Ecology, 75(6), 1994, pp. 1780–1793

AIC MODEL SELECTION IN OVERDISPERSED CAPTURE–RECAPTURE DATA[1]

D. R. ANDERSON AND K. P. BURNHAM
Colorado Cooperative Fish and Wildlife Research Unit, U. S. Fish and Wildlife Service,[2] Fort Collins, Colorado 80523 USA

G. C. WHITE
Department of Fishery and Wildlife Biology, Colorado State University, Fort Collins, Colorado 80523 USA

Abstract. Selection of a proper model as a basis for statistical inference from capture–recapture data is critical. This is especially so when using open models in the analysis of multiple, interrelated data sets (e.g., males and females, with 2–3 age classes, over 3–5 areas and 10–15 yr). The most general model considered for such data sets might contain 1000 survival and recapture parameters. This paper presents numerical results on three information-theoretic methods for model selection when the data are overdispersed (i.e., a lack of independence so that extra-binomial variation occurs). Akaike's information criterion (AIC), a second-order adjustment to AIC for bias (AIC_c), and a dimension-consistent criterion (CAIC) were modified using an empirical estimate of the average overdispersion, based on quasi-likelihood theory. Quality of model selection was evaluated based on the Euclidian distance between standardized $\hat{\theta}$ and θ (parameter θ is vector valued); this quantity (a type of residual sum of squares, hence denoted as RSS) is a combination of squared bias and variance. Five results seem to be of general interest for these product-multinomial models. First, when there was overdispersion the most direct estimator of the variance inflation factor was positively biased and the relative bias increased with the amount of overdispersion. Second, AIC and AIC_c, unadjusted for overdispersion using quasi-likelihood theory, performed poorly in selecting a model with a small RSS value when the data were overdispersed (i.e., overfitted models were selected when compared to the model with the minimum RRS value). Third, the information-theoretic criteria, adjusted for overdispersion, performed well, selected parismonious models, and had a good balance between under- and overfitting the data. Fourth, generally, the dimension-consistent criterion selected models with fewer parameters than the other criteria, had smaller RSS values, but clearly was in error by underfitting when compared with the model with the minimum RSS value. Fifth, even if the true model structure (but not the actual parameter values in the model) is known, that true model, when fitted to the data (by parameter estimation) is a relatively poor basis for statistical inference when that true model includes several, let alone many, estimated parameters that are not significantly different from 0.

Key words: AIC; Akaike; capture–recapture; Cormack–Jolly–Seber model; extra-binomial variation; Kullback–Leibler discrepancy; model selection; overdispersion.

INTRODUCTION

Open models based on the Cormack–Jolly–Seber model

Pollock et al. (1990) summarize the state of the science for the general Jolly–Seber model, which includes time-specific parameters for population size (N_i), probability of capture (p_i), and number of new recruits (B_{i-1}) at time i for $i = 1, \ldots, k$ and the probability of survival (ϕ_i), during the interval i to $i + 1$. Anderson et al. (1993) suggest some general trends in capture–recapture modeling in open populations. In particular, in recent years, interest has increasingly focused on the Cormack–Jolly–Seber (CJS) model in recognition of Cormack's (1964) paper as well as the papers by Jolly (1965) and Seber (1965). The CJS model is a product-multinomial model incorporating only time-specific survival (ϕ_i) and recapture (p_i) probabilities and since the late 1980s it has seen several major extensions. Burnham et al. (1987) generalize the CJS model to allow analysis of multiple data sets, in particular the cases where there are treatment vs. control group contrasts. Lebreton et al. (1992) extend the modeling of the survival and recapture probabilities in the framework of a general "analysis of variance" philosophy with an emphasis on the analysis of multiple data sets and logit-linear modeling of external variables. Burnham (1991, 1993) provides a major synthesis and unification and gives the general theory for the joint analysis of capture–recapture and band recovery data (also see Seber 1982).

Closed-form parameter estimators exist for only a

[1] Manuscript received 21 June 1993; revised 11 November 1993; accepted 1 December 1993.

[2] The Unit is now part of the U.S. Department of Interior National Biological Survey.

few capture–recapture models and those that do exist are computationally intensive. Computation of test statistics also involves a substantial amount of calculation that is quite error-prone if done by hand. Thus computer software is essential in data analysis, especially in the case of multiple data sets. Sophisticated software now exists but is not always easy to use. Programs RELEASE, SURGE, and SURVIV (see summary of these programs and others in Lebreton et al. 1992:86) are most relevant here.

Model selection

The primary issue in the analysis of capture–recapture data is that of proper model selection (Burnham and Anderson 1992). This is especially critical in the analysis of multiple, interrelated data sets (e.g., males and females each represented by 2–3 age classes). Such a study over 10–20 yr (occasions) will also usually have several "time effects." As Jolly (1965) anticipated, some of the survival (ϕ) or recapture (p) parameters might also be common across age or sex classes or years, while others must be sex- or age- or year-specific, or be a function of external covariates. The statistical naive approach to analysis of such data is to just fit the most general model that includes all reasonable effects possible in the data; we define such a general model as the "global model" (Burnham and Anderson 1992, Lebreton et al. 1992).

The global model for such a single large, multifactor data set might have on the order of 200 parameters and the global model for such data sets over several geographic areas might contain over 1000 parameters. This is too many parameters to easily interpret and usually most of the estimates of these parameters do not represent (in analysis of variance parlance) statistically significant effects in the data. For these and other reasons, the selection of a parsimonious model is very important in data analysis: a finite amount of data will only "support" a certain number of parameters and a limited model structure. Moreover, it is difficult to conceive of a "true model" in capture–recapture; rather as sample size increases, more structure ("effects") can be identified. Shibata (1989) recommends rejection of those models far from reality and the selection of a model in which the error of approximation and the error due to random fluctuations are well balanced. Burnham and Anderson (1992) provide further discussion of these issues and a detailed capture–recapture example in which the global model contained 46 parameters. It is worth noting here that these issues about models, parsimony, model selection, and so forth are common to much of empirical of experimental ecology and data-based model selection methods discussed below have far wider application than just capture–recapture.

Huggins (1991), Burnham and Anderson (1992), and Lebreton et al. (1992) recommend the use of Akaike's (1973, 1985) Information Criterion (AIC) as the basis for model selection in the analysis of capture–recapture data. Akaike reasoned that if one had an objective discrepancy measure (similar to a metric) between any approximating model and the true model, one should select the approximating model for which this measure was smallest. Moreover, there are compelling reasons to use, as Akaike suggested, the Kullback–Leibler (K–L) discrepancy between two distributions, as the basis of such model selection. Denoting the true statistical sampling distribution of the data by $f(x)$ ("truth") and the model by $g(x \mid \boldsymbol{\theta})$ (with a known form but generally unknown parameters, denoted by $\boldsymbol{\theta}$), then the Kullback–Leibler discrepancy is

$$I(f, g) = \int f(x)\log\left[\frac{f(x)}{g(x \mid \boldsymbol{\theta})}\right] dx.$$

The K–L discrepancy has its roots deep in information theory (see, e.g., Kullback 1959) and is by no means an arbitrary choice of metric here; essentially, $I(f, g)$ is a unique metric to use in the context of maximum likelihood theory. Kapur and Kesavan (1992) provide a current review of information-theoretic measures such as $I(f, g)$ (but without reference to the model selection problem in statistical data analysis). Even if a true (unknown) model exists, the K–L discrepancy is not observable, nor can it be computed directly from the sample data. Akaika found a relation between the K–L discrepancy and an expected log-likelihood, and this finding has allowed major practical and theoretical advances in model selection and the analysis of complex data sets (Bozdogan 1987 provides a statistical review of AIC).

Noting that $I(f, g)$ can be written as

$$I(f, g) = \int f(x)\log[f(x)]\, dx - \int f(x)\log[g(x \mid \theta)]\, dx,$$

leads to

$$I(f, g) = \mathrm{E}_x\{\log[f(x)]\} - \mathrm{E}_x\{\log[g(x \mid \boldsymbol{\theta})]\}.$$

The first expectation depends only on the unknown true distribution, $f(x)$, and is not dependent on any approximating model, $g(x \mid \boldsymbol{\theta})$, and its parameters (e.g., Sakamoto et al. 1986:45–48). This first expectation can be considered as an unknown (and not estimable) constant. Akaike (1973, 1985) showed that, apart from this constant term, the K–L discrepancy is equal to an expected log-likelihood. Thus to find the model that maximizes $I(f, g)$ one need only maximize $\mathrm{E}_x\{\log[g(x \mid \boldsymbol{\theta})]\}$, which is directly related to the log-likelihood of the model $g(x \mid \boldsymbol{\theta})$ [the likelihood $\mathcal{L}(\boldsymbol{\theta} \mid \text{data } x)$ is proportional to $g(x \mid \boldsymbol{\theta})$]. Akaike derived a consistent estimator of $\mathrm{E}_x\{\log[\mathcal{L}(\boldsymbol{\theta} \mid \text{data } x)]\} = \mathrm{E}_x\{\log[g(x \mid \boldsymbol{\theta}]\}$ + an unknown constant, when $\boldsymbol{\theta}$ is estimated by maximum likelihood estimation (MLE) (for

more specifics see, for example, Akaike 1973, Bozdogan 1987, Burnham et al. 1994).

