The various assumption tests and distributional goodness-of-fit tests are summarized in Table 5. In all cases, a significant rejection of the null hypothesis is indicated when $p<0.05$ (bold figures in Table 5). The first column contains the results of the closure test, and in no case is the assumption of closure rejected. (However, Otis *et al.* 1978 emphasise that this test has very low power). The next three columns (tests 1—3) consider whether M_h M_b or M_t are necessary, or whether the simpler M_o is adequate. The null hypothesis is that M_o is an adequate model. The next six columns (test 4—7) consider whether the hypothesised model M_h, M_b or M_t is sufficient to account for the data, or whether other assumptions are necessary. (Note that M_o is a special case of each of these). Tests 5A and 5B consider whether individual heterogeneity and/or time affect the probability of first capture or recapture, respectively, over and above the effects of behavioural response. Test 5 is the sum of these two tests. In each case the null hypothesis is that the model at the top of the column is sufficient. The final goodness-of-fit tests (tests 8—11) consider whether the data fit the hypothesised distributions in their zero-truncated form. The null hypothesis is that they do. If all individuals are equally trappable, f_j should fit a zero-truncated Binomial distribution. However, when capture probability is not too high, Eberhardt (1969) has suggested that the zero-truncated Poisson is a reasonable approximation. Therefore the fit of the data to these distributions provides a further test of equal probabilities of capture between individuals (in addition to test 1).

Several test statistics, notably those concerning the sufficiency of M_t and goodness-of-fit tests for session A, often fail due to inadequate data. Of the goodness-of-fit tests (8—11) there is evidence to reject the hypothesised distribution for only two data sets. The Binomial test suggests unequal p between individuals for the whole population (A+B). Also the total population data for session A only do not fit the Geometric distribution well. The fact that each of four quite different distributions generally fail to be rejected for the same data illustrates the

Table 5

Summary of hypothesis tests concerning closure, model necessity, model sufficiency, goodness-of-fit to truncated distributions and indication of selected models and estimators.

Test no		1	2	3	4	5	5A	5B	6	7	8	9	10	11		
Null hypothesis	Closed	M_o	M_o	M_o	M_h	M_b	M_b	M_b	M_t	M_h	Goodness-of-fit tests				Appropriate	Selected
Alternative hypothesis	Open	M_h	M_b	M_t	Not null hypothesis					M_{bh}	Geom	N.Bin	Binom	Poiss	model	estimator
Session A+B																
Adult male	.81	.95	.35	**.00**	**.01**	**.04**	**.04**	.20	—	.29	.13	—	.91	.80	M_{th}	Darroch
Adult female	.82	.12	.49	.27	.51	.66	.58	.61	—	.38	.15	.20	.47	.23	M_o	Null
Juvenile	.52	.27	—	**.01**	**.02**	.24	—	.24	—	**.02**	.60	.23	.13	.41	M_{tbh}/M_o	Null
Total	.94	**.04**	**.00**	**.00**	**.00**	**.02**	**.01**	.24	—	**.00**	.08	.47	**.04**	.07	M_{tb}	Zippin
Session A																
Adult male	.98	—	.50	.34	.29	.11	.24	.12	—	.06	—	—	—	—	M_o	Null
Adult female	.97	.07	.79	.23	.30	.16	.15	.30	—	.07	—	—	—	—	M_o	Null
Juvenile	.79	—	—	.33	.15	.18	—	.18	—	.08	—	—	—	—	M_o	Null
Total	.89	.18	.29	.08	.13	.34	.28	.43	—	**.00**	**.01**	—	.18	.08	M_o	Null
Session B																
Adult male	.99	.88	**.02**	**.00**	**.02**	.16	.06	.59	—	**.02**	.07	—	.85	.42	M_{th}	Zippin
Adult female	.99	.43	.09	.56	.72	.63	.89	.28	—	.42	.35	.44	.54	.38	$M_{tbh}/M_{bh}/M_o$	Removal
Juvenile	.98	.38	.67	**.01**	.26	.53	.57	.42	—	.38	.60	—	.37	.50	M_o	Null
Total	.99	.68	**.02**	**.00**	**.01**	**.00**	**.00**	.07	.52	**.00**	.10	—	.36	.80	M_{tb}/M_{th}	Darroch
Session A+B (evening only)																
Total	.08	.31	**.03**	.44	.47	.68	.87	.35	—	.26	.36	.40	.14	.58	M_{tbh}/M_{bh}	Removal

