The Set of Candidate Models

Formulation of critical hypotheses and models to adequately reflect these hypotheses is conceptually more difficult than estimating the model parameters and their precision. Model formulation is the point where the scientific and biological information formally enter the investigation. Building the set of candidate models is partially subjective--that is why scientists are trained, educated and experienced in their discipline. The published literature and experience in the biological sciences can be used to help formulate a set of candidate models, prior to data analysis. The most original, innovative part of scientific work is the phase leading to the proper question. Good approximating models in conjunction with a good set of relevant data can provide deep insights into the underlying biological process and structure.

Where Do Models Come From? [1]

Lehmann (1990) asks "where do models come from" and cites some biological examples (also see Ludwig 1989, Walters 1996). Models arise from questions about biology and the manner in which biological systems function. Relevant theoretical and practical questions arise from a wide variety of sources (see O'Connor and Spotila 1992). Traditionally, these questions come from the scientific literature, results of manipulative experiments, personal experience, or contemporary debate within the scientific community. More practical questions stem from resource management controversies, biomonitoring programs, quasi-experiments, and even judicial hearings.

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 analysis, with too little thought about the reason for the study in the first place (see Hayne 1978 for convincing examples). Worse yet, there is a tendency to rush to the computer and begin to rummage around in the data, hoping that a statistical algorithm will find the "significant" things for you (i.e., data dredging).

The philosophy and theory presented in FW663 must rest on well designed studies and careful planning and execution of field or laboratory protocol. Many good books exist 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,

[1] Much of this material is taken from Burnham and Anderson (1998).

Skalski and Robson 1992, Scheiner and Gurevitch 1993, and Thompson et al. 1998). Chatfield (1991) reviews statistical pitfalls and ways that these might be avoided.

In the following material we will assume that the animal marking data are "sound" and that inference to some larger population is reasonably justified by the nature in which the data were collected. In an important sense, the statistical inference being made is from the sample data to the marked population. If we want to make a further inference from the marked population to the total (i.e., unbanded, unmarked), population, then the inference is often somewhat weaker.

The Global Model

Development of the *a priori* set of candidate models often should include a global model; a model that has many parameters, includes all potentially relevant effects and reflects causal mechanisms thought likely, based on *the science of the situation.* Model $\{\theta_{g*t}, \psi_{g*t}\}$, where θ is a vector of survival probabilities and ψ is a vector of sampling probabilities, is often used as a global model. The global model should also reflect the study design and attributes of the system studied. Specification of the global model should not be based on a probing examination of the data to be analyzed. At some early point, one should investigate the fit of the global model to the data (e.g., examine residuals and measures of fit such as deviance, or formal χ^2 goodness-of-fit tests) and proceed with analysis only if it is judged that the global model is an acceptable fit to the data. Over-dispersion is a common feature of marked animal data and care should be exercised to quantify this issue (e.g., estimation of a variance inflation factor ($\hat{c}$) and use this to adjust the estimated variance-covariance matrix and the model selection criterion (e.g., QAIC, instead of AIC).

Models with fewer parameters can then be derived as special cases of the global model. This set of reduced models represent plausible alternatives based on what is known or hypothesized about the process under study. Generally, alternative models will involve differing numbers of parameters; the number of parameters will often differ by at least an order of magnitude across the set of candidate models. It is only necessary to assess the goodness-of-fit for the global model because model selection methods will not select a more parsimonious model that does not fit.

The more parameters used, the better the fit of the model to the data that can be achieved. Large and extensive data sets are likely to support more complexity and this should be considered in the development of the set of candidate models. If a particular model (parameterization) does not make biological sense, it should not be included in the set of candidate models. In developing the set of candidate models, one must recognize a certain balance between keeping the set 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 that modest over-fitting is less damaging than under-fitting (Shibata 1989). We recommend and encourage a considerable amount of careful, *a priori* thinking in arriving at a set of candidate models (see Peirce 1955, Burnham and Anderson 1998, Chatfield 1995). Using AIC and other information-theoretic methods one can only hope to select the best model from this set. If good models are not in the set of candidates, they cannot be discovered by model selection (i.e., data analysis) algorithms.

A Strategy for Analysis and the A Priori Considerations

The underlying philosophy of analysis is important here. We advocate a conservative approach to the overall issue of *strategy* (see Anderson and Burnham 1999) in the analysis of data in the biological sciences with an emphasis on *a priori* considerations and models to be considered. Careful, *a priori* consideration of alternative models will often require a major change in emphasis among many people. This is often an unfamiliar concept to both biologists and statisticians where there has been a tendency to use either a traditional model or a model with associated computer

software, making its use easy (Lunneborg 1994). 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 initial investigation.