The maximized log-likelihood is a biased estimator of this "expected log-likelihood" and the asymptotic bias equals K, the number of free parameters in the model (Akaike 1973). Thus, a consistent estimator of Akaike's expected log-likelihood is $\log[\mathcal{L}(\hat{\theta})] - K$ (Akaike 1973, Bozdogan 1987), where θ is a vector of the model parameters (here, these are the ϕ_i and p_i in CJS models). Akaike then defined AIC by multiplying through by -2 (for "historical reasons," according to Akaike, namely because the multiplier "2" appears as part of all likelihood ratio tests and the "$-$" makes AIC almost always a positive number), hence

$$\text{AIC} = -2 \log[\mathcal{L}(\hat{\theta})] + 2K.$$

A key point is that $-\log[\mathcal{L}(\hat{\theta})] + K$ is a consistent estimator of the K–L discrepancy (plus an additive constant that does not depend upon any aspect of the candidate models examined) and therefore has a solid theoretical basis. We emphasize that model selection based on a AIC has the goal of selecting a model based on the simple yet compelling idea of minimizing the Kullback–Leibler discrepancy between the unknown "true model" (i.e., truth) and the approximating data-based model. Truth could be very complex, even having an infinite number of parameters, thus not be a useful model. Yet if truth were known we could, and generally would, select a best approximating model by minimizing the K—L metric over a suitable class of useful models. AIC model selection allows us to find this data-based best approximation to truth even though truth is unknown to us. In essence, using AIC, we have a well-defined, meaningful target to aim for in model selection.

It is only heuristically that the first term in AIC can be seen as a measure of lack of model fit while the second term is a penalty for increasing the "size" of the model. In this sense, AIC attempts to identify a parsimonious model as a trade-off between increasing K to achieve a good model fit (hence low bias in estimated parameters) and decreasing K to minimize the penalty for having too many parameters (hence increasing precision of estimated parameters) (Burnham and Anderson 1992 comment more on this bias-precision trade-off). AIC as computed for a number of candidate models and the model with the lowest AIC is selected as a basis for inference (Atkinson 1980, Sakamoto et al. 1986). Stone (1977) shows a relationship between AIC and cross validation. Hurvich and Tsai (1989) give a small-sample (second order) bias correction, termed AIC_c, where,

$$\text{AIC}_c = \text{AIC} + \frac{2(K + 1)(K + 2)}{n - K - 2},$$

and n = sample size (also see the earlier work by Sugiura 1978 on second-order bias adjustments in a similar context). AIC_c has a somewhat larger penalty term than AIC, particularly if n is small with respect to K. This version of AIC_c was derived in a time series context, but was shown to be reasonable for CJS models by Burnham et al. (1994).

A criterion for asymptotically consistent model selection was developed by Schwarz (1978) and termed BIC, for Bayesian Information Criterion. Bozdogan (1987) presents a good review of information-theoretic methods in model selection, including what he terms a dimension-consistent criterion (termed CAIC for consistent AIC), which is asymptotically the same as BIC. In this study we investigate Bozdogan's dimension-consistent criterion. Hence we use

$$\text{CAIC} = -2 \log[\mathcal{L}(\hat{\theta})] + K[\log(n) + 1],$$

(see also Shibata 1976). This model selection criterion is useful when a true model exists that has a finite, in fact small, order (K) that does not increase with sample size. This CAIC criterion does not enjoy the theoretical link with the K–L discrepancy, rather it provides an alternative penalty, derived from a Bayesian viewpoint, such that the dimension (order) of the true model is consistently estimated as $n \to \infty$. Thus, AIC (and AIC_c) relates to finite samples and is an estimate of the K–L discrepancy (except for a constant that is unrelated to the data).

As sample size increases, AIC will select increasingly more complex models. The underlying idea is that more data generally does contain more structure, hence merits a more complex model (to collect more data one either must include more areas, species, times, etc., or get increased sample sizes on factors such as age, condition, etc., of animals). In contrast, CAIC focuses on a penalty term that is chosen to allow an asymptotically consistent estimator of the true model dimension (assuming K is fixed as $n \to \infty$). CAIC has a larger penalty term and, thus, selects models with the same or fewer parameters than AIC and AIC_c.

If one takes the K–L discrepancy as the logical starting point it is, then AIC and AIC_c have a deep level of theoretical support. CAIC is popular in the literature, but its foundation and even objective is disputed (see, e.g., Akaike 1981, Bozdogan 1987). The need for asymptotic dimension consistency seems less supported, especially when no true, low-dimensional model is likely to exist. Perhaps such a true model exists in some of the physical sciences. In field biology, as sample size increases the number of factors possible in the model also increases, hence, more model structure is revealed and increasing numbers of parameters are required. Thus, CAIC seems inappropriate a priori for many biological modeling issues.

Burnham et al. (1994) evaluate the adequacy of the first- and second-order bias terms for use in estimating the K–L discrepancy, based on the maximized log-likelihood. They found these bias terms to be adequate

over a wide range of sample sizes, number of sampling occasions, and parameter values. These results paved the way for increased evaluation of K–L-based methods in model selection among candidate capture–recapture models.

Burnham et al. (*in press*) compared the use of AIC, AIC_c, and CAIC with the use of likelihood ratio tests, under four different α values, in a nested sequence of treatment-control models of the CJS type. Their comparison focused not only on the ability to select the "true model," but more on the quality of the resulting inference as measured by a type of expected residual sum of squares (RSS). Let $\boldsymbol{\theta}$ be the vector of model parameters (i.e., the ϕ_i and p_i) under the global (i.e., most general) model; any model fit represents some smoothing restriction on the estimators of $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_G)$. Then,

$$\text{RSS} = \sum_{i=1}^{G} [(\hat{\theta}_i - \theta_i)/\theta_i]^2$$

measures both bias and sampling variation in $\hat{\boldsymbol{\theta}}$. The objective of Burnham et al. (*in press*) was to identify the model selection strategy that picks models with a small RSS. It was shown that AIC, AIC_c, and CAIC consistently outperformed the hypothesis-testing approach in selecting models with the low RSS values. Furthermore, AIC and AIC_c tended to achieve a balance between a model with too few vs. too many parameters (see Shibata 1989). Surprisingly, CAIC performed well, often selecting models with a lower RSS than those selected by AIC and AIC_c. Although the data were simulated from models with a small, fixed K, the CAIC-selected models provided poor estimates of the dimension of the true model (Burnham et al., *in press*).

Overdispersion

Akaike (1973) considers his information-theoretic-based method to be a generalization of Fisher's likelihood theory (which is fundamental to most of statistical theory). More recently, Kapur and Kesavan (1992) also consider likelihood theory to be a special case of information theory. While our evaluation of the information-theoretic methods in model selection in the product-multinomial models in capture–recapture is encouraging (Burnham et al. 1994, *in press*), real capture–recapture data often seem to be overdispersed. The CJS model and its many extensions assume only binomial variation; however, our experience suggests that frequently capture–recapture data are overdispersed. Count data have a long history of not conforming to simple variance assumptions (e.g., Bartlett 1936, Fisher 1949, Armitage 1957, Finney 1971).

The reasons for overdispersion or "extra-binomial variation" are many in biological populations (Eberhardt 1978). The focus here is on a lack of independence in the data. Banded Canada Geese (*Branta* spp.) frequently mate for life and the pair behaves almost as an individual, rather than as two independent "trials." The young of some species continue to live with the parents for a period of time, which also can cause a lack of independence. Further reasons for overdispersion in biological systems include species whose members exist in schools or flocks. Members of such populations can be expected to have positive correlations among individuals; such dependence causes overdispersion. The effect of overdispersion is to cause the variances to be underestimated. An alternative reason for overdispersion, not addressed here, is that parameter heterogeneity in the survival or capture probabilities, or in both, may cause overdispersion.