Values indicate probability of the null hypothesis being correct. Figure for tests which reject the null hypothesis are in bold type. Where more than one model has a selection value >0.95, these are all indicated. Where no estimator exists for the appropriate model, the model with the highest selection value for which there is an estimator is chosen. All selected estimator models had selection values >0.75. Dashes indicate failure of test due to inadequate data.

low power of such tests, as emphasised by Roff (1973b), Carothers (1973), Cormack (1979).

Interpretation of tests 1—7 is not straightforward, since, as pointed out above, one cause of variation in p may affect the test statistic designed to examine an alternative cause. Accordingly the multivariate algorithm of Otis *et al.* (1978) has been used to aid interpretation of this table. Nevertheless, certain general results are clear. The session A tests, with one exception, all fail to reject any of the null hypotheses. This undoubtedly is a reflection of poor sample sizes. This comment probably applies in lesser degree to the session B data.

The final two columns of Table 5 present the results of the model selection procedure of Otis *et al.* (1978) and, based upon this, the recommended estimator. Generally, if tests 1, 2 or 3 fail to reject the null hypothesis, M_o is considered appropriate and estimator "Null" selected. This estimator is chosen in seven out of thirteen data sets. Tests for Session B reveal somewhat similar patterns to those for corresponding groups in Sessions A+B, reflecting the non-independence of these data. This in turn implies that the causes of variation in p over the whole period are generally not reduced by considering only the latter period. In particular, there is evidence from Table 5 of both temporal and behavioural effects on p for A+B and B, total. However, when only evening traprounds are included in the analysis, the temporal effect seems less strong. This indicates that the major source of variation in captures with time was not in Session A compared to B, but between morning and evening traprounds. This is suggested by comparing the number of captures (n_j from Table 3) per morning trapround ($\bar{x}=16.8$) to the evening ($\bar{x}=22.3$) — paired test, $t=2.65$, $p<0.05$. In most cases, where a more complex model than M_o is selected, it seems only safe to state that there is evidence for some interaction between M_t, M_h and M_b.

3.5. Population Estimates

The estimates of N (generically referred to as $\hat{N}_x$) from each method for each sampling scheme are presented in Table 6. Beneath each value of $\hat{N}_x$ is its percentage relative deviation (RD) defined as $100.(\hat{N}_{CH}-\hat{N}_x)/\hat{N}_{CH}$. It has already been noted that the precision of an estimate is likely to be a function of sampling intensity, hence those values of N based on restricted groups or time periods might be expected, on average, to yield larger RD values. (Note that in this context bias and error are confounded.) In order to provide some yardstick to which the RD of each estimate may be compared, Table 6 also includes the total number of different individuals trapped in each sampling scheme (M_{j+1}) and the RD of this figure.

Not surprisingly, for each case, the three estimators of M_o show rather similar patterns to one another, as do the two for M_t. All five are ge-

Table 6

Population estimates ($\hat{N}_x$) (above) and their percentage relative deviation (RD) (below) from trapipng data for each subgroup and each time period.

Selected estimates and RDs are in bold type.

