GENERAL STRATEGIES FOR THE COLLECTION AND ANALYSIS OF RINGING DATA

David R. Anderson and Kenneth P. Burnham
Colorado Cooperative Fish and Wildlife Research Unit
Room 201 Wagar Building, Colorado State University
Fort Collins, Colorado USA 80523-1484

1. Introduction

Little has appeared on general *strategies* for either the collection or analysis of bird ringing data. Strategy is important as a biologist contemplates a comprehensive analysis of data, either exploratory or somewhat more confirmatory. Such considerations are particularly important in large-scale studies involving perhaps several study areas, objectives, and several investigators. Our general objective is to provide biologists and analysts with practical advice (i.e., a consistent strategy) for issues surrounding the collection and analysis of ringing data. The emphasis here is on bird ringing data where (1) encounters relate to dead birds (Brownie et al. 1985) and the main parameters of interest are conditional survival (S_j) and reporting (r_j) probabilities and (2) capture-recapture (C-R) studies of the Cormack-Jolly-Seber type (Lebreton et al. 1992) where the parameters of interest are the conditional apparent survival (ϕ_j) and recapture (p_j) probabilities.

2. Planning Prior to Data Collection

We first assume that an exciting biological question has been carefully posed. Secondly, we must assume the investigator(s) knows the biology of the species to be studied. If the study involves a formal experiment, then Burnham et al. (1987), Pollock et al. (1990), and Skalski and Robson (1992) provide theory, practical advice, and examples to guide nearly all design aspects to be considered. Program RELEASE (Burnham et al. 1987) is useful in various stages of planning such experiments.

It is more difficult to offer such detailed prescriptions if an observational study is called for (but see papers in Lebreton and North 1993 and volume 22, nos. 5 & 6 of the *Journal of Applied Statistics* and references therein). Adequate sample size of released cohorts is an important consideration and programs ESTIMATE (Brownie et al. 1985:153-154 and 183-193), RELEASE, and BAND2 (Wilson et al. 1989) provide options that will aid in determining approximate sample size required to attain a given level of precision. Statistical design formally considers trade-offs such as optimal allocation of ringing vs. recapture effort. While equal allocation of ringing effort across years is not terribly non-optimal, it is better to have increased effort in the early years of the study. This is often the opposite of what is typically done. Similar allocation issues arise in estimating age-specific survival probabilities in that more of the younger age classes should be ringed as they typically have lower survival than older birds. High probability of reporting (r_j) is important and it is often possible to model these parameters during the analysis, particularly if the species is hunted and hunting pressure varies across the years of the study.

In capture-recapture studies it is important to measure several covariates related to the ringing effort (e.g., number of person days, seasonal weather, number of trapping days). In both capture-recapture and recovery studies potential covariates related to survival (e.g., environmental conditions or habitat fragmentation) should be recorded each year, in addition to the more obvious variables such as age, sex and membership in other categories (e.g., breeding vs. nonbreeding). Measures of ringing effort can be incorporated directly into C-R models of the recapture probabilities, thus reducing the number of estimated parameters. Depending on the study objectives, covariates (either group covariates for a time or space variable or individual based) thought to affect survival should also be measured (e.g., winter severity or habitat change, or fragmentation). This might include the mass of individual birds, body condition or the parasite load of each bird as these can be used as an individual covariates, thus allowing the estimation of survival distributions (see Skalski et al. 1993, Smith et al. 1994 for further examples). Individual covariates that change over time (e.g., parasite load) are problematic because this variable is not observed in years when a particular bird is not encountered or when the bird was never encountered, following release.

Precision of estimated parameters increases as the number of released birds increase and as encounter probabilities (i.e., capture or recovery probabilities) increase. Thus, REWARD rings might be considered to increase ring recovery probabilities (Nichols et al. 1995). In general there is more information in increasing the encounter rate relative to increasing the numbers ringed (see Burnham et al. 1987: 315). Ringing studies are best when one achieves high encounter rates of each bird every year it is alive. These issues can be considered fairly easily with available software. Analysis methods assume that rings do not fall off or corrode such that the number becomes illegible.

Observational studies are most informative if conducted over long time frames (say, 20 or 30 years). The value of the data increases with the length of the study. Relatively little can be learned from only 3-4 years of ringing, except perhaps in an experiment involving an acute treatment effect. Large-scale studies require careful coordination, probably best done through a written protocol (Forsman et al. 1996). Of course, cause and effect can only safely be established from experimental studies; only correlational results can be inferred from observational studies. Often, little can be salvaged if data collection has been seriously flawed or if the question was poorly posed (Hand 1994). Field assistants that lack proper training and motivation risk ruining the study and should be avoided. We realize, of course, that these issues are never as ideal as one would like, however, proper attention must be placed on the collection of data (Chatfield 1991, 1995a).

An important issue it the degree to which the ringed birds represent the birds in the population of interest. Can inferences be made from the ringed birds to the species in a larger geographic area (i.e., to birds that were not ringed) or only to the ringed birds? This relates to the trapping methods used (e.g., baited traps), timing and site of ringing. Ringing at several sites might often be preferable to ringing at a single site in the hope that the population will be better represented by the ringed birds. Mallard ducks in eastern Colorado were banded during the winters of 1963-86 in 7 large geographic areas. This allowed some spatial “replication" as well as the ability to assess differences in the spatial areas (Wotawa 1993). While ringing sites are usually not selected at random, one must guard against ringing at sites having some unusual property that might make birds at the site nonrepresentative. It is not desirable to miss ringing in a year once the study has begun.

3. Analysis Philosophy

Our intention is to present an analysis strategy that treats model formulation, model selection, estimation of model parameters and their associated uncertainty in a unified manner, under a common framework (information and likelihood theory). The ideas here stem partially from Akaike's papers and our experience, but also consider the critique by Chatfield (1995b) and other similar literature. We stress inferences concerning understanding the structure and function of the ringed population, estimators of relevant parameters, and valid measures of precision. We assume the reader has a general familiarity with literature on methods and software for the analysis of ringing data.