The estimators of model parameters often remain unbiased in the presence of overdispersion, but the model-based, theoretical variances are underestimated (McCullagh and Nelder 1989). To properly cope with overdispersion one needs to model the overdispersion and then use generalized likelihood inference methods. Quasi-likelihood theory (Wedderburn 1974) provides such a means to handle the analysis of overdispersed data (also see Williams 1982, McCullagh and Pregibon 1985, Moore 1987, McCullagh and Nelder 1989).

In general, if the random variable m represents count data under some simple distribution (e.g., Poisson or binomial) the expectation, $\mu(\boldsymbol{\theta})$, and the variance, $\sigma^2(\boldsymbol{\theta})$, are known functions of the unknown parameter $\boldsymbol{\theta}$. In an overdispersion model the expectation of m is not changed, but the variance model must be generalized, such as by a multiplicative factor, $\gamma(\boldsymbol{\theta})$, hence $\text{var}(m) = \gamma(\boldsymbol{\theta})\sigma^2(\boldsymbol{\theta})$. The form of the factor $\gamma(\boldsymbol{\theta})$ can be partly determined by theoretical considerations and can be complex (see, e.g., McCullagh and Nelder 1989). For CJS models, and in general, the data constitute many interrelated counts, m_{ij} (captures and recaptures, by cohorts) so there are (conceptually) many different overdispersion factors to be modeled: γ_{ij}. This is a daunting task; however, there overdispersion factors typically are small, ranging from just above one to two or three, if the model structure is correct and overdispersion is due to small violations of assumptions such as independence and parameter homogeneity over individuals. Hence, a first approximation for dealing with overdispersion is to use a constant c (conceptually $c = \bar{\gamma}$) in place of each γ_{ij}. Burnham et al. (1987:243–246) and Lebreton et al. (1992:106–107) discuss the estimation of empirical variances and covariances in capture–recapture models using a constant c and quasi-likelihood methods.

Cox and Snell (1989) discuss modeling of overdispersed variances for count data and note that the first useful approximation is based on a single variance inflation factor (c), which can be estimated from the standard goodness-of-fit chi-square statistic (χ^2) and its degrees of freedom for count data, hence

$$\hat{c} = \chi^2/df$$

(χ^2 here is the usual $[O - \hat{E}]^2/\hat{E}$ statistic summed over the *O*bserved count data based on a fitted model giving the *E*xpected counts). Furthermore, Cox and Snell (1989) assert that this simple approach should often be adequate, as opposed to the much more arduous task of seeking a model for the γ_{ij}. In a study of these competing approaches on five data sets, Liang and McCullagh (1993) found that modeling overdispersion was clearly better than use of a single c in only one of five cases examined. Future research on overdispersion in CJS models may try the difficult task of modeling the overdispersion (separately for each multinomial cohort); however, in these initial studies we are using the simplest approach, a single variance inflation factor.

Given $\hat{c}$, empirical estimates of sampling variances $[\text{var}_e(\hat{\theta}_i)]$ and covariances $[\text{cov}_e(\hat{\theta}_i, \hat{\theta}_j)]$ can be computed by multiplying the theoretical (model-based) variances and covariances by $\hat{c}$ (a technique that has long been used, see, e.g. Finney 1971). These empirical measures of variation [i.e., $\hat{c} \cdot \text{var}_e(\hat{\theta}_i)$] must be treated as having the degrees of freedom used to compute $\hat{c}$ for purposes of setting confidence limits or testing hypotheses (Finney 1971).

Under the CJS model theory, $c = 1$; however, with real data we expect $c > 1$, but we do not expect c to exceed ≈ 4 (see Eberhardt 1978). Substantially larger values of c (say, 6–10) are usually caused partly by a model structure that is inadequate, that is the fitted model does not actually represent all the explainable variation in the data. Quasi-likelihood methods of variance inflation are appropriate only after the structural adequacy of the model has been achieved (Burnham et al. 1987:243–254). Lebreton et al. (1992) discuss at length strategies for analysis of CJS capture–recapture data so as to determine an adequate structural model (but with $c = 1$ assumed). The issue of the model's structural adequacy is at the very heart of good data analysis (i.e., the reliable identification of the structure vs. residual variation in the data) so we do not herein attempt a discussion of this matter.

Objectives

The first objective is to evaluate the estimation of a single variance inflation factor based on quasi-likelihood methods from overdispersed capture–recapture data. Secondly, we examine model selection using information-theoretic methods in the presence of overdispersion and, finally, determine if a quasi-likelihood adjustment to the information-theoretic approaches improves their performance with overdispersed data.

Methods

Capture–recapture model and simulated data

Burnham et al. (1987) present a series of nested CJS-type models for the analysis of treatment-control experiments involving marked animals captured over k occasions. This series starts with model H_0 (no treatment effect, survival and capture parameters in both groups are equal) and ends with model $H_{k-1,\phi}$ in which all the survival and recapture probabilities are affected by the treatment and all parameters differ between the two groups. This sequence of nested models is denoted as H_0, $H_{1\phi}$, H_{2p}, $H_{2\phi}$, H_{3p}, $H_{3\phi}$, . . . , $H_{k-1,\phi}$. We assume the reader is familiar with this material. Monte Carlo data were generated using program RELEASE. (Burnham et al. 1987) under models H_0, H_{2p}, $H_{3\phi}$, and $H_{9\phi}$. SAS (SAS 1985) was used for the analysis of the results from program RELEASE.

Model H_0 is a CJS model for two (as used here) groups of marked animals, but there is no treatment effect on survival or recapture probabilities, thus all parameters are equal between the groups (a null model). Parameters and sampling effort constants used in Monte Carlo simulations of the 81 cases (3^4) were $\phi = \{0.5, 0.7, \text{and } 0.9\}$, $p = \{0.4, 0.6 \text{ and } 0.8\}$, u (the number of unmarked animals caught, marked, and released on each occasion) = {50, 100, and 300}, and $k = \{5, 10, \text{and } 20\}$. In each repetition, the ϕ, p, and u were constant over the k sampling occasions. The basic design and parameter values were chosen to be similar to those used by Burnham et al. (*in press*) in the no overdispersion case.

Model H_{2p} is here a CJS model for each of two groups, a treatment and a control group, denoted by subscripts t and c, respectively. The treatment is assumed to affect the first survival probability (ϕ_{t1} vs. ϕ_{c1}) and the first recapture probability (p_{t2} vs. p_{c2}), while the remaining parameters ($\phi_2, \phi_3, \ldots, \phi_{k-1}$ and $p_3, p_4, \ldots, p_k$) are the same for the two groups. Data under model H_{2p} were simulated as under model H_0 but with the following relationships among parameters: $\phi_{t1} = \phi_{c1} - 0.15$ and $p_{t2} = p_{c2} - 0.15$. The treatment effects on ϕ_1 and p_2 are relatively moderate and acute. One might expect that a parsimonious model for real data could identify a treatment effect on ϕ_1 but perhaps fail to identify a treatment effect on p_2 (i.e., the selected model would correspond to concluding the treatment effect does not extend beyond ϕ_1); still the statistical inference that is supported by the data might be quite useful (while not achieving full reality).

Model $H_{3\phi}$ is here a CJS model for a treatment and control group, in which the treatment is assumed to have chronic effects on ϕ_{t1}, ϕ_{t2}, ϕ_{t3}, and p_{t2}, and p_{t3} while the remaining parameters ($\phi_4, \phi_5, \ldots, \phi_{k-1}$ and $p_4, p_5, \ldots, p_k$) are the same for both groups. Data under model $H_{3\phi}$ were simulated as with model H_0 but with the following relationships among parameters: $\phi_{t1} = \phi_{c1} - 0.15$, $\phi_{t2} = \phi_{c2} - 0.15/2$, $\phi_{t3} = \phi_{c3} - 0.15/4$, and $p_{t2} = p_{c2} - 0.15$, $p_{t3} = p_{c3} - 0.15/2$. Here a parsimonious model for real data might be substantially different from the "true" model, especially when sample size is small.

Model $H_{9\phi}$ was investigated less intensely using the

values for ϕ, p, and u as in model H_0, but only for $k = 10$. Thus, there were 27 cases (3^3) and dampened chronic treatment effects were simulated using the following relationships: $\phi_{ti} = \phi_{ci} - (0.1)(0.8)^{i-1}$ for $i = 1, \ldots, 9$ and $p_{ti} = p_{ci} - (0.1)(0.8)^{i-2}$ for $i = 2, \ldots, 9$.

Sample size is $n = \Sigma_{i=1}^{k-1} R_i$, where R_i is the total number of animals released at occasion i; in the underlying multinomial models, these R_i's are the sample sizes of each released cohort. When there are no losses on capture, R_i is the sum of the newly marked animals released on occasion i, u_i, and animals recaptured and rereleased on occasion i (Burnham et al. 1994). The u_i are taken here as given, rather than being generated as random variables because the population dynamics process that produces the u_i (and hence their partially stochastic nature) is irrelevant to our purposes. We assume no losses on capture in the simulations. We believe these models and parameters generally reflect many biological situations that occur in practice, at least for vertebrate species sampled (and reproducing) on an annual basis.