	Model			M_o			M_h		M_b	M_{bh}		M_t
	Estimator	M_{j+l}	Binom	Poiss	Null	Geom	N.Bin	Jack	Zippin	Removal	Darro	Sch-Esch
Session A+B	Adult male	52	60	62	59	95	60	65	64	53	**59**	59
		—20.4	—8.4	—5.8	—9.6	+45.3	—8.4	—0.2	—2.0	—18.8	**—9.6**	—10.3
	Adult female	66	74	76	**73**	115	78	82	76	76	73	73
		—33.1	—24.9	—23.0	**—26.0**	+16.5	—21.3	—16.7	—22.9	—22.9	—26.0	—26.2
	Juvenile	42	52	54	**51**	86	124	65	—	158	42	46
		—40.6	—26.4	—24.0	**—27.9**	+21.8	+75.2	—8.2	—	+123.5	—40.6	—34.8
	Total	160	185	190	184	293	205	214	**224**	191	183	182
		—31.3	—20.7	—18.5	—21.0	+26.0	—12.1	—8.2	**—3.8**	—18.0	—21.4	—21.8
Session A	Adult male	32	66	74	**64**	134	45	57	150	154	63	62
		—51.0	+0.3	+12.7	**—2.0**	+105.8	—31.7	—12.7	+129.7	+135.8	—3.5	—5.1
	Adult female	43	74	82	**73**	146	56	67	66	43	72	71
		—56.4	—24.8	—16.9	**—26.0**	+47.8	—43.7	—32.0	—33.1	—56.4	—27.0	—28.5
	Juvenile	17	32	36	**31**	65	—	57	—	—	17	36
		—76.0	—54.9	—49.1	**—56.2**	—7.8		—19.4			—76.0	—49.5
	Total	92	170	189	**169**	341	149	175	388	396	168	166
		—60.5	—27.0	—18.7	**—27.4**	+46.4	—36.2	—24.8	+66.7	+70.1	—27.8	—28.7
Session B	Adult male	44	61	66	60	111	59	64	**47**	50	59	58
		—32.6	—6.4	+0.3	—8.1	+69.7	—9.2	—1.4	**—9.2**	—23.4	—9.6	—11.3
	Adult female	54	71	76	70	127	78	82	60	**60**	70	70
		—45.2	—27.6	—22.7	—29.0	+28.6	—20.5	—16.8	—39.1	**—39.1**	—29.0	—29.4
	Juvenile	34	45	48	**44**	80	52	52	41	41	44	45
		—51.9	—36.2	—31.8	**—37.8**	+13.4	—26.4	—27.2	—42.0	—42.0	—37.8	—36.8
	Total	132	178	190	176	317	189	216	150	219	175	173
		—43.3	—23.8	—18.6	—24.4	+36.3	—19.0	—7.3	—35.6	—5.9	**—24.8**	—25.7
Session A+B	Total	28	189	205	188	352	231	255	347	**348**	187	190
(evenings)		—45.0	—18.9	—12.1	—19.2	+51.4	—0.9	+9.5	+49.0	**+49.5**	—19.7	—18.4

nerally negatively biased, often quite considerably. The Geometric generally overestimates $\hat{N}_{CH}$, with a tendency to greatly overestimate with sparse data sets. Both the Negative Binomial and Jackknife generally show a negative bias, but this is usually more extreme in the former. Also the Negative Binomial's behaviour is erratic for poor data sets (juvenile A+B and A only). This comment applies even more so to the Zippin and Removal estimators. Although there is no consistent direction to their RD values (i.e., no evidence overall for true bias) their magnitude is often great.

For the four estimators based upon zero-truncated distributions, no ınference may be drawn from the goodness-of-fit p values (Table 5) about the truthfulness of a particular estimate. For instance, a better fit to the Poisson is found for total (B) than adult male (B) ($p=0.80$ and 0.42 respectively) and yet the latter estimate shows a much closer correspondence to $\hat{N}_{CH}$ (Table 6).

As an arbitrary convention, I shall consider an estimate as "reasonable" if its |RD| is$<$0.5 |RD| of M_{j+1}, and "poor" if |RD| $>$0.5 | RD| of M_{j+1}. The "best" estimate is that which is colsest to $\hat{N}_{CH}$ (smallest RD). The "selected" estimate is that estimate, of the five considered by Otis *et al.* (1978), which their parsimonious selection algorithm indicates is the most appropriate. Table 5 shows that the most frequently selected estimator is "Null" (7/13 cases). However, this estimator was the best in only 2/13 cases. Altogether, out of thirteen cases, the selected estimator was poor in seven, reasonable (but not best) in three and best (and reasonable) in three. Therefore if one were using assumption tests to choose an estimator (from the five in Otis *et al.* 1978) in 77% of cases another estimator (from the ten considered here) would have done better than that chosen. Over thirteen data sets the number of times each estimator was the best is: Jackknife 5$^1/_2$ (one test joint best), Geometric 2$^1/_2$, Null 2, Zippin 1, Removal 1, Negative Binomial 1. Therefore the Jackknife yielded better results more often than any other estimator. Moreover, the Jackknife estimate was reasonable in 10/13 cases (cf. Null above). Although the RD of the Jackknife sometimes exceeded 30% (worse for sparse data showing low recapture rates) it is never more than 14% worse than the best estimate available. Of the six truly independent data sets Null is poor in three and reasonable in three; Jackknife is poor in two, reasonable in two and best (and reasonable) in two. However, these six sets of necessity include the weakest data (few individuals and few recaptures).