The philosophy and theory presented here must rest on well designed studies and careful planning and execution of ringing protocol. Many good books exist giving information on these important issues (Burnham et al. 1987, Cook and Campbell 1979, Mead 1988, Hairston 1989, Desu and Roghavarao 1991, Eberhardt and Thomas 1991, Manly 1992, Skalski and Robson 1992, and Scheiner and Gurevitch 1993). In the following material we will assume that the data are “sound" and that inference to a larger population is justified by the manner in which the data were collected.

A philosophy of thoughtful, science-based, *a priori* modeling is advocated when possible. Science and biology play a lead role in this *a priori* model building and careful consideration of the problem. The modeling and careful thinking about the problem are critical elements that have often received relatively little attention in ornithology. Instead, there has often been a rush to “get to the data analysis" and begin to rummage through the ringing data and compute various quantities of interest. We advocate a deliberate, focused effort on *a priori* model building, as this tends to avoid “data dredging" which leads to over-fitted models (Freedman 1983) and finding spurious effects. We realize other, more liberal, philosophies may have their place, especially in more exploratory phases of investigation.

The size or dimension (K=number of parameters) of some ringing models can be quite high and this has increased substantially over the past two decades. Open capture-recapture and ring recovery models commonly have 30 – 60 estimable parameters for a single data set and might have well over 200 parameters for the joint analysis of several data sets (see Burnham et al. 1996). These are applications where objective modeling and model selection is essential to answer the question “*what inferences do the data support*?"

Careful, *a priori* consideration of alternative models will often require a change in emphasis among many biologists. This *a priori* strategy is in contrast to strategies advocated by others where they view modeling and data analysis as a highly iterative and interactive exercise. Such a strategy, to us, represents deliberate data dredging and should be reserved for early exploratory phases of an initial investigation.

4. Models and Data

Of necessity data analysis will need to be based on a model to represent the information in the data; these models involve parameters (e.g., S, ϕ, p, f, r, N, B). Omnibus methods have been developed to achieve valid inference from models that are good approximations to ringing data (e.g.,

various extensions of the Cormack-Jolly-Seber model – note Jolly (1965) anticipated many of these extensions some 33 years ago). A broad definition of ringing data is employed here; the data might be a single, simple data set as the subject of analysis. Increasingly, ringing data collected from several field sites might be the subject of a more comprehensive analysis (e.g., Burnham et al. 1996). These large-scale studies might commonly be extensive and partitioned by age, sex, species, treatment group, within several habitat types or geographic areas. There are often factors (variables) with large effects in these rich data sets as well as a myriad of smaller effects, both fixed and random. Parameters in the model represent these factors or effects.

Fisher's *Maximum Likelihood* method has been the primary, omnibus approach to parameter estimation in ringing studies (relatively few alternatives have been proposed), but it *assumes the model structure is known* and that only the parameters in that structural model are to be estimated. That is, if one assumes or somehow chooses a particular model structure, methods exist that are objective and asymptotically optimal for estimating model parameters and the model-based sampling covariance structure if the model is "correct." Given an appropriate model and if the sample size is "large" then maximum likelihood provides estimators of parameters (MLEs) that are consistent (i.e., asymptotically unbiased, variance tending to zero), fully efficient (i.e., minimum variance among consistent estimators), and normally distributed. In addition, profile likelihood intervals or log-based intervals can be used to achieve asymmetric confidence intervals with good coverage properties. In general, ML provides an objective, omnibus theory for estimation of model parameters and the sampling covariance matrix, *given an appropriate model*.

5. The Critical Issue: "What Model To Use?"

While hundreds of books and countless journal papers deal with estimation of model parameters and their associated precision, relatively little has appeared concerning model specification (what set of candidate models to consider) and model selection (what model(s) to use for inference) (see Peirce 1955). In fact, R. A. Fisher believed at one time that model specification was outside the field of mathematical statistics and this attitude prevailed within the statistical community for several decades until, at least, the early 1970s. "*What model to use?"* is *the* critical question in making valid inference from data in the biological sciences.

If a poor or inappropriate model is used, then inference based on the data and this model will often be poor. Thus, it is clearly important to select an appropriate model for the analysis of a specific data set; however, this is not the same as trying to find truth (i.e., the "true model" – there is no *true* model). Model selection methods with a deep level of theoretical support are required and, particularly, methods that are easy to use and widely applicable in practice.

6. Science Inputs – Formulation of Set of Candidate Models

Model formulation is the point where the scientific and biological information formally enter the investigation. The published literature, field experience, and the biological question posed must be used to formulate a set of *a priori* candidate models. Good approximating models in conjunction

with a good set of relevant data can provide insight into the underlying population process of interest. Starting with a good global model will help protect the analyst from selecting a bad fitting model.

Interesting biological questions come from the scientific literature, prior results of manipulative experiments, personal experience while taking the field data, or contemporary debate within the scientific community. More practical questions stem from management controversies, biomonitoring programs, quasi-experiments, and even judicial hearings. Such questions are represented as models, or differences between models.

Development of the *a priori* list of candidate models should include a global model; a model that has many parameters, includes all relevant effects and reflects causal mechanisms thought likely, based on the science of the situation. Specification of the global model should not be based on a probing examination of the data to be analyzed. Models with fewer parameters can then be derived as special cases of the global model. This set of reduced models represent plausible hypotheses based on what is known or hypothesized about the process under study. Generally, alternative models will involve differing numbers of parameters; often the number of parameters can differ by at least an order of magnitude across the candidate models. Chatfield (1995b) writes concerning the importance of subject-matter considerations such as accepted theory, expert background knowledge, and prior information in addition to known constraints on both the model parameters and the variables in the models. All these factors should be brought to bear on the makeup of the set of candidate models, prior to actual data analysis.