We do not feel that our results are tightly linked to this specific sequence of models. In particular, our results are not specific to models of a treatment effect on the parameters. Rather, these models were chosen because they are a nested sequence and they have closed-form parameter estimators, thus providing a convenient basis for exploring model selection issues in capture–recapture data analysis. (Closed-form MLEs are important because if numerical methods were required there would be a 5–10 fold increase in computing time).

Monte Carlo study

We generated 1000 repetitions for each of the 81 basic cases, that is, four factors (u, k, p, and ϕ) each at three levels (hence this aspect of the design is a 3^4 factorial), for each of three true models (H_0, H_{2p}, $H_{3\phi}$) (thus overall we have a 3^5 factorial design). Data were generated at two levels of overdispersion ($c = 2$ and 4) and a null case with no overdispersion ($c = 1$), thus a total of 729 cases (3^6 factorial) were simulated, each with 1000 repetitions. The models allow ϕ_i and p_i to vary by occasion (i), but we used, with no loss of generality, $\phi_i \equiv \phi$ and $p_i \equiv p$.

A modified version of Program RELEASE (Burnham et al. 1987) was used to generate the simulated CJS data. Given an "animal" is marked and released on occasion i, its subsequent capture history (next $k - i$ occasions) is produced by generating a series of independent Bernoulli events: did the animal survive the next time interval (probability $\equiv \phi_i$), if not, all remaining capture events are set to 0; if it survived, was it caught (probability $\equiv p_{i+1}$), if yes, capture event is set at 1 for occasion $i + 1$; then the animal is released and the process is repeated for occasion $i + 2$.

The generated data are stored (represented) as a capture history matrix, X (Burnham et al. 1987:28–33). Each row (i) of this matrix corresponds to a marked individual and columns (j) in this matrix correspond to sampling occasions. If the individual is captured at occasion j, then a 1 is recorded in the j^{th} column, otherwise a 0 denotes that the individual was not captured on occasion j. To generate overdispersed data (i.e., $c = 2$ or 4) such that $E(n)$ is the same as that where there is no overdispersion ($c = 1$), the u_i (number of animals first captured at time i) were set to μ_i/c, and the X matrix was generated on this basis but then "cloned" c times. By cloned we mean each generated capture history was present c times in the final X matrix used in the data analysis.

A minor complication is that when $c = 4$ and $u_i = 50$, the ratio u_i/c is not an integer. Then, this ratio was randomly rounded to either 12 or 13 for each repetition, hence, the expectation for the number of new releases was still 12.5.

This procedure produced Monte Carlo generated data where the overdispersion parameter was known theoretically to be c for the specific values of u_i (i.e., we simulated data this way because we then would know the true value of c). The simulated data can be further summarized in a compact (compared to the X matrix) array m_{ij}, where m_{ij} is the number of animals first captured at occasion j from releases at occasion i. Rows in the m_{ij} matrix are then multinomial-distributed random variables (Burnham et al. 1987:45–47).

Quasi-likelihood corrections

Quasi-likelihood estimates of the variance inflation factor (c) are computed from the full goodness-of-fit test for the CJS model for each group. Here, two groups are used, corresponding to a treatment and control group. The theory and notation for these tests are nontrivial and are not given here (see Burnham et al. 1987: 64–77, 174–177). The full goodness-of-fit test is the sum of two components (i.e., TEST2 + TEST3). The degrees of freedom is the sum of the degrees of freedom for the two test components. Substantial pooling of the data is required to avoid small expectations (e.g., to avoid expectations <2). While such pooling is somewhat arbitrary, program RELEASE accomplishes a reasonable pooling algorithm to obtain a test statistic that is asymptotically chi-square-distributed with appropriate degrees of freedom. The Monte Carlo data that were generated to achieve overdispersion can be represented as cm_{ij}, where these hypothetical m_{ij} fit the CJS model with no overdispersion. Thus, conceptually, the terms in the goodness-of-fit test are (Observed data are the m_{ij})

$$[cm_{ij} - c\hat{E}(m_{ij})]^2/c\hat{E}(m_{ij}),$$
$$= c^2[m_{ij} - \hat{E}(m_{ij})]^2/c\hat{E}(m_{ij}),$$
$$= c[m_{ij} - \hat{E}(m_{ij})]^2/\hat{E}(m_{ij}).$$

This is merely c times the usual chi-squared terms in the goodness-of-fit test. The sum of these terms is thus

c times an asymptotically central χ^2-distributed variable under the null hypothesis. Then, $E(\chi^2) = c(df)$ and $E(\chi^2/df) = c$, except that pooling makes this only an approximation. Finally, $\hat{c} = \chi^2/df$; the simulation results allow this estimator to be evaluated regarding its bias and precision.

Quasi-likelihood theory suggests simple modifications to AIC, AIC_c, and CAIC (Lebreton et al. 1992: 106–107); we denote these modifications as

$$QAIC = -\{2 \log(\mathcal{L}(\hat{\theta})/\hat{c}\} + 2K,$$

$$QAIC_c = QAIC + \frac{2(K + 1)(K + 2)}{n - K - 2},$$

and

$$QCAIC = -\{2 \log[\mathcal{L}(\hat{\theta})]/\hat{c}\} + K[\log(n) + 1].$$

If no overdispersion exists, then $-2 \log[\mathcal{L}(\hat{\theta})]$ is a measure of lack of fit. However, in the presence of overdispersion, this quantity exaggerates the lack of fit and thus a modification is needed to better enforce parameter parsimony.

RSS metric

Quality of inference was measured by the expected residual sum of squares [E(RSS)] of parameter estimates about the true parameter values, for a given model selection method, over 1000 Monte Carlo repetitions:

$$RSS = \sum_v \left[\sum_{i=1}^{k-2} \left(\frac{\hat{\phi}_{vi} - \phi_{vi}}{\phi_{vi}} \right)^2 + \sum_{i=2}^{k-1} \left(\frac{\hat{p}_{vi} - p_{vi}}{p_{vi}} \right)^2 + \left(\frac{\hat{\beta}_{vk} - \beta_{vk}}{\beta_{vk}} \right)^2 \right],$$

(the parameter estimates vary by replicate, but the notation used here is not elaborated to show that variation) where $v = \{c, t\}$, for control and treatment and β_k is the product $\phi_{k-1}p_k$; only β_k is identifiable here. For each selection method, the RSS value is computed for the selected model, then these 1000 values are averaged to give

$$\hat{E}(RSS) = \frac{\sum_{rep=1}^{1000} RSS_{rep}}{1000}.$$

This (estimated) expected RSS is computed for the AIC-, AIC_c-, and CAIC-selected models and denoted AICRSS, AIC_cRSS and CAICRSS, respectively. Likewise, $\hat{E}$(RSS) is computed for models selected using the quasi-likelihood corrections (QAIC, $QAIC_c$, and QCAIC) and denoted QAICRSS, $QAIC_c$RSS, and QCAICRSS, respectively. If data are generated under a particular model and analyzed under this same (true) model for all 1000 repetitions, then its $\hat{E}$(RSS) is called TRUERSS.

Finally, during the analysis of each specific Monte Carlo repetition for each of the 729 cases, all models in the set $[H_0, H_{1\phi}, H_{2p}, \ldots,$ and $H_{k-1,\phi}\}$ are fitted to the data (via MLE) and the RSS is computed for each of these fitted models. The model with the lowest RSS is determined and the $\hat{E}$(RSS) of this selection procedure is denoted MINRSS. This model selection procedure represents the best possible selection outcome (under the RSS metric) in the entire set of models (not just the three models from which data were generated). Many comparisons are made using MINRSS as the standard, rather than TRUERSS from the true model (e.g., Figs. 1 and 2).

The numerical evaluation of 1000 repetitions of each of the 729 cases took ≈14 d of CPU (Central Processing Unit) time on Sun SPARC Model IPX running UNIX at Colorado State University. This large effort was necessary to meet the objectives over a broad range of parameter values, sample sizes, sampling occasions, and underlying models for overdispersed capture–recapture data, at least for open models of the CJS type.

Results

The patterns in the results were generally similar across, at least, the four basic factors (ϕ, p, u, and k), thus the material in Table 1 and Figs. 1 and 2 is pooled over these 81 cases. When patterns differed by true model (H_0, H_{2p}, or $H_{3\phi}$), we present the results partitioned by these models. The results presented are always partitioned by the dispersion parameter (c = 1, 2, and 4).