3.6. Coefficients of Variation

The preceding analysis has considered all estimates (including $\hat{N}_{CH}$) as a single value. However, since each is an estimate it should have an associated confidence interval. Although theoretical variance estimates

Table 7

Estimated coefficients of variation for population estimates presented in Table 5, defined as 100 $(se\hat{N})/\hat{N}$.

		Binom	Poisson	Null	Geom	N.Bin	Jack	Zippin	Removal	Darro	Sch-Esch
Session A+B	Adult male	11.6	11.7	6.0	12.6	7.9	7.4	14.1	3.6	5.5	7.3
	Adult female	9.6	9.8	4.8	10.6	8.9	6.6	9.0	9.0	4.4	5.3
	Juvenile	14.4	15.0	8.7	15.8	123.4	11.0	—	153.3	0.0	10.1
	Total	6.9	6.7	3.5	7.2	8.0	4.7	12.3	18.6	3.4	3.9
Session A	Adult male	33.2	31.3	24.3	31.6	17.5	13.8	229.5	237.8	23.1	36.3
	Adult female	23.2	23.2	16.6	23.5	14.2	20.3	32.1	3.0	15.9	21.8
	Juvenile	35.1	40.1	29.4	40.8	—	23.3	—	—	0.1	40.0
	Total	14.7	16.6	12.7	17.2	16.8	9.1	117.3	121.4	12.4	13.4
Session B	Adult male	16.5	17.5	11.5	18.6	15.5	9.7	6.7	12.7	10.8	14.3
	Adult female	14.4	14.9	9.3	15.8	18.9	9.7	7.8	7.8	9.0	12.4
	Juvenile	15.9	19.1	11.7	20.0	27.9	11.7	16.5	16.6	10.8	8.6
	Total	7.9	9.8	6.2	10.3	11.4	7.4	5.8	39.0	6.1	11.0
Session A+B (evening only)	Total	10.4	11.1	7.6	11.7	20.8	9.4	53.2	54.0	7.7	10.9

are available for each $\hat{N}_x$, their usefulness should be questioned following the simulation studies of Manly (1971) and Roff (1973a). In particular these studies have revealed that variance estimates and the parameter estimate to which they relate may be highly correlated. Hence if $\hat{N}_x$ is biased, so will be Var (N_x), leading to a false impression of precision. Nonetheless, variance estimates provide one way of describing the effect of sampling intensity for a particular estimate.

The last column of Table 2 expresses the 95% confidence interval of $\hat{N}_{CH}$ as a percentage of $\hat{N}_{CH}$. These values can be directly compared to the RD of the appropriate $\hat{N}_x$ in Table 6. This comparison reveals that in the majority of cases the point estimates $\hat{N}_x$ lie outside the 95% confidence limits of $\hat{N}_{CH}$. In order to examine the error associated with $\hat{N}_x$ values, their coefficients of variation (CV) are presented in Table 7. The obvious result is that as numbers of individuals, and numbers of recaptures diminish, CVs increase. Hence, in general, CV for $A+B<B<A$ for all groups and within a session CVs for total<female<male <juvenile. Therefore we have the lowest confidence in male and juvenile estimates, session A, and it is in these that RD is also at its greatest (Table 6), in general. The trends in CV for all M_o estimators, the Jackknife and the Schumacher-Eschmeyer are very similar, although the magnitude varies up to twofold for any particular case. The CV estimates for Zippin, Removal and Darroch are rather erratic, the latter tending to extreme (unjustified) conservatism in two cases, whilst the former may produce such a large CV as to make the estimate worthless, or no CV can be calculated at all. In such cases Otis *et al.* (1978) suggest that $\hat{N}_x$ should be ignored.