The more parameters used, the better the fit that can be achieved to the particular data at hand. Large and extensive data sets are likely to support more complexity (larger K) and this should be considered in the development of the list of candidate models. If a particular model structure or parameterization does not make biological sense, it should *not* be included in the list of candidate models. In developing the list of candidate models, one must recognize a certain balance between keeping the list small and focused on plausible hypotheses, while making it big enough to guard against omitting a very good, *a priori* model. While this balance should be considered, we advise the inclusion of all models that seem to have a reasonable justification, prior to data analysis. While one must worry about errors due to both under-fitting and over-fitting, it seems clear that over-fitting is less damaging than under-fitting (Shibata 1989). Here we judge under- and over-fitting relative to a best approximating model – an optimal trade-off between bias and variance. Then under-fitting is failing to recover information that is supported (as an inference to the population) by the data, while over-fitting is making an inference to the population about a feature of the sample that is quite unique to that sample. Under-fitting leads to a model with too few parameters and an overestimation of precision, while confidence interval coverage is likely to be well below the nominal level. Over-fitting leads to finding spurious effects and an unnecessary loss of precision.

Here, we advocate the deliberate exercise of carefully developing a set of, say, 4-20 alternative models as potential approximations to the data available and the scientific questions being addressed. Some practical ringing studies might have as many as 70-100 or more models that one might want to consider, but we hope these are exceptions rather than the rule. This set of models, developed without first deeply examining the data constitute the set of “candidate models."

Chatfield (1995b) suggests there is a need for more *careful thinking* (than is usually evident) and a *better balance* between the problem (biological question), analysis theory, and data. Too often, the emphasis is focused on the analysis theory and data available, with too little thought about the reason for the study in the first place (see Hayne 1978 for convincing examples). The set of models should be understandable in terms of study objectives and design and any model structure and its parameters should be interpretable.

7. Model Selection

After several years of investigating methods for use in selecting a good approximating model for capture-recapture and ring recovery data, we have concluded that only the information theoretic methods have a deep theoretical foundation and philosophical basis, while being generally applicable and easy to use in the analysis of ringing data. In particular, Akaike's Information Criterion (AIC) (Akaike 1973, 1974, 1977, 1978, 1981a and b, 1983, 1985, 1992, 1994, Bozdogan 1987, and deLeeuw 1992) is recommended here:

$$\text{AIC} = \quad -2\log(\mathcal{L}(\hat{\theta}) \mid x) + 2K,$$

where $\log(\mathcal{L}(\hat{\theta}) \mid x)$ is the value of the maximized log-likelihood function and K is the number of estimable parameters in the model. AIC can only select the best model from the *a priori* set; if good models are not in this set, they will remain out of consideration.

If K is large with respect to the sample size (n), then a second order version (Sugiura 1978, Hurvich and Tsai 1989, 1991 1994, 1995) is useful (when, say $n/K < 40$),

$$\text{AIC}_c = \quad -2\log(\mathcal{L}(\hat{\theta})) + 2K + \frac{2K(K+1)}{n-K-1},$$

alternatively,

$$\text{AIC}_c = -2\log(\mathcal{L}(\hat{\theta})) + 2K\left(\frac{n}{n-K-1}\right).$$

At some point, one should investigate the fit of the global model to the data (e.g., formal χ^2 goodness-of-fit tests). If the fit is adequate (say a nonsignificant χ^2 at $\alpha = 0.15$) is seems safe to proceed. If the fit is poor, then a variance inflation factor $\hat{c}$ can be computed from the global model as χ^2_{gof}/df (Wedderburn 1974). If $\hat{c}$ is less than, say, 3-4 one might conclude there is some effect of heterogeneity, a slight violation of independence, or some small failure of the structural component of the model and proceed to use QAIC or QAIC_c (below) for model selection. If $\hat{c} > 4$ or 5 one might want to reconsider the global model and what was thought to be understood about the problem at hand. Possibly the data are in error or the structure of the model has not been interpreted correctly by the computer software being used. At this point the analysis may take a more exploratory nature.

Many sets of ringing data exhibit overdispersion (Eberhardt 1978, Burnham et al. 1987) or some structural failure that cannot be easily modeled and in these cases, model selection should be based on criteria that account for this,

$$\text{QAIC} = \quad - \left[2 \log(\mathcal{L}(\hat{\theta})) / \hat{c} \right] + 2K$$

or

$$\text{QAIC}_c = \; - \left[2 \log(\mathcal{L}(\hat{\theta})) / \hat{c} \right] + 2K \; + \; \frac{2K(K+1)}{n-K-1} \; ,$$

where $\hat{c}$ is an estimated variance inflation factor. When no overdispersion exists, c = 1, then the formulae for QAIC and QAIC_c reduce to AIC and AIC_c, respectively (see Anderson et al. 1994).

AIC, AIC_c and QAIC_c are estimates of the Kullback-Leibler (K-L) distance or discrepancy, a fundamental quantity in information theory (Kullback and Leibler 1951, Kullback 1959). In the case of iid observations, AIC is equivalent to the computationally intensive cross validation method (see Stone 1977). There are other estimators of the Kullback-Leibler distance, but these are more complicated (see Takeuchi 1976 and reviews by Shibata 199_). We warn against the use of the so-called dimension-consistent methods (see Anderson and Burnham 1999). In general, we are not denying the value of advanced Bayesian methods, only to note their computational complexity and the fact that these methods remain relatively unknown to biologists.

The use of AIC has had many practical and theoretical advantages and we have continued to explore this general issue in the context of ring recovery and capture-recapture data (Burnham and Anderson 1992, Anderson et al. 1994, Burnham et al. 1994, 1995a and b, Anderson et al. in press). However, Akaike's paradigm goes beyond merely the computation and interpretation of AIC to select a parsimonious model for inference from empirical data. Akaike suggested increased attention on a variety of considerations and modeling *prior* to the actual analysis of data. These considerations have refocused on the need for careful planning prior to data collection and the reasons for the study in the first place.