We advocate the deliberate exercise of carefully developing a set of, say, perhaps 4-20 alternative models as potential approximations to the population-level information in the data available and the scientific question being addressed. Some practical problems might have as many as 70-100 or more models that one might want to consider. The number of candidate models is often larger with large data sets (e.g., long-term data with the multi-strata models). We find that people include many models that are far more general than the data could reasonably support (e.g., models such as $\{\theta_{g*t}, \psi_{g*t}\}$, when the additive models $\{\theta_{g+t}, \psi_{g+t}\}$ might be both more biologically realistic and parsimonious). This set of models, developed without first deeply examining the data constitute the "set of candidate models." The science of the issue enters the analysis through the *a priori* set of candidate models.

Problems with Using Example Data Sets in FW-663

Consider a hypothetical team of biologists working on the effect of a long-line fishery on survival of black-footed albatross (*Deomedea nigripes*). These people have a good understanding of albatross biology and some understanding of the developing fishery and its potential indirect effects on albatross populations. They begin to plan a sampling program, involving thousands of banded albatross and a design to attempt to isolate potential impacts of the fishery on albatross survival and recapture probabilities. Data collection is carried out for 12 years and a strategy is developed for the analysis of the data, including various covariates that have been measured. This team can likely formulate a good set of candidate models prior to analysis. Particularly if a team member has quantitative expertise, the set of models will almost surely be useful, leading to results and inferences being at the confirmatory end of the spectrum. Finally, if they then want to add some additional models, after studying the results for the *a priori* efforts, they can certainly chase these added insights (hopefully, they will fully acknowledge the two types of inference in preparing their written report).

In FW-663 the situation is very different and this difference can lead to confusion. In FW-663, students are given a data set on marked animals and a paragraph or two explaining the biological setting. There is not an intimate biological understanding of the data in the classroom or computer laboratory, like there is in the real world. The student has very limited time to get deeply immersed in the data, underlying hypotheses, why the data were actually collected in the first place, etc., etc.

The result is that students are rightfully leery of developing a meaningful list of *a priori* models or they might even feel that it is unrealistic to develop a model set without some data dredging. For the student, the problem statement is short and relatively little time (perhaps 4-6 hours at the most) is available to fully consider the various issues, develop some hypotheses and models corresponding to the hypotheses. Even if 4-6 pages of detail on the problem setting and the data available might help in only a marginal way – reality is a bit distant in a class-room setting! In the real world, investigators

are often working in teams, have intimate knowledge of the biological problem, and might often develop the candidate model set over a period of several months.

Thus, we suggest that students learn the statistical theory and gain proficiency in the computer software (e.g., MARK and DISTANCE) making good use of the exercises and examples, but realize that there is a certain level of artificiality in the set of models derived in classroom settings. Finally, we hope that students involved in the analysis of real data (thesis or dissertation research, helping colleagues, or researching biological systems upon graduation) pay particular attention to carefully and deliberately developing a good set of candidate models. Here it is important to have a solid rational to include some models in the set, as well as having a rational for other models to be omitted from the candidate model set.

Literature Cited

Anderson, D. R., and Burnham, K. P. (1999). General strategies for the analysis of ringing data. *Bird Study* **46**, Supplement, S261-270.

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.

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

Chatfield, C. (1995). 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 Company, Boston, MA, USA.

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

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

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

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

Lehmann, E. L. (1990). Model specification: the views of Fisher and Neyman, and later developments. *Statistical Science* **5**, 160-168.

Ludwig, D. (1989). Small models are beautiful: efficient estimators are even more beautiful. Pages 274-284 *in* C. Castillo-Chavez, S. A. Levin, and C. A. Shoemaker (Eds.) *Mathematical approaches to problems in resource management and epidemiology*. Springer-Verlag, London.

Lunneborg, C. E. (1994). *Modeling experimental and observational data.* Duxbury Press, Belmont, CA, USA. 506pp.

Madigan, D., and Raftery, A. E. (1994). Model selection and accounting for model uncertainty in graphical windows using Occam's window. *Journal of the American Statistical Association* **89**, 1535-1546.

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, NY.

O'Connor, M. P., and Spotila, J. R. (1992). Consider a spherical lizard: animals, models, and approximations. *American Zoologist* **32**, 179-193.

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

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

Shibata, R. (1989). Statistical aspects of model selection. Pages 215-240 *in* J. C. Willems (Ed.) *From data to model*. Springer-Verlag. London.

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

Thompson, W. L., White, G. C., and Gowan, C. 1998. *Monitoring vertebrate populations.* Academic Press, New York.

Walters, C. J. (1996). Computers and the future of fisheries. Pages 223-238 *in* B. A. Megrey, and E. McKsness (Eds.) *Computers in fisheries research.* Chapman and Hall, London. 254pp.