The RD estimates of Table 6 show no general positive relationship with CV values of Table 7. Hence, comparison of CVs between estimates for a particular case could not be used to identify the most "precise" $\hat{N}_x$ value. This result is in accordance with the discussion of bias, above.

4. DISCUSSION

The preceding analysis indicates that the model selection procedures used often fail to recommend the best possible estimator. This must partly reflect the unsatisfactory behaviour of several assumption tests (see Introduction). Undoubtedly, more complete data (higher p and/or N) would be less likely to retain the tests' null hypotheses. This is particularly true for session A only. However, the session A+B data are considered to represent data as good as many field studies are able to produce.

We might examine, in a post-hoc fashion, which assumptions are likely to have been important in the present analysis, by comparing overall biases with expected bias from violation of model assumptions (see simulation results of Otis *et al.*, 1978). The five estimators from models

M_o and M_t generally show a negative bias, which is expected when either M_h is appropriate or there is "trap-happiness" in M_b. If M_b or M_{hb} were appropriate, the Zippin or Removal estimates might be expected to yield good results, yet they are generally disappointing. Zippin's method is expected to show negative bias when individual heterogeneity is present, and positive bias when p is low (Otis *et al.*, 1978, p 30—31). If these factors were responsible for the poor performance of M_b, the Removal estimator should do better, especially for session A+B together. However, if u_j does not exhibit a "definite decrease" through the course of the trapping session, the Removal estimator performs poorly (Otis *et al.*, 1978, p 42). Table 3 shows that, for most cases, whilst there is an overall decline in u_j, this is not monotonic, and specifically there are increases in u_j following prebaiting and trap relocation after occasion 8, and also on occasion 18. Therefore although M_{bh} is considered one of the most realistic models discussed, the sensitivity of its estimator to temporal changes in u_j gave rise to poor performance in the present study.

The estimators for model M_h are of two distinct types. The Geometric and Negative Binomial are parametrically based, whilst the Jackknife is nonparametric. The former two (of which the Geometric is a special case of the Negative Binomial) are founded on a model of heterogeneity in p deriving from unequal trap access. Although these methods have been justified on theoretical grounds (Eberhardt, Peterle & Schofield, 1963; Gates & Smith, 1972) support for them comes largely from goodness-of-fit tests and empirical evaluation (Edwards & Eberhardt, 1967; Nixon *et al.*, 1967; Tanton, 1965). Cormack (1979) and Seber (1982) concur that these methods should be regarded as statistical descriptions which might be generated by various, quite different, models. The results of the present study reject the ability of goodness-of-fit tests to support a particular model. Empirical evaluation suggests a general overestimation by the Geometric (often by over 40%), and underestimation by the Negative Binomial. Nonetheless the Geometric was the second most common "best" estimator. Perhaps Overton's (1971, p 445) comment that the Geometric can only be expected as a "transient" distribution is relevant here. Under certain sampling constraints the data may approximate this distribution well (including the zero class) but as the proportion of recaptures increases, the lower capture frequency classes contain relatively fewer individuals and the Geometric becomes less appropriate. Maybe this is the reason for the empirical recommendation of Edwards and Eberhardt (1967) for "...a 50 percent capture and an average capture of 1½ to 2 times per livetrapped animal...". These general conditions were met in only six of the present data sets (see Table 4). Even within these (e.g. each session B analysis) the Geometric was not noticeably better than in other analyses. A good deal of caution is therefore necessary in applying zero-truncated distributions to estimate population size. This is of particular relevance to grey squirrel studies, since Nixon *et al.* (1967) have recommended the Geometric method for this species.

Simulation studies of the Jackknife method have revealed it to be quite robust to variation in p through time or as a consequence of learned trap response (Otis *et al.*, 1978, p 35; Burnham & Overton, 1979). The foregoing discussion indicates that individual capture heterogeneity, temporal and behavioural variation were probably all present in some degree. It is unlikely that a simple model (such as M_h) is ever true in real populations, and so robustness is a highly desirable property of any estimator. This conclusion suggests a potential weakness in the model selection procedure of Otis *et al.* (1978).