A best approximating model is achieved by properly balancing the errors of under-fitting and over-fitting (see Linhart and Zucchini 1986). Stone and Brooks (1990) comment on "... straddling pitfalls of underfitting and overfitting" – the Principle of Parsimony. The proper balance is the trade-off such that bias and variance are controlled to achieve confidence interval coverage at approximately the nominal level. Proper model selection rejects a model that is far from reality and attempts to identify a model in which the error of approximation and the error due to random fluctuations are well balanced (Shibata 1983, 1989).

After a carefully defined set of candidate models has been developed, one is left with the evidence in the data; the task of the analyst is to interpret this evidence from analyzing the data. Questions such as "what effects are supported by the data?" can be answered objectively. This modeling approach allows a clear place for experience (i.e., prior knowledge and beliefs), the results of past studies, the biological literature and current hypotheses to formally enter the modeling process.

8. Model Selection Uncertainty

Only in recent years has it been widely realized that the estimated standard errors from the selected model are frequently too small (i.e., the precision has been overestimated). This is because the uncertainty due to model selection has not been incorporated as a variance component in measuring precision. Often, there is substantial uncertainty in the best approximating model to select for inference. Practical methods for dealing with model selection uncertainty are at the state of the science and a full review of this subject here would take us too far astray. Instead, we mention two approaches that provide insight into the amount of uncertainty in model selection. In the first approach, differences in AIC are computed as Δ_i = AIC(i) – AIC(min), over all candidate models. This simple rescaling allows one to rank the candidate models. In fact, the Δ-AIC values provide a ratio scale, a true zero relative to the best model in the set. These values are estimates of the K-L distance between the best model and other models in the set and allow a quantitative measure of model plausibility. Preliminary research suggests that for models having Δ_i < about 2-3, these models are nearly tied and inference should be based on that subset of models. Models with Δ_i values > about 7-10 are relatively poor and should not receive any substantial consideration. Models with Δ_i values in the range of about 3-7 have some support.

The second approach goes further in that models can be calibrated to provide the relative plausibility via normalized AIC weights computed as

$$w_i = \{\exp(-\Delta_i/2)\}/\sum\{\exp(-\Delta_i/2)\}.$$

Because –AIC/2 is an approximately (the approximation gets better if there are some good models in the set) asymptotically unbiased estimate of the expected log-likelihood of the model, then exponentiating $-\Delta/2$ is an estimate of the relative, expected likelihood of each model. This calibration goes beyond just ranking the models according to their K-L distances from the best model.

In some cases, a particular parameter occurs in all models; this is often the case when prediction is of interest. Here, an analytic expression (Buckland et al. 1997) allows unconditional estimates of sampling variance of that parameter, using,

$$\hat{\text{var}}(\hat{\bar{\theta}}) = \left(\sum w_i \sqrt{\hat{\text{var}}(\hat{\theta}_i \mid \theta) + (\hat{\theta}_i - \hat{\bar{\theta}})^2} \right)^2,$$

where the first term under the radical is the MLE of the sampling variance, conditional on the model, the second term represents the variance in the parameter of interest, across the models, and $\hat{\bar{\theta}} = \sum w_i \hat{\theta}_i$. This unconditional variance can be used in setting confidence limits on either $\hat{\theta}$ or the model-averaged $\hat{\bar{\theta}}$. This method also has applicability in cases where a specific parameter (i.e., S_4,

the conditional survival in year 4) may not appear explicitly in each model in the set of candidates, however, $\hat{S}_4$ can be gotten from a model that assumes conditional survival is constant across years.

The third approach is to drawn a large number (at least 1,000) bootstrap samples from the ringing data, analyze each of these samples using the full list of candidate models, use AIC to select the best model for each bootstrap sample, and store the MLEs for the parameters. Averaging the estimates from each bootstrap sample provides the bootstrap estimate and confidence intervals can be set by using the lower and upper percentiles of the ordered, individual bootstrap estimates for each parameter. This approach, while computer intensive, allows model selection uncertainty to be directly incorporated into the inference. Shibata (199_) provides extended bootstrap methods in model selection.

9. Models vs. Full Reality

We believe that "truth" in the biological sciences has essentially infinite dimension and that full reality cannot be revealed with only finite samples of data and a "model" of those data. Bird populations are complex with many small effects (e.g., individual heterogeneity) and interactions; we can only hope to identify a model that provides a good *approximation* to the data available (Anderson and Burnham 199_). Full truth (reality) is elusive! Proper modeling tells what the data support, not what full reality might be (White et al. 1982:14-15). Increased sample size (information) allows us to chase full reality, but never quite catch it.

10. Analysis

Many ringing studies will address models including age (a), sex (s), location/site (l), years (t), or treatment/control group (v) and this might lead to consideration of models with some interactions ($t*a$) or linear submodels without interactions ($t+a$) (see Lebreton et al. 1992 for other examples). Software such as MARK (White and Burnham 199_), POPAN5 (Arnason and Schwarz 199_), SURGE (Pradel and Lebreton 1991), and SURPH (Smith et al. 1994) allow such models to be fit easily. Models with age-specific parameters are relatively difficult, especially if there are more than 2 age classes. The analysis of ringed birds where individual covariates have been taken is a relatively new an exciting methodology and should be given careful attention when initiating a ringing study.

In Lebreton et al. (1992) we suggested first modeling and testing to arrive at a parsimonious structure for the sampling probabilities and secondly, using an appropriate submodel for these, turning to address a parsimonious structure for the survival probabilities. We (DRA and KPB) no longer recommend this two stage approach nor the use of formal hypothesis tests to arrive at a parsimonious model for inference.