Bias in ĉ

Before turning to the central issue of model selection, it is of interest to evaluate the estimation of c using quasi-likelihood methods. Clearly, $\hat{c}$ is positively biased: $\hat{E}(\hat{c})$ = 1.057 ± 0.002 (mean ± 1 SE), 2.210 ± 0.006, and 4.689 ± 0.020 for c = 1, 2, and 4, respectively. The relative bias increases with c (6, 10, and 17% for c = 1, 2, and 4, respectively). These values are based on all 243 cases, but there was only minor variation in $\hat{E}(\hat{c})$ across ϕ, p, u, and k. In general, the bias in $\hat{c}$ decreases as (1) ϕ increases, (2) p increases, and (3) u increases, but these decreases are generally small (<10%). Bias in $\hat{c}$ is unaffected by changes in k. Pooling of the data to compute the goodness-of-fit χ^2 test statistic may be a partial cause of the positive bias in $\hat{c}$. If no overdispersion was present (i.e., c = 1) in the data but one used a quasi-likelihood adjustment when it was, indeed, not needed (i.e., $\hat{c} \sim 1$), then this practice seems to have little effect on model selection (Table 1, compare the two columns under c = 1).

Model selection using uncorrected criteria

If no correction for overdispersion is made then AIC and AIC_c tend to select overfitted (i.e., $\hat{K} > K$) models and to select models with large RSS values when the data are overdispersed (Table 1, Fig. 1). If no overdispersion exists, these methods often select models with RSS near the best possible (MINRSS); however,

TABLE 1. Summary average of RSS (residual sum of squares) values for models H_0, H_{2p}, and $H_{3\phi}$ for capture–recapture data generated with overdispersion. Uncorrected (AIC, AIC_c and CAIC) and corrected (QAIC, $QAIC_c$, and QCAIC) are to be compared with the model with the minimum RSS, (MINRSS) for each repetition and the RSS under the true model (TRUERSS). CV ≤ 0.6% for all the table entries.

	$c = 1$		$c = 2$		$c = 4$	
	Uncorrected	Corrected	Uncorrected	Corrected	Uncorrected	Corrected
Model H_0 81 cases						
MINRSS	0.757	0.757	1.005	1.005	1.555	1.555
AIC	0.796	0.795	1.397	1.078	2.974	1.723
AIC_c	0.793	0.791	1.343	1.070	2.822	1.704
CAIC	0.762	0.762	1.022	1.013	1.772	1.589
TRUERSS	0.761	0.761	1.011	1.011	1.588	1.588
Model H_{2p} 81 cases						
MINRSS	0.799	0.799	1.083	1.083	1.670	1.697
AIC	0.877	0.874	1.545	1.220	3.245	1.965
AIC_c	0.874	0.871	1.485	1.212	3.108	1.953
CAIC	0.870	0.870	1.171	1.166	2.035	1.812
TRUERSS	0.844	0.844	1.192	1.192	1.975	1.975
Model $H_{3\phi}$ 81 cases						
MINRSS	0.828	0.828	1.131	1.131	1.788	1.788
AIC	0.935	0.932	1.679	1.333	3.542	2.143
AIC_c	0.931	0.927	1.618	1.321	3.385	2.113
CAIC	0.911	0.911	1.246	1.226	2.231	1.911
TRUERSS	0.937	0.937	1.430	1.430	2.542	2.542
All models 243 cases						
MINRSS	0.795	0.795	1.073	1.073	1.680	1.680
AIC	0.870	0.867	1.540	1.210	3.254	1.944
AIC_c	0.866	0.863	1.482	1.200	3.105	1.923
CAIC	0.847	0.847	1.146	1.135	2.013	1.771
TRUERSS	0.847	0.847	1.211	1.211	2.035	2.035

the presence of overdispersion in the data severely weakens the ability of these uncorrected methods to select a proper parsimonious model. The increase in RSS for $c = 4$ is striking (Table 1, compare lst two columns) and clearly some correction to these model selection criteria is needed to cope with overdispersed data. In addition, these uncorrected methods select the best model (i.e., that model producing the MINRSS) relatively infrequently with overdispersed data (Fig. 1, top). These factors cause the average RSS to be large as the models selected are not appropriately parsimonious. CAIC performs well with overdispersed data and selects models with relatively small RSS values. The large penalty term in CAIC results in the selection of more simple models ($\hat{K} < K$) than the MINRSS model (Fig. 1) and a good balance is achieved between underfitting and overfitting models in the presence of overdispersion.

Model selection using corrected criteria

Quasi-likelihood corrections (QAIC and $QAIC_c$) allow improved model selection and lower RSS values with overdispersed capture–recapture data (Table 1). As expected, QAIC and $QAIC_c$ are similar in performance. In general, QCAIC selects models with the lowest RSS (Table 1). Except for the null model (model H_0), QAIC, $QAIC_c$, and QCAIC tend to select models with a smaller RSS than if one knew, and used, the true model.

The quasi-likelihood corrected criteria QAIC and $QAIC_c$ not only select MINRSS models frequently with overdispersed data, but the models selected tend to have Shibata's (1989) balance between underfitting ($\hat{K} < K$) and overfitting ($\hat{K} > K$) models (Fig. 1). This balance in errors in selecting a parsimonious model is further illustrated in Fig. 2 for data generated under models H_0, H_{2p}, and $H_{3\phi}$. Here, the MINRSS model is frequently selected using QAIC or $QAIC_c$. In contrast, the use of QCAIC selects models with too few parameters compared to the MINRSS model and little balance in under and overfitting errors is achieved. Still, the RSS from QCAIC-selected models is less than for those selected by QAIC or $QAIC_c$ (Table 1).

The general pattern in the above results carries over when data analyses are partitioned by ϕ (0.5, 0.7, 0.9), p (0.4, 0.6, 0.8), u (50, 100, 300), k (5, 10, 20), or model (H_0, H_{2p}, $H_{3\phi}$). Factorial ANOVA was performed on the 243 cases using ϕ, p, u, k, and the true model as factors. The response variable was the RSS for a specific corrected criterion minus the MINRSS. This analysis attempts to answer the question, "what factors are associated with a difference between the model selected vs. the best model (i.e., MINRSS)?" Nearly every factor and interaction term was "significant" due to the large sample sizes (1000 replicates), thus the magnitude of the F values was used to rank the most important factors. Few patterns were evident when using the corrected criteria. QAICRSS-MINRSS was always influ-

