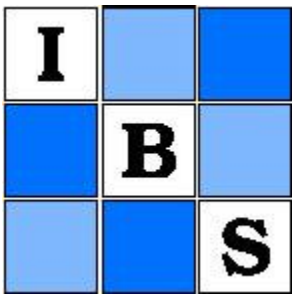

BIOMETRICS 53, 603–618
June 1997

Model Selection: An Integral Part of Inference

S. T. Buckland,[1] K. P. Burnham,[2] and N. H. Augustin[1]

[1]School of Mathematical and Computational Sciences, University of St. Andrews, North Haugh, St. Andrews, Fife KY16 9SS

[2]Colorado Cooperative Fish and Wildlife Research Unit, Fort Collins, Colorado 80523, U.S.A.

SUMMARY

We argue that model selection uncertainty should be fully incorporated into statistical inference whenever estimation is sensitive to model choice and that choice is made with reference to the data. We consider different philosophies for achieving this goal and suggest strategies for data analysis. We illustrate our methods through three examples. The first is a Poisson regression of bird counts in which a choice is to be made between inclusion of one or both of two covariates. The second is a line transect data set for which different models yield substantially different estimates of abundance. The third is a simulated example in which truth is known.

1. Introduction

There is a considerable amount of literature devoted to methods for quantifying precision of statistical estimators. Disagreement is not uncommon over, for example, alternative methods of setting confidence intervals, as illustrated by the debate in the recent bootstrap literature over whether percentile confidence intervals are backwards (Efron and Tibshirani, 1993). Often, much effort is expended in developing better confidence limits for specialized applications, and pathological examples are used to demonstrate the supposed superiority of the new method over other contenders. For moderate sample size, differences between competing methods are often trivial when compared with the potential impact of incorporating model selection into the statistical inference.

Although model selection is widely recognized as central to good inference, paradoxically, it has seldom been integrated fully into inference. For example, there are many methods in multiple regression for identifying an appropriate subset of covariates. Having identified them, subsequent inference is usually conditional on the selected model; that is, we assume that the model is correct. It is more defensible to recognize the uncertainty in model selection when quantifying the precision of an estimator. Under this philosophy, model misspecification bias is not bias at all, but merely a component of the variance. In practice, some degree of bias will remain and will be a decreasing function of the richness of the set of competing models. Several recent researchers have considered how to incorporate model selection into inference, and an excellent discussion of their work is provided by Chatfield (1995).

The reason that inference is generally conditional on the selected model is the complexity encountered when attempting inference unconditional on that model. This may be circumvented to a useful extent using simple weighting methods, by adopting computer intensive or simulated inference, or by some combination of both. In this paper, we have two goals: we demonstrate the importance of integrating model selection with statistical inference through simple examples, and we provide the applied statistician with easy-to-use tools. Integration is achieved using simple weighting methods, where the weights are obtained from information criteria or by using the bootstrap. Different philosophies suggest different methodologies for incorporating model selection uncertainty into inference. Our philosophy is that truth is high (effectively infinite) dimensional. The more information that is gathered, the greater is the model complexity that the data can

Key words: AIC; BIC; Information criteria; Model selection uncertainty; Simulated inference.

support. If data are sparse, they can support only a simple model with few parameters. In our view, model selection is the process of identifying the best approximating model, accepting that the data can never support, and we can never identify, the true model. In many circumstances, it may be possible to identify a global or maximal model; that is, one that incorporates all possible parameters of interest. It need not be a saturated model. Model selection is then the process of identifying which submodel is the best approximating model. In practice, in order to incorporate model selection uncertainty into inference, we must make simplifications. We consider two options. The first, adopted in Section 2, assumes that the fitted models are in some sense a random sample from an infinite set of possible models, each of which provides a valid estimate of the parameter in its own right. In the second approach (Section 3), a set of models is defined, and we seek to estimate which of these is the best approximating model. For some purposes, we may need to assume that one model of the set is the true model. As noted by Berk (1966), this has the shortcoming that, if the true model is not among the set of contending models, as more data accumulate, the model in the set closest to the true model will appear to be the true model. We seek methods that are robust to this shortcoming.