The study of conditional survival and sampling probabilities as functions of covariates is increasingly important. Here, it is tempting to obtain MLEs of annual survival probabilities and then regress these estimates against, say, an annual measure of winter severity as a separate analysis. This is poor procedure because the estimates often have substantial sampling correlations and unequal sampling variances, making the separate ordinary least squares approach invalid. If covariates are available, they should be incorporated into the log-likelihood as a submodel or, perhaps, treated in a

random effects context (Burnham 199_). Computer software allows these models to be easily built and analyzed (e.g., MARK, SURPH and SURGE).

Some overdispersion may be common in ringing data, caused by ringing brood mates, pairs at the nest, or members of social groups. This lack of independence causes the estimated sampling variances and covariances to be underestimated (i.e., the actual variability is larger than that estimated). The simple variance inflation approach mentioned in section 7 (see Burnham et al. 1987:243-246) is useful. The sampling variances and covariances of the parameter estimates from the selected model are inflated by $\hat{c}$. Generally, $1 \leq \hat{c} \leq 3$ for many sets of ringing data that we have seen (see Anderson et al. 1994). A $\hat{c} > 1$ should be supported by some biological reason for overdispersion (e.g., banding of brood mates); otherwise the value may reflect some structural inadequacy in the global model. Even when structural inadequacy is present, it may be appropriate to use a variance inflation factor to reflect the added uncertainty when it is not possible to model this residual structure.

Sometimes there exists special interest or controversy over a particular effect or factor. For example, this might be an effect on survival caused by the attachment of a radio transmitter. Thus, in the simplest case, birds in a control group are ringed and released simultaneously with the ringed treatment group, each with a radio transmitter attached. We denote this special effect as ψ. Traditionally one might try to "test" for the "significance" of such a survival effect, ψ. We recommend, instead, including models with and without this effect in the set of candidate models. If the (best) selected model does not include ψ, then some inference about ψ can be gotten from the model that includes this effect, but is otherwise the same as the selected model, by examining its estimate ($\hat{\psi}$) and its estimated standard error. Here, it is likely that the effect size will be relatively small and its estimated standard error will be relatively large, although perhaps the result will not be so clear. If the selected model includes ψ, then $\hat{\psi}$ and the unconditional $\widehat{se}(\hat{\psi})$ can be examined and inferences made concerning the effect or factor of interest.

It is important to perform comprehensive analyses when ringing data reflect multiple factors such as sites, gender, age, or breeding status. Such data sets should not be addressed piecemeal, but rather as a single, coherent analysis. Here an overall parsimonious model using link functions to allow parallelism in sex effects over years will often provide deep insights into the process of interest (Catchpole et al. 1995, Brownie et al. 1993, Nichols et al. 1993). Software such as MARK, SURPH and SURGE allow easy model building and model selection.

11. Data Dredging

The process of analyzing data with few or no *a priori* questions, by subjectively and iteratively searching the data for patterns and "significance," is often called by the derogatory term of "data dredging." Here the data are "submitted for analysis" in the hope that the computer and several hypothesis test results will provide information on "what is significant?". We believe that such a dredged model is probably over-fitted and unstable (highly variable results obtained over different replicate samples) (Freedman 1983).

Examples of data dredging include the examination of annual estimates of conditional survival probabilities and noting that two are quite less than the others. Then one begins to search (fish) for possible explanations and incorporate these into new models, after having noticed the pattern in the estimates. Similar data-dependent *post hoc* considerations can often suggest linear or nonlinear relationships and interactions and, therefore, lead the investigator to consider additional models. These activities should be avoided, at least until after the *a priori* considerations have been put in place, as they probably lead to over-fitted models with spurious effects and variables. This type of data-dependent, exploratory data analysis might have a place in the earliest stages of understanding a scientific issue and should probably remain unpublished. We can only recommend that the results of such approaches be treated as possible hypotheses, new data collected to address these effectively, and then submitted for a comprehensive and largely *a priori* strategy of analysis such as we advocate here. To a large degree, data dredging represents poor science.

Two types of data dredging might be distinguished. The first is that described above; a highly interactive, data dependent, iterative *post hoc* approach where the analyst makes subjective choices at each iteration or branching point in the process. Here, there seems little hope to be able to quantify the precision of the estimates or objectively judge model structure. The second is also common and also leads to likely over-fitting. In this second type, the investigator also has little (or does not chose to pursue) *a priori* information, thus "all possible models" are considered as candidates. For many ringing studies, the number of candidate models in this approach can be large. At least this second type is not explicitly investigator-dependent, whereby the investigator is making subjective decisions based on prior analysis results. Also, it is usually a one pass strategy; rather than taking the results of one set of analyses and inputting some of these results into the consideration of new models. In some applications, computer software often can systematically search all such models nearly automatically and, thus, the strategy of trying all possible models (or, at least, a very large number of models) continues to be popular. At least, if this approach is combined with estimates of unconditional sampling variances, and perhaps model averaging, one might expect to obtain valid estimates of precision. Still, we wonder if many inferential situations might be substantially improved if the researcher tried harder to focus on the science of the situation before proceeding with such a blind approach.

We certainly encourage people to understand their data and the science question of interest. We advocate some examination of the data prior to the formal analysis to detect obvious outliers and outright errors. One might examine the goodness of fit tests from a carefully-chosen global model to look for overdispersion (Lebreton et al. 1992, Anderson et al. 1994). However, if a particular pattern is noticed while examining the residuals and this leads to including another variable, then we would suggest caution concerning data dredging and its likely outcome. Often, there can be a fine line between a largely *a priori* approach and some degree of data dredging.

12. Hypothesis Tests

Over the past 20 years, modern statistical practice has been moving away from traditional methodologies based on formal statistical hypothesis testing (Yoccoz 1991, Bozdogan 1994, Johnson 1995, Stewart-Oaten 1995, Nester 1996). The historic emphasis on statistical hypothesis

testing will continue to diminish in the years ahead (e.g., see Bozdogan 1994), with increasing emphasis on estimation of effect sizes and associated confidence intervals (Graybill and Iyer 1994:35). Thorny theoretical issues arise when testing hypotheses in the presence of nuisance parameters; and this is of particular concern in ringing studies where the sampling probabilities (the p_j, f_j or r_j) are an integral component of the models. In particular, hypothesis testing for model selection is often poor (Akaike 1981b) and will surely diminish in the years ahead.