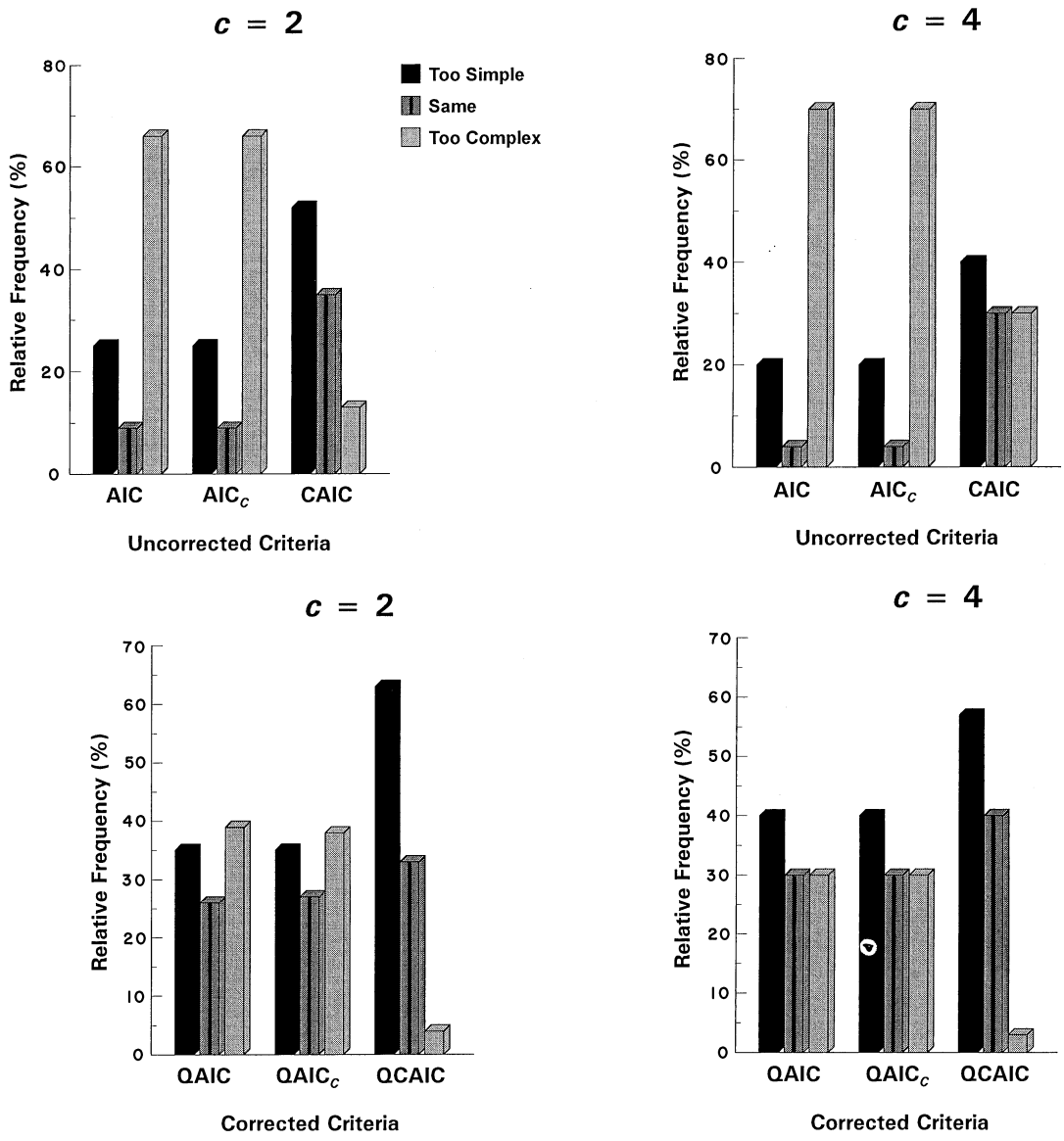


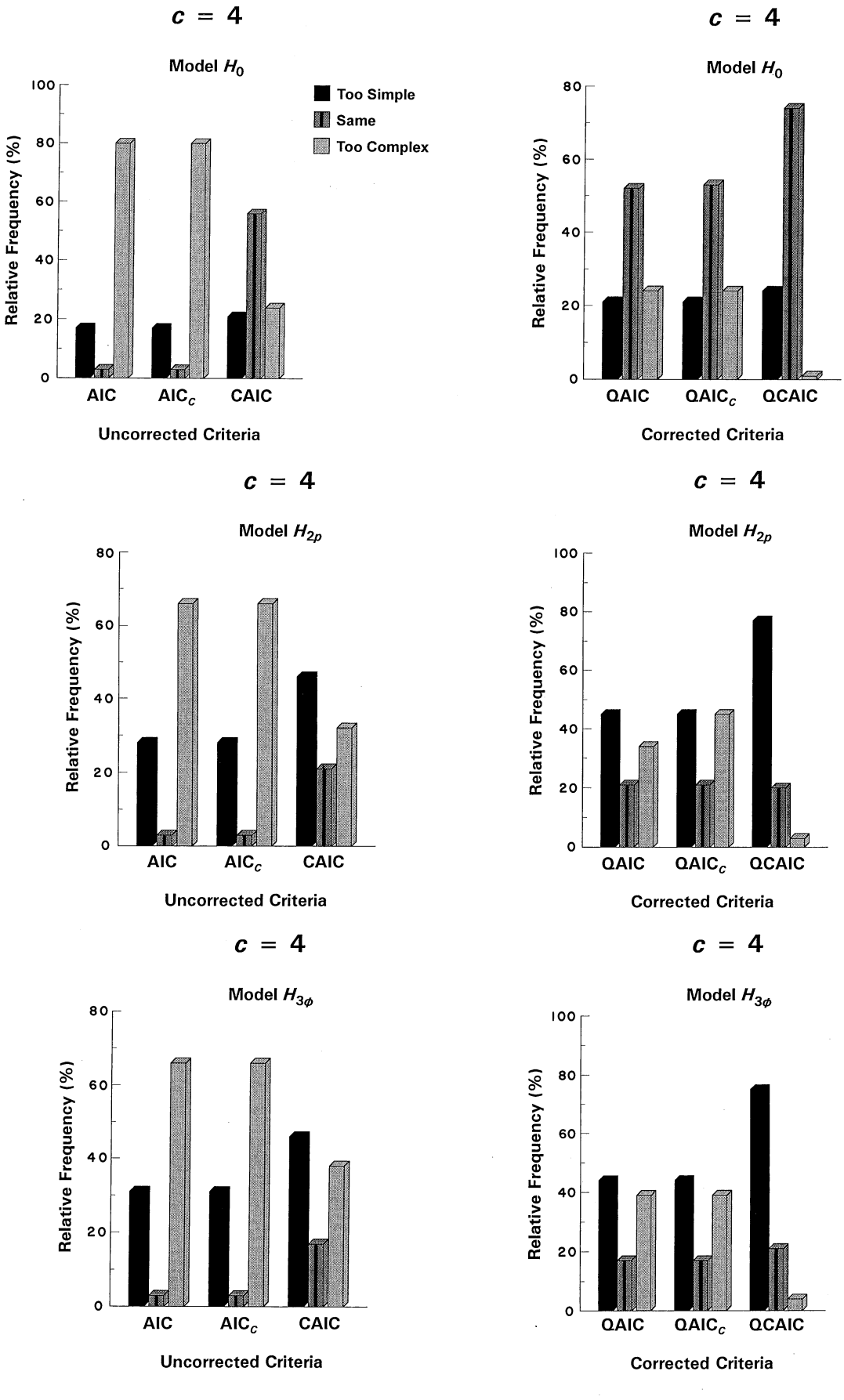
FIG. 1. Relative frequency of models selected by uncorrected and corrected information-theoretic criteria compared to the MINRSS model. Histograms at the left are for mild overdispersion ($c = 2$), while the figures to the right are for more severe overdispersion ($c = 4$). Categories shown as "Too Simple" are underfit models ($\hat{K} < K$), while "Too Complex" are overfit models ($\hat{K} > K$) when compared to the MINRSS model.

enced by ϕ, p, and u (in that order), but true model and interactions such as $\phi \times p$ and $\phi \times u$ appeared occasionally. Similar factors and interactions affected QCAICRSS-MINRSS, but even less pattern was revealed here.

It seems clear that in the analysis of overdispersed data, factors such as ϕ, p, u, and true model will often influence the degree to which these corrected criteria will achieve the selection of the best (MINRSS) model. At a larger perspective, QAIC and QAIC$_c$ do well in model selection with overdispersed capture–recapture data, while the uncorrected criteria do poorly. A complete understanding of the relative merits of QAIC and QAIC$_c$ vs. QCAIC remain unknown. QAIC and QAIC$_c$ achieve a balance between under- and overfitting, at the expense of a higher RSS when compared to QCAIC.

→

FIG. 2. Relative frequencies of models selected by uncorrected (left) and corrected (right) information-theoretic criteria compared to the MINRSS model for fairly severe overdispersion ($c = 4$). Information for models H_0, H_{2p}, and $H_{3\phi}$ is shown. Categories shown as "Too Simple" are underfit models ($\hat{K} < K$), while "Too Complex" are overfit models ($\hat{K} > K$) when compared to the MINRSS model.


c = 4
Model H_0
Too Simple
Same
Too Complex
Relative Frequency (%)
100
80
60
40
20
0
AIC
AIC_c
CAIC
Uncorrected Criteria
c = 4
Model H_0
Relative Frequency (%)
80
60
40
20
0
QAIC
$QAIC_c$
QCAIC
Corrected Criteria
c = 4
Model H_{2p}
Relative Frequency (%)
80
60
40
20
0
AIC
AIC_c
CAIC
Uncorrected Criteria
c = 4
Model H_{2p}
Relative Frequency (%)
100
80
60
40
20
0
QAIC
$QAIC_c$
QCAIC
Corrected Criteria
c = 4
Model $H_{3\phi}$
Relative Frequency (%)
80
60
40
20
0
AIC
AIC_c
CAIC
Uncorrected Criteria
c = 4
Model $H_{3\phi}$
Relative Frequency (%)
100
80
60
40
20
0
QAIC
$QAIC_c$
QCAIC
Corrected Criteria

At some point it will be necessary to investigate confidence interval coverage for models selected under these differing criteria. In some cases QAIC and $QAIC_c$ select the MINRSS model relatively frequently (Figs. 1 and 2).