2. Model Weighting

We outline here a philosophy for weighting contending models, in preference to selecting between them. This can be done within a Bayesian framework, so that model selection is replaced by estimated probabilities that different models are correct. However, such an approach raises the issue of how to set priors. Raftery (1996) and Kass and Raftery (1995) use Bayes factors to incorporate model selection uncertainty in a Bayesian context, but the method can be sensitive to the choice of prior. The method is also problematic if there are many possible models. Solutions to both problems, for example using intrinsic Bayes factors (Berger and Pericchi, 1996) for the first or implementing a Markov chain Monte Carlo algorithm for moving through the model space (Madigan et al., 1994) for the second, add to the complexity of the Bayesian method. The complexity also forces the user to adopt computer-intensive methods, often substantially more computer intensive than the bootstrap.

Here, we explore simpler philosophies that allow applied statisticians to integrate model selection into inference routinely. We seek weights that can be associated with estimates derived under each of the contending models, $M_k, k = 1, \ldots, K$. Using these weights w_k, scaled so that $\Sigma\, w_k = 1$, the estimate of a parameter θ (assumed to be common to all models) is taken to be

$$\hat{\theta} = \sum_k w_k \hat{\theta}_k, \tag{1}$$

where $\hat{\theta}_k$ is the estimate of θ under model k.

How we estimate the variance of $\hat{\theta}$ depends on our philosophy. Consider first the unrealistic case that the $\hat{\theta}_k$ are identically distributed with expectation θ and that the weights w_k are known constants. Then we obtain

$$\mathrm{var}(\hat{\theta}) = \sum_k w_k^2 \mathrm{var}(\hat{\theta}_k) + \sum_k \sum_{l \neq k} w_k w_l \mathrm{cov}(\hat{\theta}_k, \hat{\theta}_l). \tag{2}$$

The problem arises of how to estimate the covariance. We could resort to simulated inference (below) and estimate the sample covariance between estimates from analyses of bootstrap resamples. Alternatively, we might argue that the covariance will be high because each model is fitted to the same data set and then choose the conservative strategy of setting it equal to its maximum possible value, which is the geometric mean of the variances of the estimates under models k and l. In that case, we obtain an upper bound for $\mathrm{var}(\hat{\theta})$ of

$$\mathrm{var}(\hat{\theta}) \leq \left\{ \sum_k w_k \sqrt{\mathrm{var}(\hat{\theta}_k)} \right\}^2. \tag{3}$$

However, this variance fails to incorporate a component representing model misspecification bias. Suppose we define $\theta_k = \theta + \beta_k$, where β_k is the misspecification bias that arises in estimating θ under model k. Suppose further that $\mathrm{E}(\beta_k) = 0$, where expectation is over all possible models. Then we denote

$$\mathrm{E}(\hat{\theta}_k \mid \beta_k) = \theta + \beta_k = \theta_k. \tag{4}$$

If we take the expectation over all possible models of this expression, we obtain $\mathrm{E}(\hat{\theta}_k) = \theta$. If we denote

$$\mathrm{var}(\hat{\theta}_k \mid \beta_k) = \mathrm{E}\left[(\hat{\theta}_k - \theta_k)^2\right] \tag{5}$$

and

$$\mathrm{var}(\hat{\theta}_k) = \mathrm{E}\left[(\hat{\theta}_k - \theta)^2\right], \tag{6}$$

then

$$\mathrm{var}(\hat{\theta}_k) = \mathrm{var}(\hat{\theta}_k \mid \beta_k) + \beta_k^2. \tag{7}$$

Hence, we have

$$\mathrm{var}(\hat{\theta}) = \sum_k w_k^2 \mathrm{var}(\hat{\theta}_k) + \sum_k \sum_{l \neq k} w_k w_l \mathrm{cov}(\hat{\theta}_k, \hat{\theta}_l), \tag{8}$$

which, if we assume perfect correlation, becomes

$$\mathrm{var}(\hat{\theta}) = \left\{ \sum_k w_k \sqrt{\mathrm{var}(\hat{\theta}_k \mid \beta_k) + \beta_k^2} \right\}^2. \tag{9}$$