The hypothesis testing approach has a variety of problems, many of which are well known, but often ignored in the practical analysis of ringing data. These include the arbitrary α levels (e.g., 0.05 level), multiple testing problems, various tests within a data set have differing power, interpretation of the observed significance levels (there may be several dozen of these *P*-values), and the fact that likelihood ratio tests between nonnested models do not exist. Ideally, the α-level should be a function of sample size, the covariance matrix, and the number of tests to be performed. Tests of various hypotheses within a single data set are not independent, making inference difficult. The order of testing is arbitrary and differing test order (e.g., stepup vs. stepdown) will often lead to different final models. Likelihood ratio tests often employ a model as the null hypothesis in one test and the same model as the alternative hypothesis in another test; this raises the issue of the distribution of the test statistic under the ever changing "null." Finally, there seems to be no theory saying how the results of such tests are to be used to build a parsimonious model with good inferential properties; instead, there are only *ad hoc* rules with little or no theoretical basis.

Often, effort is spent testing hypotheses that are *obviously* false (see Johnson 1995). For example, testing the hypothesis that conditional survival probabilities over 18 years are equal (i.e., $S_1 = S_2 = \cdots = S_{18}$) seems pointless; the hypothesis is obviously false, so why test it? Similarly, formal tests for differences in conditional survival or sampling probabilities by age or sex seem misguided (e.g., surely survival differs between young and adult birds over a period of several years!). The central issues here are twofold: first, one must know if the differences are large enough to justify inclusion in a model to be used for inference (this is a *model selection problem*) and, secondly, one wants to know the *magnitude* of the difference (the "effect size" – e.g., trivial, small, medium, large) and a valid measure of precision (this is an *estimation problem*). These central issues are not ones of hypothesis testing for many problems in ringing studies. To our knowledge, none of the testing-based selection methods allow a ranking of the candidate models nor a calibration to selection uncertainty; these issues pose severe limitations to inference from ringing data.

13. Summary

We hope the material here does not appear to be too dogmatic or idealized. We have tried to synthesize concepts that we believe are important and incorporate these as recommendations or general advice. We realize there are other approaches; some people may still wish to test hypotheses to build models of empirical data, and that others may have a more lenient feeling towards data dredging than we advocate here. However, we are compelled by careful planning of the ringing study, the *a priori* approach of building candidate models, the use of information theoretic criteria for selecting a best approximating model, the use of likelihood theory for deriving parameter estimators, and incorporating model selection uncertainty into the statistical inferences.

Information theoretic methods that are estimates of Kullback-Leibler information (1) identify an estimated best approximating model from the set of *a priori* models considered, (2) provide a ranking of the candidate models, (3) allow the means to quantify model selection uncertainty and (4) allow estimates of unconditional precision. Such information allows a relatively deep level of insight and tends to avoid trying to pretend that issues are “black or white" when, in fact, they are more complex that such a simple classification.

Computational restrictions prevented biologists from evaluating alternative models until the past two decades or so. Thus, people tended to use an available model, without careful consideration of alternatives. Present computer hardware (Pentium processors) and software (MARK, POPAN5, RELEASE, SURPH and SURGE) make it possible to consider a number of alternative models as an integral component of comprehensive data analysis.

Perhaps we cannot totally overcome problems in estimating precision, following a data-dependent selection method such as AIC (e.g., see Dijkstra 1988). This limitation certainly warrants exploration as model selection uncertainty is a quite difficult area of inference. However, we must also consider the “cost" of *not* selecting a good parsimonious model for the analysis of a particular data set. That is, a model is just somehow “picked" independent of the data and used to approximate the data as a basis for inference. Then, one does not know (or care?) if it fits, over-fits, under-fits, etc. This naive strategy certainly will incur substantial costs. Alternatively, one might be tempted into an iterative, highly interactive strategy of data analysis (unadulterated data dredging) – again there are substantial costs in this approach (e.g., subjectivity and lack of valid measures of precision). The cost of not selecting a reasonably parsimonious model for data analysis is rarely considered.

14. Literature Cited

Akaike, H. (1973). Information theory as an extension of the maximum likelihood principle. Pages 267-281 in *Second International Symposium on Information Theory*. B. N. Petrov and F. Csaki, (eds.). Akademiai Kiado, Budapest.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control AC* **19**,716-723.

Akaike, H. (1977). On entropy maximization principle. Pages 27-41 *in* P. R. Krishnaiah (editor). *Applications of Statistics*. North-Holland, Amsterdam.

Akaike, H. (1978). A Bayesian analysis of the minimum AIC procedure. *Annals of the Institute of Statistical Mathematics* **30**, 9-14.

Akaike, H. (1981a). Likelihood of a model and information criteria, *Journal of Econometrics* **16**, 3-14.

Akaike, H. (1981b). Modern development of statistical methods. Pages 169-184 *in* P. Eykhoff (editor). *Trends and progress in system identification*. Pergamon Press, Paris.

Akaike, H. (1983). Statistical inference and measurement of entropy. Pages 165-189 In *Scientific inference, data analysis, and robustness*. G. E. P. Box, T. Leonard, and C-F. Wu (eds.) Academic Press, London.

Akaike, H. (1985). Prediction and entropy. In *A celebration of statistics*. Ed. A. C. Atkinson and S. E. Fienberg, pp. 1-24. Springer, New York.

Akaike, H. (1992). Information theory and an extension of the maximum likelihood principle. Pages 610-624. *in* S. Kotz and N. L. Johnson (editors). *Breakthroughs in statistics*, Vol. 1. Springer-Verlag, London.

Akaike, H. (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, Vol. 3, Proceedings of the First US/Japan Conference on the Frontiers of Statistical Modeling: An Informational Approach. Kluwer Academic Publishers, Dordrecht, Netherlands.