Model $H_{9\phi}$

The results for model $H_{9\phi}$ are of interest because this model represents the case where there are many small treatment effects in addition to some larger effects. As sample size increases, one might expect to detect more of these minor effects. The results for model $H_{9\phi}$ with $c = 1$, 2, or 4 show that QAIC and $QAIC_c$ perform well over the range of ϕ, p, k, and u studied. These two criteria select a parsimonious fitted model with fewer (estimated) parameters than the true model and maintain a balance between under- and overfitting models when compared to the MINRSS model. Generally, QCAIC selected models with substantially fewer parameters and had smaller RSS values than did QAIC and $QAIC_c$ but, curiously, at $\phi = 0.9$ of $u = 300$ (i.e., large sample sizes) QCAIC performed poorly, producing large RSS values compared to QAIC and $QAIC_c$, and much larger RSS values than the MINRSS model.

Averaging over ϕ, p, and u, the RSS values for the true model (TRUERSS = 0.615, 1.394, and 3.267 for $c = 1$, 2, and 4, respectively) were substantially larger than the RSS values for QAIC (0.465, 0.895, and 1.759 for $c = 1$, 2, and 4, respectively), $QAIC_c$ (0.459, 0.879, and 1.737 for $c = 1$, 2, and 4, respectively) or QCAIC (0.389, 0.689, and 1.415 for $c = 1$, 2, and 4, respectively) selected models. This illustrates the advantages of a proper parsimonious model in data analysis (i.e., model fitting via parameter estimation) when the true model has many "effects" that are near zero.

Discussion

General

The motivation behind this paper is the issue of model selection in open model capture–recapture (CJS models) when there is overdispersion, relative to theory, in the data. Such overdispersion is a common occurrence. Before discussing specifics, however, there is a more general message we wish to convey: data-based model selection must be done in an objective manner. Model selection may be viewed as answering the question "how complex a model will the data support?" Essentially, the more data one has, the bigger the model may be and the selection of model "size" represents a trade-off of precision vs. bias in parameter estimation to achieve parsimony (see Sakamoto et al. 1986, Lehmann 1990, Burnham and Anderson 1992). These concepts have a long history in statistics (Linhart and Zucchini 1986, Lehmann 1990).

A key principle (essentially, this is parsimony) motivating model selection is that even if the true model structure is known, but parameters must be estimated from the data, use of the true model structure is likely to provide a relatively poor basis for statistical inference when compared to models selected by the information criteria. This result is particularly true and important when sample size is relatively small or when there are many relatively small effects (that must be estimated) in the true model.

Theory for statistical point estimation, confidence intervals, and hypothesis testing is well developed and accepted. It is as yet virtually unknown outside of the statistical discipline that there is sound theory for model selection when multiple alternative models (especially if they are all special cases of a global model) are fit to the same data. In particular, there is information theory based model selection (AIC), which has a deep foundation and may be considered as an extension of likelihood theory. The theory for AIC model selection dates back to only 1973 (Akaike 1973) and actual usage is only now starting to be widespread. Intensive investigations of performance and properties of AIC in ecological usage are in their infancy. Investigation of AIC model selection modified for overdispersion with multinomial count data is an unexplored subject.

We are aware that there is resistance to the idea of data-based model selection. Seventy years ago there was resistance to maximum likelihood fitting of a single model. Eventually it came to be accepted that analysis of one's data should consist of fitting (i.e., estimating the parameters of) a single model to the data by some well-defined optimality criterion (often maximum likelihood). Model "specification" (i.e., what model to use if there were options) was ignored (Lehmann 1990, Akaike 1994); one was just supposed to know the correct (hence "true" or "best") model to use. Initially in statistics the analysis paradigm of using only one model was promoted and was popular with users, because computing resources were very limited, often being confined to hand calculation with simple formulae. Given such a paradigm and no computers, there was little motivation to consider fitting many models, any or all of which required intense numerical calculations. Hence, there was no real motivation to extend theory to model selection as part of the data analysis. Statistical theory has been so extended in the past 20 yr (with more to come, we expect) and given the ubiquity of computers, computational difficulties are no longer an excuse to fit only a single, assumed-true model.

Resistance to the idea of model selection also exists when the only selection methods of which researchers are aware are poor. In particular, methods based on hypothesis testing, such as stepwise variable selection in regression, have terrible performance (Flack and Chang 1987). It is not the concept of model selection that is flawed but the selection methodologies: they have often been ad hoc, sometimes in the (inappropriate) spirit of "data mining." Not much better has

been the casting of model selection as an hypothesis testing problem with its inherent weaknesses: arbitrary choice of α level, need for nested models, asymmetry of null vs. alternative hypotheses, and the problems of making multiple tests, especially if the number of tests made is random.

Proper model selection must start with an objective criterion to be achieved, and then model selection becomes a problem in optimization. Viewed this way, model selection is just like parameter point estimation (as for example by maximum likelihood or least squares), a matter of optimizing a suitable objective function. Information theory provides an objective function; AIC implements the selection. There also needs to be a recognition that there is uncertainty in the selected "best" model, just as there is uncertainty in a point estimator. AIC-based model selection also has the potential to provide a measure of this model selection uncertainty (Kishino et al. 1991) in the spirit of confidence intervals.

We maintain that in complicated studies, such as long-term multiple capture–recapture data sets, it is a contradiction to think there is a true model (cf. Burnham and Anderson 1992:28); a model is a simplification of reality, hence will not reflect all the "truth" underlying a data set. Fundamental to this issue of model selection is that the size of the data set affects the "size" of the model (i.e., the number of parameters, of "effects") that should be used to represent the information in that data. Only through objective model selection can we let the data tell us what size of a model those data will support (Lehmann 1990; see also Burnham and Anderson 1992, Lebreton et al. 1992 and references therein).

Specific

The investigation of overdispersion here deals only with the lack of independence and not with parameter heterogeneity, although both statistical dependence and heterogeneity are probably common in the analysis of real capture–recapture data. The statistical sampling distribution used here for the CJS data is that of $k - 1$ independent multinomials for each data subset (this is the commonly used approach for CJS data, see Lebreton et al. 1992). The representation of overdispersion as just c times the theoretical sampling variance of each multinomial in the model is the simplest possible approach. However, any more general approach gets quickly much more complicated, for example, see O'Hara Hines and Lawless (1993) concerning overdispersion in multinomial distributions. There are a variety of models for the generalized multinomial variance–covariance matrix; different models correspond to different causes of overdispersion.

While it would be interesting to know the likely causes of overdispersion, the form of the model used may not matter much as regards getting a single estimated average overdispersion, $\hat{c}$, to use in modified AIC-based model selection. In a treatise on analysis of count data, Cox and Snell (1989) assert that the simple approach, such as we have studied here, should often be adequate, as opposed to the much more arduous task of seeking an explanatory model for the overdispersion. This comment of Cox and Snell is supported by the results of Liang and McCullagh (1993), who found that causal modeling of overdispersion was clearly better than use of a single overdispersion parameter, c, in only one of five cases examined.

The magnitude of the overdispersion as measured by c is relevant to this issue. The intent of fitting a model is to interpret the structural variation (patterns) in the data (McCullagh and Nelder 1989). In CJS models we have good reason to think the theoretical multinomial models have dispersion structure that is approximately correct, but a perfect match of theory and reality is too much to ask. Once one has found an adequate model structure, overdispersion, $\hat{c}$, seems often to be just above one to as much as three. Sophisticated modeling of overdispersion may well be unnecessary at these low levels of c. Conversely, if $\hat{c}$ is as big as 10 (and perhaps if it is as much as 5), or more, it is our opinion that important structural variation remains to be extracted from the data (i.e., the model selected is not structurally adequate). Our opinion on this matter is based on experience in capture–recapture and other areas (e.g., Eberhardt 1978, Burnham et al. 1980, Buckland et al. 1993).

Our first result was that the variance inflation factor c can be fairly well estimated using quasi-likelihood methods. The estimator $\hat{c}$ has a positive bias and the relative bias increases with the degree of overdispersion. If $\hat{c}$ is much above 1 it may be recommended that empirical variances and covariances be computed as $\hat{c}$ times the theoretical (i.e., model-based) variance and covariance estimates. Confidence intervals should be based on these inflated standard errors and the degrees of freedom must relate to the degrees of freedom associated with the estimation of c.

The most important results of this study about model selection for open capture–recapture models are (1) if there is overdispersion then unmodified AIC-based model selection performs very poorly. However, (2) simple corrections to AIC, AIC_c, and CAIC made based on quasi-likelihood principles are relatively effective in selecting a parsimonious model for the analysis of, and inference to, overdispersed data. These criteria are simple to compute and we recommend their use in the analysis of capture–recapture data. Because of the strong similarity between open capture–recapture models and band recovery models (Brownie et al. 1985), we suggest that these results are likely to apply in the analysis of band recovery data.

Further research might allow a less biased estimator of c to be derived. More important, perhaps, is that additional research might be done to evaluate the relative merits of QAIC and $QAIC_c$ vs. QCAIC. However,

because of the major conceptual difference in the objective underlying AIC and CAIC (and hence QAIC vs. QCAIC) we do not recommend CAIC (the use of which assumes a fixed true model independent of sample size). In contrast, AIC usage assumes that it is justified to estimate more parameters as sample size increases.