This variance may be estimated by substituting $\hat{\beta}_k = \hat{\theta}_k - \hat{\theta}$ and $\widehat{\mathrm{var}}(\hat{\theta}_k \mid \beta_k)$. The estimates $\hat{\theta}_k$ and $\widehat{\mathrm{var}}(\hat{\theta}_k \mid \beta_k)$ are found by normal inference methods, assuming that model k is the true model, and $\hat{\theta}$ is given by equation (1).

How can we avoid assuming that the estimators are perfectly correlated? The bootstrap avoids the problem by generating resamples from the data and assuming that these resamples are independent data sets. If there are sufficient data, a less computer-intensive solution is to divide the data into K equal (or nearly equal) groups, where K is the number of models. Each model is fitted to one of these data sets to obtain the estimates $\hat{\theta}_k$ and corresponding variances $\widehat{\mathrm{var}}(\hat{\theta}_k \mid \beta_k)$. The above analysis conditions on the weights w_k, and these will be estimated more reliably by fitting each model to the full data set. Independence of the $\hat{\theta}_k$ can now be assumed, so that equation (9) is replaced by

$$\mathrm{var}(\hat{\theta}) = \sum_k w_k^2 \left\{ \mathrm{var}(\hat{\theta}_k \mid \beta_k) + \beta_k^2 \right\}. \tag{10}$$

Note that the above inference can only be applied to parameters that are in common to all contending models (e.g., $\mathrm{E}(y_i)$, the expectation of the ith observation, or $\mu = \mathrm{E}(\bar{y})$). For parameters that are not present in all models, we can apply the above methods for the subset of models in which the parameter occurs and use the sum of weights for that subset as an indicator of the importance of that parameter.

In the above development, we have assumed the weights are known constants, whereas in practice they will be estimated. How should we do this? Choice is intrinsically linked with choice of model selection criterion. Key components of model selection are expert (or subjective) opinion and model availability. The latter cannot be integrated with inference; if none of the available models (including composite models) adequately reflect reality, inference will be poor irrespective of whether model selection uncertainty is incorporated. Expert opinion can be integrated by adopting a Bayesian framework, using subjective priors, or by creating bootstrap resamples and using expert opinion to select a model 'independently' for analyzing each resample. The latter strategy is only implementable if an expert system is developed, allowing the thought processes of the expert to be automated. Hypothesis testing is widely used for model selection, and there are examples of specific applications in the literature for which the hypothesis testing has been fully incorporated into inference. However, there are problems in developing a general approach based on hypothesis tests. We prefer instead to work with information criteria of the form

$$I = -2\log(L) + q, \tag{11}$$

where L is the likelihood function, evaluated by substituting the maximum likelihood estimates of the parameters, and q is a penalty that is a function of the number of parameters and/or the number of observations. Two examples are Akaike's Information Criterion AIC (Akaike, 1973; Burnham and Anderson, 1992), for which $q = 2p$, where p is the number of parameters, and the Bayes Information Criterion BIC, for which $q = p\log(n)$, where n is the number of observations.

Apart from a constant, AIC is a consistent estimator of the Kullback–Leibler discrepancy between the distribution that generated the data and the model that approximates it. A small-sample bias adjustment was investigated by Hurvich and Tsai (1989, 1995):

$$\mathrm{AIC_c} = \mathrm{AIC} + \frac{2p(p+1)}{n-p-1}. \tag{12}$$

The philosophy underlying AIC and $\mathrm{AIC_c}$ is that 'truth' is high-dimensional, requiring many (possibly infinitely many) parameters to describe it. Sakamoto, Ishiguro, and Kitagawa (1986) note that 'AIC is not a criterion for the estimation of the true order but the one for the best model fit.' In other words, we seek the best approximating model. The dimension of this model will be low if there are few data and will increase as more information becomes available. By contrast, BIC is dimension consistent. It provides a consistent estimate of the true order of the model, at the expense of assuming that a true model exists and is low-dimensional. (For further discussion on these issues, we refer the reader to the following papers: Akaike, 1978; Bozdogan, 1987; Rissanan, 1989; Schwarz, 1978; Sclove, 1987; Shibata, 1989.)