Anderson, D. R., and Burnham, K. P. (199_). Understanding information criteria for selection among capture-recapture or ring recovery models. EURING97.

Anderson, D. R., Burnham, K. P., and White, G. C. (1994). AIC model selection in overdispersed capture-recapture data. *Ecology* **75**, 1780-1793.

Anderson, D. R., Burnham, K. P., and White, G. C. (1998). Comparison of AIC and CAIC for model selection and statistical inference from capture-recapture studies. *Journal of Applied Statistics*25:263-282.

Arnason, N., and Schwarz, C. (199_). Using POPAN to analyze banding data. EURING97.

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

Bozdogan, H. (1994). Editor's general preface. Pages ix-xii *in* H. Bozdogan, editor. Engineering and Scientific Applications, Vol. 3, Proceedings of the First US/Japan Conference on the Frontiers of Statistical Modeling: An Informational Approach. Kluwer Academic Publishers, Dordrecht, Netherlands.

Brownie, C., Anderson, D. R., Burnham, K. P., and Robson, D. S. (1985). *Statistical inference from band recovery data – a handbook*. 2nd Ed. U. S. Fish and Wildl. Serv., Resour. Publ. 156. 305pp.

Brownie, C, Hines, J. E., Nichols, J. D., Pollock, K. H., and Hestbeck, J. B. (1993). Capture-recapture studies for multiple strata including non-Markovian transitions. *Biometrics* **49**, 1173-1187.

Buckland, S. T., Burnham, K. P., and Augustin, N. H. (1997). Model selection: an integral part of inference. *Biometrics* **53**, 603-618.

Burnham, K. P. (1999). Random effects models in ringing and capture-recapture studies. EURING97.

Burnham, K. P., and Anderson, D. R. (1992). Data-based selection of an appropriate biological model: the key to modern data analysis. In *Wildlife* 2001*: Populations*. Ed. D. R. McCullough and R. H. Barrett, pp. 16-30. London, Elsevier Sci. Publ., Ltd.

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

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

Burnham, K. P., White, G. C., and Anderson, D. R. (1995a). Model selection in the analysis of capture-recapture data. *Biometrics* **51**, 888-898.

Burnham, K. P., Anderson, D. R., and White, G. C. (1995b). Selection among open populations capture-recapture models when capture probabilities are heterogeneous. *Journal of Applied Statistics* **22**, 611-624.

Burnham, K. P., Anderson, D. R., and White, G. C. (1996). Meta-analysis of vital rates of the Northern Spotted Owl. *Studies in Avian Biology* **17**, 92-101.

Catchpole, E. A., Freeman, S. N., and Morgan, B. J. T. (1995). Modelling age variation in survival and reporting rates for recovery models. *Journal of Applied Statistics* **22**, 597-609.

Chatfield, C. (1991). Avoiding statistical pitfalls (with discussion). *Statistical Science* **6**, 240-268.

Chatfield, C. (1995a). *Problem Solving: A Statistician's Guide*. Chapman and Hall, London. 325 pp.

Chatfield, C. (1995b). Model uncertainty, data mining and statistical inference. *Journal of the Royal Statistical Society, Series* A. **158**, 419-466.

Cook, T. D., and Campbell, D. T. (1979). *Quasi-experimentation: design and analysis issues for field settings*. Houghton Mifflin Co., Boston.

deLeeuw, J. (1992). Introduction to Akaike (1973) information theory and an extension of the maximum likelihood principle. Pages 599-609 *in* S. Kotz and N. L. Johnson (editors). *Breakthroughs in statistics*, Vol. 1. Springer-Verlag, London.

Desu, M. M., and Roghavarao, D. (1991). *Sample size methodology*. Academic Press, Inc., New York. 135pp.

Dijkstra, T. K. (ed). (1988). *On model uncertainty and its statistical implications*. Lecture Notes in Economics and Mathematical Systems, Springer-Verlag, New York, NY, USA. 138pp.

Eberhardt, L. L. (1978). Appraising variability in population studies. *Journal of Wildlife Management* **42,** 207-238.

Eberhardt, L. L., and Thomas, J. M. (1991). Designing environmental field studies. *Ecological Monographs* **61**, 53-73.

Forsman, E. D., DeStefano, S., Raphael, M. G., and Gutierrez, R. J. (Eds.) (1996). Demography of the Northern Spotted Owl. *Studies in Avian Biology*, No. 17. 122pp.

Freedman, D. A. (1983). A note on screening regression equations. *The American Statistician* **37**, 152-155.

Graybill, F. A., and Iyer, H. K. (1994). *Regression analysis: concepts and applications.* Duxbury Press, Belmont, CA. 701pp.

Hairston, N. G. (1989). *Ecological experiments: purpose, design and execution*. Cambridge University Press, Cambridge, UK.

Hand, D. J. (1994). Statistical strategy: step 1. Pages 1-9. In P. Cheeseman and R. W. Oldford (eds.) *Selecting models from data.* Springer-Verlag, New York.

Hayne, D. (1978). Experimental designs and statistical analyses. Pages 3-13 *in* D. P. Synder (editor). *Populations of small mammals under natural conditions.* Pymatuning Symposium in Ecology, Univ. of Pittsburgh, Vol 5.

Hurvich, C. M., and Tsai, C-L. (1989). Regression and time series model selection in small samples. *Biometrika* **76,** 297-307.

Hurvich, C. M., and Tsai, C-L. (1991). Bias of the corrected AIC criterion for underfitted regression and time series models. *Biometrika* **78,** 499-509.

Hurvich, C. M., and Tsai, C-L. (1994). Autoregressive model selection in small samples using a bias-corrected version of AIC. Pages 17-157 *in* H. Bozdogan, editor. Engineering and Scientific Applications, Vol. 3, Proceedings of the First US/Japan Conference on the Frontiers of Statistical Modeling: An Informational Approach. Kluwer Academic Publishers, Dordrecht, Netherlands.