The need for a selection algorithm is evident from the difficulty in directly interpreting test statistics as presented in Table 5. The strategy of Otis *et al.* (1978) is to find "...the simplest model that "fits" the data". This parsimonious procedure combined with low-powered hypothesis tests results in model M_o frequently being selected, especially for sparse data sets (see Table 5). Unfortunately, the discussion above indicates that estimators for this model may show considerable bias when its assumptions are not strictly met. A more useful selection procedure might be to choose the model with the most robust estimator which "fits" the data. Such an algorithm would undoubtedly have led to the Jackknife estimates, which were "best" most commonly, being selected more frequently.

Until more powerful assumption tests are available the field biologist is forced to make a choice of estimator on an often inadequate theoretical basis. On the other hand, simply to adopt the pragmatic approach begs the question of generality of the present findings for future work on the same species, other species, other sampling schemes, etc. It is encouraging to note that the reanalysis of Carothers' (1973) known size taxi population by Otis *et al.* (1978, p 81) also found the Jackknife to estimate N well. If the Jackknife method performed best in these studies because it is robust, then it can be cautiously recommended for future studies. It is, of course, desirable to see further studies in this vein to support the generality of this statement. Perhaps the most important general recommendation, however, is that further effort should be directed towards developing generalized robust estimators, and using robustness as a selection criterion, since we are probably never likely to be able to define all the influences on capture probabilities for a single wild population.

Acknowledgements: Above all, I am indebted to Monica Shorten and Fred Courtier for allowing me access to their original data. Gary White kindly provided me with a copy of the program "CAPTURE" by which many of the calculations were achieved. I am grateful to Jeremy Franklin for developing the solutions to certain variance estimates. Mark Boyce, Murray Efford and John Gurnell offered many helpful comments on an earlier draft. I thank John Phillipson for the facilities provided to me through the AERG. This work was conducted whilst the author was supported by NERC.

REFERENCES

1. Bouffard S. H. & Hein D., 1978: Census methods for Eastern gray squirrels. J. Wildl. Manage., *42:* 550—557.
2. Brady J. R. & Pelton M. R., 1976: A comparison of some census techniques for the cottontail rabbit. Proc. Ann. Conf. SE. Assoc. Fish. Wildl. Agencies, *30:* 546—551.
3. Burnham K. P. & Overton W. S., 1979: Robust estimation of population size when capture probabilities vary among animals. Ecology, *60:* 927—936.
4. Carothers A. D., 1973: Capture-recapture methods applied to a population with known parameters. J. Anim. Ecol., *42:* 125—146.
5. Chapman D. G., 1951: Some properties of the hypergeometric distribution with applications to zoological censuses. Univ. Calif. Public. Stat., *1:* 131—160.
6. Cormack R. M., 1979: Models for capture-recapture. [In: Cormack R. M. Patil G. P. & Robson D. S. (Eds.), "Sampling biological populations"]. Statistical Ecology Series., 5: 217—255. International Cooperative Publ., Maryland.
7. Don B. A. C., 1981: Spatial dynamics and individual quality in a population of the grey squirrel *(Sciurus carolinensis).* D. Phil. Thesis. Oxford Univ.
8. Eberhardt L. L., 1969: Population estimates from recapture frequencies. J. Wildl. Manage., *33:* 28—39.
9. Eberhardt L. L., Peterle T. J. & Schofield R., 1963: Problems in a rabbit population study. Wildl. Monogr., *10:* 1—51.
10. Edwards W. R. & Eberhardt L. L., 1967: Estimating cottontail abundance from live-trapping data. J. Wildl. Manage., *31:* 87—96.
11. Flyger V. F., 1959: A comparison of methods for estimating squirrel populations. J. Wildl. Manage., *23:* 220—223.
12. Gates C. E. & Smith W. B., 1972: Estimation of density of mourning doves from aural information. Biometrics, *28:* 345—349.
13. Hartley H. O., 1958: Maximum likelihood estimation from incomplete data. Biometrics, *14:* 174—194.
14. Keith L. B. & Meslow E. C., 1968: Trap response by snowshoe hares. J. Wildl. Manage., *32:* 795—801.
15. Manly B. F. J., 1970: A simulation study of animal population estimation using the capture-recapture method. J. Appl. Ecol., *7:* 13—39.
16. Manly B. F. J., 1970: A simulation study of Jolly's method for analysing capture-recapture data. Biometrics, *27:* 415—424.
17. Mares M. A., Streilein K. E. & Willig M. R., 1981: Experimental assessment of several population estimation techniques on an introduced population of Eastern chipmunks. J. Mammal., *62:* 315—328.
18. Mosby H. S., 1969: The influence of hunting on the population dynamics of a woodlot gray squirrel population. J. Wildl. Manage., *33:* 59—73.
19. Nixon C. M., Edwards W. R. & Eberhardt L., 1967: Estimating squirrel abundance from livetrapping data. J. Wildl. Manage., *31:* 96—101.
20. Otis D. L., Burnham K. P., White G. C. & Anderson D. R., 1978: Statistical inference from capture data on closed animal populations. Wildl. Mongr., *62:* 1—135.