Suppose there are K models, with $I_k = -2\log(L_k) + q_k$ for model k. Then the model with the smallest value for I is in some sense the best model. If we compare model i with model j, we find

$$\frac{L_i \exp(-q_i/2)}{L_j \exp(-q_j/2)} = \frac{\exp(-I_i/2)}{\exp(-I_j/2)}. \tag{13}$$

If the penalties are equal for the two models ($q_i = q_j$), this is just the ratio of the likelihoods, which is the Bayes factor for comparing simple models. If, further, the prior odds ratio is one, this expression represents the posterior odds ratio of the respective models. If BIC is used, equation (13) gives Schwarz's (1978) approximation of the Bayes factor. A plausible choice for weight w_k is thus

$$w_k = \frac{\exp(-I_k/2)}{\sum_{i=1}^{K} \exp(-I_i/2)}, \qquad k = 1, \ldots, K. \tag{14}$$

By defining the weights in this way, we ensure that two models with the same value for I are given the same weight, whether or not they have the same penalty q. The Bayes factor might more descriptively be termed the relative likelihood factor. We term the ratio $\{L_i \exp(-q_i/2\}/\{L_j \exp(-q_j/2)\}$ the relative penalized likelihood factor.

3. Simulated Inference

A key tool of simulated inference is the bootstrap (Efron, 1979), which allows resamples to be generated. Each resample is analyzed exactly as if it were the original sample. Thus, a simple method of incorporating model selection uncertainty into inference is to apply the model selection procedure independently to each resample (Buckland, 1982; Sauerbrei and Schumacher, 1992; Hjorth, 1994; Norris and Pollock, 1996).

To implement the nonparametric bootstrap, observations are sampled with replacement from the original data set until sample size is equal to that for the real data. These observations comprise the first bootstrap resample. The process is repeated to give, say, b resamples. Any estimator of interest is evaluated using the data from each resample in turn. The sample variance of the bootstrap estimates provides an estimated variance, either for the estimate from the real data or for the mean of the bootstrap estimates. An approximate $100(1-2\alpha)\%$ 'percentile' confidence interval is given by ordering the bootstrap estimates from smallest to largest and selecting the rth and sth values from the list, where $r = (b+1)\alpha$ and $s = (b+1)(1-\alpha)$ (Buckland, 1984). It is convenient to choose b so that r and s are integer. Otherwise, if b is large, they may be rounded to the nearest integer values; for smaller b, linear interpolation between the rth and $[r+1]$st values may be used for the lower limit and similarly for the upper limit.

For regression problems, Efron (1979) noted that resampling should be from the residuals r_i, so that analysis remains conditional on the covariate values. If the bootstrap is used to allow for model selection uncertainty, this strategy assigns too much weight to the model from which the residuals were obtained. Possible solutions to this are considered in our first example.

The nonparametric bootstrap assumes that the sampling units (usually observations or residuals) are independently and identically distributed. For generalized linear models, observations are assumed to be independently distributed, but their variance is a function of their expectation. This difficulty may be resolved using the parametric bootstrap, in which the ith bootstrap

observation is generated from the assumed parametric distribution fitted at the covariate values associated with observation y_i. However, if counts are overdispersed, this procedure fails to recreate the overdispersion in the bootstrap resamples. Bravington (1993) has proposed a generalization of the nonparametric bootstrap that requires observations to be independently, but not identically, distributed and that preserves any overdispersion. His proposal is to transform observation y_i to $u_i = \hat{F}_i(y_i)$, where $F_i(y_i)$ is the cumulative distribution function of y_i. A bootstrap resample is now generated by selecting a sample of size n with replacement from these u_i. If the ith value in the resample is u_j, then the ith bootstrap observation is calculated as $\hat{F}_i^1(u_j)$. The process is repeated to generate b bootstrap resamples. In the absence of overdispersion, the method replicates the parametric bootstrap, and the sample of u values would be uniform on $(0, 1)$ if the cumulative distribution function were known rather than estimated.