A very important subject for further research is the bias in standard errors and confidence interval coverage of the parameter estimators of the selected model. The usual procedure is to use the theoretical standard errors from the model selected and adjust these standard errors for overdispersion; this is sound as far as it goes. However, no adjustment is made for the uncertainty induced by model selection. It is not known how to account for the model selection uncertainty (in this frequentist framework), but research on the matter may lead to the needed methods.

ACKNOWLEDGMENTS

We thank the reviewers for their helpful comments on the submitted version of this paper and Douglas H. Johnson for diligent and helpful editorial and technical suggestions. The efforts of these people improved our paper. Two of us (D. R. Anderson and K. P. Burnham) thank the U.S. Fish and Wildlife Service for its years of support to us on this general research topic of capture–recapture.

LITERATURE CITED

Akaike, H. 1973. Information theory as an extension of the maximum likelihood principle. Pages 267–281 *in* B. N. Petrov and F. Csaki, editors. Second International Symposium on Information Theory. Akademiai Kaido, Budapest, Hungary.

———. 1981. Likelihood of a model and information criteria. Journal of Econometrics **16**:3–14.

———. 1985. Prediction and entropy. Pages 1–24 *in* A. C. Atkinson and S. E. Fienberg, editors. A celebration of statistics. Springer, New York, New York, USA.

———. 1994. Implications of the informational point of view on the development of statistical science. Pages 27–38 *in* H. Bozdogan, editor. Engineering and Scientific Applications. Volume 3. Proceedings of the First U.S./Japan Conference on the Frontiers of Statistical Modeling: An Informational Approach. Kluwer Academic, Dordrecht, The Netherlands.

Anderson, D. R., M. A. Wotawa, and E. A. Rexstad. 1993. Trends in the analysis of recovery and recapture data. Pages 373–386 *in* J.-D. Lebreton and P. M. North, editors. Marked individuals in the study of bird population. Birkhauser Verlag, Basel, Switzerland.

Armitage, P. 1957. Studies in the variability of pock counts. Journal of Hygiene **55**:564–581.

Atkinson, A. C. 1980. A note on the generalized information criterion for choice of a model. Biometrika **67**:413–418.

Bartlett, M. S. 1936. Some notes on insecticide tests in the laboratory and in the field. Journal of the Royal Statistical Society, Supplement **3**:185–194.

Bozdogan, H. 1987. Model selection and Akaike's information criterion (AIC): the general theory and its analytical extensions. Psychometrika **52**:345–370.

Brownie, C., D. R. Anderson, K. P. Burnham, and D. S. Robson. 1985. Statistical inference from band recovery data—a handbook. Second edition. U.S. Fish and Wildlife Service, Resource Publication Number **156**.

Buckland, S. T., D. R. Anderson, K. P. Burnham, and J. L. Laake. 1993. Distance sampling: estimating abundance of biological populations. Chapman and Hall, London, England.

Burnham, K. P. 1991. On a unified theory of release-resampling of animal populations. Pages 1–35 *in* M. T. Chao and P. E. Cheng, editors. Proceedings of the 1990 Taipei Symposium in Statistics. Institute of Statistical Science, Academia Sinica, Taipei, Taiwan.

———. 1993. A theory for combined analysis of ring recovery and recapture data. Pages 199–213 *in* J.-D. Lebreton and P. M. North, editors. Marked individuals in the study of bird population. Birkhauser Verlag, Basel, Switzerland.

Burnham, K. P., and D. R. Anderson. 1992. Data-based selection of an appropriate biological model: the key to modern data analysis. Pages 16–30 *in* D. R. McCullough and R. H. Barrett, editors. Wildlife 2001: populations. Elsevier, London, England.

Burnham, K. P., D. R. Anderson, and J. L. Laake. 1980. Estimation of density from line transect sampling of biological populations. Wildlife Monographs **72**.

Burnham, K. P., D. R. Anderson, and G. C. White. 1994. Evaluation of the Kullback-Leibler discrepancy for model selection in open population capture–recapture models. Biometrical Journal **36**:299–315.

Burnham, K. P., D. R. Anderson, G. C. White, C. Brownie, and K. H. Pollock. 1987. Design and analysis methods for fish survival experiments based on release-recapture. American Fisheries Society, Monograph **5**:1–437.

Burnham, K. P., G. C. White, and D. R. Anderson. *In press.* Model selection in the analysis of capture-recapture data. Biometrics.

Cormack, R. M. 1964. Estimates of survival from the sighting of marked animals. Biometrika **51**:429–438.

Cox, D. R., and E. J. Snell. 1989. Analysis of binary data. Second edition. Chapman and Hall, New York, New York, USA.

Eberhardt, L. L. 1978. Appraising variability in population studies. Journal of Wildlife Management **42**:207–238.

Finney, D. J. 1971. Probit analysis. Third edition. Cambridge University Press, London, England.

Fisher, R. A. 1949. A biological assay of tuberculins. Biometrics **5**:300–316.

Flack, V. F., and P. C. Chang. 1987. Frequency of selecting noise variables in subset regression analysis: a simulation study. American Statistician **41**:84–86.

Huggins, R. M. 1991. Some practical aspects of a conditional likelihood approach to capture-recapture experiments. Biometrics **47**:725–732.

Hurvich, C. M., and C. Tsai. 1989. Regression and time series model selection in small samples. Biometrika **76**: 297–307.

Jolly, G. M. 1965. Explicit estimates from capture-recapture data with both death and immigration—stochastic model. Biometrika **52**:225–247.

Kapur, J. N., and H. K. Kesavan. 1992. Entropy optimization principles with applications. Academic Press, London, England.

Kishino, H., H. Kato, F. Kasamatsu, and Y. Fujise. 1991. Detection of heterogeneity and estimation of population characteristics from the field survey data: 1987/88 Japanese feasibility study of the Southern Hemisphere minke whales. Annals of the Institute of Statistical Mathematics **43**:435–453.

Kullback, S. 1959. Information theory and statistics. John Wiley, New York, New York, USA.

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 Monograph **62**:67–118.

Lehmann, E. L. 1990. Model specification: the view of Fish-

er and Neyman, and later developments. Statistical Science **5**:160–168.
Liang, K-Y., and P. McCullagh. 1993. Case studies in binary dispersion. Biometrics 49:623–630.
Linhart, H., and W. Zucchini. 1986. Model selection. John Wiley & Sons, New York, New York, USA.
McCullagh, P., and J. A. Nelder. 1989. Generalized linear models. Second edition. Chapman and Hall, New York, New York, USA.
McCullagh, P., and D. Pregibon. 1985. Discussion comments on the paper by Diaconis and Efron. Annals of Statistics **13**:898–900.
Moore, D. F. 1987. Modelling the extraneous variance in the presence of extra-binomial variation. Journal of the Royal Statistical Society **36**:8–14.
O'Hara Hines, R. J., and J. F. Lawless. 1993. Modelling overdispersion in toxicological mortality data grouped over time. Biometrics **49**:107–121.
Pollock, K. H., J. D. Nichols, C. Brownie, and J. E. Hines. 1990. Statistical inference for capture-recapture experiments. Wildlife Monographs **107**.
Sakamoto, Y., M. Ishiguro, and G. Kitagawa. 1986. Akaike information criterion statistics. KTK Scientific, Tokyo, Japan.
Schwarz, M. 1978. Estimating the dimensions of a model. Annals of Statistics **6**:461–464.
Seber, G. A. F. 1965. A note on the multiple recapture census. Biometrika **52**:249–259.
———. 1982. The estimation of animal abundance and related parameters. Second edition. MacMillan, New York, New York, USA.
Shibata, R. 1976. Selection of the order of an autoregressive model by Akaike's information criterion. Biometrika **63**: 117–126.
———. 1989. Statistical aspects of model selection. Pages 215–240 *in* J. C. Willems, editor. From data to model. Springer-Verlag, London, England.
Stone, M. 1977. An asymptotic equivalence of choice of model by cross validation and Akaike's criterion. Journal of the Royal Statistical Society, **B 39**:44–47.
Sugiura, N. 1978. Further analysis of the data by Akaike's information criterion and the finite corrections. Communications in Statistics, Theory and Methods **A7**:13–26.
Wedderburn, R. W. M. 1974. Quasi-likelihood functions, generalized linear models, and the Gauss-Newton method. Biometrika **61**:439–447.
Williams, D. A. 1982. Extra-binomial variation in logistic linear models. Applied Statistics **31**:144–148.