21. Overton W. S., 1971: Estimating the numbers of animals in wildlife populations. [In: Giles R. H. (ed.), "Wildlife Management Techniques"]: 403—455. 3rd ed. The Wildlife Society, Washington, DC.
22. Roff D. A., 1973a: On the accuracy of some mark-recapture estimators. Oecologia (Berl.), *12:* 15—34.
23. Roff D. A., 1973b: An examination of some statistical tests used in the analysis of mark-recapture data. Oecologia (Berl.), *12:* 35—34.
24. Romersburg H. C. & Marshall K., 1979: Fitting the geometric distribution to capture frequency data. J. Wildl. Manage., *43:* 79—84.
25. Seber G. A. F., 1982: The estimation of animal abundance and related parameters. 2nd Edition. Griffin: 1—654. London.
26. Shorten M., 1951: Some aspects of the biology of the grey squirrel *(Sciurus carolinensis)* in Great Britain. Proc. zool. Soc. Lond., *121:* 427—459.
27. Shorten M. & Courtier F. A., 1955: A population study of the grey squirrel *(Sciurus carolinensis)* in May 1954: Ann. Appl. Biol., *43:* 495—510.
28. Tanton M. T., 1965: Problems of live-trapping and population estimation for the wood mouse, *Apodemus sylvaticus* (L.). J. Anim. Ecol., *34:* 1—22.
29. Zarnoch S. J., 1979: Simulation of effects of learned trap response of three estimators of population size. J. Wildl. Manage., *43:* 474—483.

Accepted, January 25, 1984.

Bruce A. C. DON

EMPIRYCZNE SPRAWDZANIE RÓŻNYCH ESTYMATORÓW LICZEBNOŚCI NA PRZYKŁADZIE *SCIURUS CAROLINENSIS*

Streszczenie

Dokonano ponownej analizy (Tabele 3—7) wyników wcześniej opublikowanych (Shorten & Courtier, 1955) badań nad liczebnością populacji *Sciurus carolinensis.* Użyto dziesięciu różnych estymatorów liczebności populacji (Tabela 1) a oznaczenia liczebności oparto też na próbnych odstrzałach wiewiórek (Tabela 2), co pozwoliło uniknąć zmian wywołanych zróżnicowaniem w łowności. Wartości populacyjne obliczone przez dane z odstrzałów porównano do wyników z odłowów. Użyto także kilku testów do oceny założeń przyjmowanych przy odłowie, wypuszczaniu i ponownym łowieniu tych zwierząt. Wyniki wskazują, że procedura przyjmowania założeń i doboru modelu często daje nietrafne wskazania estymatora, który byłby najlepszym z dostępnych. Uważa się, że jest to odzwierciedlenie faktu, iż założenia niezbyt dobrze precyzują uwarunkowania danego estymatora. Największą zgodność danych uzyskano, wśród użytych estymatorów dla wskaźnika zwanego „Jacknife".