The above provides a computer intensive framework for incorporating model selection bias into inference. It also provides an alternative method of obtaining weights, a method which recognizes that the weights are not known constants: the weight for model k is estimated by the proportion of resamples in which model k is identified as the best approximating model. These weights differ in one important respect from those of equation (14). Suppose that two models, A and B, are indistinguishable in that their respective likelihoods are identical (whatever the data). The bootstrap selects a 'winner' for each resample (in this case, we might choose to pick one of the two models at random), so that the total weight assigned to these two models would be the same as the weight assigned to either one if the other were omitted. By contrast, adopting equation (14), the total weight $w_A + w_B$ would be larger (up to double) by including both models rather than just one.

The bootstrap, with model selection applied independently to each resample, allows us to estimate variance and to generate robust confidence intervals when we do not wish to condition our analyses on a single selected model. Suppose we wish to assign weight w_k to model k, where w_k is as found from equation (14). The weight w_k will not in general equal the proportion of times model k was selected in the analyses of the resamples, so we must reweight the bootstrap samples. In the unweighted percentile method, bootstrap estimates are ordered and each is given equal weight to determine the required percentiles of the distribution. In what we term the weighted percentile method, a weight of $v_i = w_k b / b_k$ is assigned to bootstrap estimate i, where k indicates the model selected when analyzing resample i, and b_k is the number of resamples for which model k was selected. To obtain the lower percentile limit, find the largest integer r such that

$$\sum_{1}^{r} v_i \leq (b+1)\alpha. \tag{15}$$

In the weighted case, it is not possible to choose b such that equality holds, and linear interpolation may be preferred. The upper limit may be found similarly.

4. Examples

4.1 *Poisson Regression*

The first example considers a simple multiple regression problem with two correlated explanatory variables, an assumed Poisson error distribution and a log link function (Table 1). In these fictitious data, transect counts of singing males of the songbird *Troglodytes invisibilis* were made on consecutive days, and we wish to predict future counts, given temperature and wind speed. In practice, date might also be a useful predictor but, to keep the example simple, we ignore it.

The problem is to predict the count on day 19. Both covariates are highly correlated with the count and with each other (Figure 1), making model selection difficult. Traditional inference based on selecting the model that explains most of the variation, subject to excluding covariates that offer no significant improvement over the fit without them, is compromised. Unless the analyst knows that one of the covariates is irrelevant, prediction for day 19 will appear to be more precise than is justified because we assume that the true model is known.

How can inference be improved? One solution is to adopt a Bayesian framework. In the absence of better information, equal prior probabilities for the two regression models with a single covariate might be defensible, but what of the model with no covariates or both covariates?

A second option is to use relative penalized likelihood factors as defined above to provide weights for the estimates from each contending model (see equation (14)), from which the weighted average and its estimated variance are easily calculated.

Instead, we might opt to generate bootstrap resamples and apply our model selection criterion separately to each resample. If we adopt Bravington's (1993) method, the Poisson error distribution

Table 1
Transect counts of a species of songbird in a study area on consecutive days. Covariates temperature (°C) and wind speed (m/s) were also recorded.