Hurvich, C. M., and Tsai, C-L. (1995). Model selection for extended quasi-likelihood models in small samples. *Biometrics* **51,** 1077-1084.

Johnson, D. K. (1995). Statistical sirens: the allure of nonparametrics. *Ecology* **76**, 1998-2000.

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

Kullback, S. (1959). *Information theory and statistics*. John Wiley & Sons, New York.

Kullback, S., and Leibler, R. A. (1951). On information and sufficiency. *Annals of Mathematical Statistics* **22**, 79-86.

Lebreton, J-D., Burnham, K. P., Clobert, J., and Anderson, D. R. (1992). Modeling survival and testing biological hypotheses using marked animals: a unified approach with case studies. *Ecological Monograph* **62,** 67-118.

Lebreton, J.-D., and North, Ph. M. (Eds.) (1993). *Marked individuals in the study of bird populations.* Birkhasuer Verlag, Boston.

Linhart, H., and Zucchini, W. (1986). *Model selection.* John Wiley and Sons, New York.

Manly, B. F. J. (1992). *The design and analysis of research Studies.* Cambridge University Press, Cambridge, UK. 353pp.

Mead, R. (1988). *The design of experiments: statistical principles for practical applications.* Cambridge University Press, New York.

Nester, M. R. (1996). An applied statistician's creed. *Applied Statistician* **45**, 401-410.

Nichols, J. D., Brownie, C., Hines, J. E., Pollock, K. H., and Hestbeck, J. B. (1993). The estimations of exchanges among populations or subpopulations. Pages 265-280 *in* Lebreton, J.-D., and North, Ph. M. (editors) *Marked individuals in the study of bird populations.* Birkhasuer Verlag, Boston.

Nichols, J. D., Reynolds, R. E., Blohm R. J., Trost, R. E., Hines, J. E., and Bladen, J. E. (1995). Geographic variation in band reporting rates for mallards based on REWARD banding. *Journal of Wildlife Management* **59**, 697-708.

Peirce, C. S. (1955). Abduction and induction. Pages 150-156. In *Philosophical Writings of Peirce*. J. Buchler (Ed.), Dover, New York, New York.

Pollock, K. H., Nichols, J. D., Brownie, c., and Hines, J. E. (1990). *Statistical inference for capture-recapture experiments.* Wildlife Monographs, **107**, 1-97.

Pradel, R., and Lebreton, J-D. (1991). User's manual for Program SURGE, version 4.1. C.E.P.E./C.N.R.S. BP5051, 34033 Montpellier, CEDEX, France, 35pp.

Scheiner, S. M., and Gurevitch, J. (Eds.). (1993). *Design and analysis of ecological experiments*. Chapman and Hall, London.

Shibata, R. (1983). A theoretical view of the use of AIC. Pages 237-244 *in* O. D. Anderson (Ed.) *Time series analysis: theory and practice*. Elsevier Sci. Publ., North-Holland.

Shibata, R. (1989). Statistical aspects of model selection. In *From data to model*. Ed. J. C. Willems, pp. 215-40. Springer-Verlag. London.

Shibata, R. (199_). Bootstrap estimate of Kullback-Leibler information for model selection. *Statistica Sinica* __, __-__.

Shibata, R. (199_). *Statistical model selection*. Springer-Verlag, London. In Prep.

Skalski, J. R., and Robson, D. S. (1992). *Techniques for wildlife investigations: design and analysis of capture data.* Academic Press, Inc., New York, New York.

Skalski, J. R., Hoffmann, A., Smith, S. G. (1993). Testing the significance of individual- and cohort-level covariates in animal survival studies. Pages 9-28 *in* J-D. Lebreton and P. M. North (Eds.) *Marked individuals in the study of bird population.* Birkhauser Verlag, Basel.

Smith, S. G., Skalski. J. R., Schlechte. J. W., Hoffmann, A., and Cassen V. (1994). *Statistical survival analysis of fish and wildlife tagging studies: SURPH.1*. Center for Quantitative Science, School of Fisheries, University of Washington, Seattle.

Stewart-Oaten, A. (1995). Rules and judgments in statistics: three examples. *Ecology* **76**, 2001-2009.

Stone, M. (1977). An asymptotic equivalence of choice of model by cross validation and Akaike's criterion. *Journal of the Royal Statistical Society,* Series B **39,** 44-47.

Stone, M., and Brooks, R. J. (1990). Continuum regression: cross-validated sequentially constructed prediction embracing ordinary least squares, partial least squares and principle components regression (with discussion). *Journal of the Royal Statistical Society*, Series B **52**, 237-269.

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.

Takeuchi, K. (1976). Distribution of informational statistics and a criterion of model fitting. *Suri-Kagaku* (Mathematic Sciences) **153**, 12-18. (In Japanese).

Wedderburn, R. W. M. (1974). Quasi-likelihood functions, generalized linear models, and the Gauss-Newton method. Biometrika **61**, 439-447.

White, G. C., and K. P. Burnham (1999). Program MARK: survival rate estimation from both live and dead encounters. EURING97.

White, G. C., Anderson, D. R., Burnham, K. P., and Otis, D. L. (1982). *Capture-recapture and removal methods for sampling closed populations*. Los Alamos National Laboratory, LA-8787-NERP, Los Alamos, New Mexico. 235pp.

Wilson, K. R., Nichols, J. D., and Hines, J. E. (1989). A computer program for sample size computations for banding studies. U. S. Fish and Wildlife Service, Technical Report 23. 19pp.

Wotawa, M. A. (1993). Survival and recovery rates, and winter distribution patterns of mallards that winter in eastern Colorado. M.S. Thesis, Colorado State University, Fort Collins, 72pp.

Yoccoz, N. G. (1991). Use, overuse, and misuse of significance tests in evolutionary biology and ecology. *Bulletin of the Ecological Society of America* **72**, 106-111.