Day	Count	Temperature	Wind speed	Day	Count	Temperature	Wind speed
1	17	22	1.1	11	15	15	3.7
2	45	23	0.5	12	39	22	0.8
3	9	17	2.9	13	18	17	1.7
4	40	22	0.4	14	29	24	0.8
5	18	14	4.8	15	22	13	3.8
6	15	13	3.9	16	10	15	3.1
7	8	14	5.7	17	15	16	2.3
8	21	18	2.6	18	27	22	0.4
9	42	24	0.5	19	??	22	1.5
10	38	26	0.3				

is retained, as is any overdispersion in the observations. The question remains of how to choose the model from which the bootstrap resamples are generated. We consider three options: (a) generate all bootstrap resamples from the model selected when analyzing the original data; (b) generate all bootstrap resamples from the model with both covariates, regardless of significance; and (c) select each model with probability equal to its weight as calculated from equation (14), and then generate the next resample from the selected model. We also consider a fourth method: (d) resample from the sampling units; that is, bootstrap the counts along with their associated covariate values.

We show in Table 2 the weights from equation (14), using AIC, together with the proportion of times each model was selected, again using AIC, under each of the four bootstrap resampling strategies. It is clear that the different methods give substantially different weights to the three Poisson regression models. Relative to the bootstrap methods, equation (14) gives greater weight to the model with both covariates present. Bootstrap method (a) favors the model from which the bootstrap resamples were generated, as might be anticipated. This does not occur for method (b), for which the model with both covariates was selected in just 20% of cases, roughly the same as for the other bootstrap methods. Methods (c) and (d) performed very similarly to method (b). It is interesting to note that the naive method (d), which treats the observations as if they are i.i.d. and fails to condition on the values of the covariates, in the sense that it resamples the covariate values along with the observations, has a similar performance to methods (b) and (c), whereas method (a), the most obvious implementation of the bootstrap, clearly biases results in favor of the model with the smallest AIC value.

The models with temperature alone, with wind speed alone, and with both yield predicted counts for day 19 of 30.0, 26.1, and 27.8 birds, respectively. Using weights from equation (14), the predicted count from equation (1) is 27.4 birds. Equation (9) yields a variance estimate of 9.4, from which an approximate 95% confidence interval for the expected count of (22.1, 34.1) was obtained, assuming log-normality. We show the bootstrap distribution of predictions of the count on day 19 using method (b) in Figure 2. In Table 3, we show bootstrap confidence intervals for the expected count on day 19 under each model and each resampling method. The composite interval is wider than

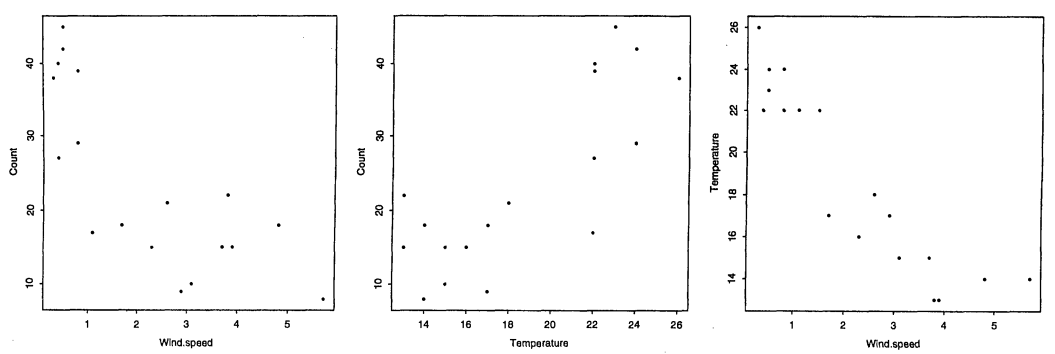

Figure 1. Scattergrams of number of singing *Troglodytes invisibilis* against wind speed (left) and temperature (middle). Also shown is a scattergram of temperature against wind speed (right).

Table 2

AIC for each Poisson regression model and the corresponding weights. The model $\mathrm{E}(y) = e^{\alpha}$ *had zero weight and is omitted. Temperature is denoted by t and wind speed by s. Also shown is the proportion of times each model was selected under each of four bootstrap sampling strategies: (a) generate all bootstrap resamples from the model selected when analyzing the original data; (b) generate all bootstrap resamples from the model with both covariates; (c) for each resample, select a model with probability equal to its weight and then generate the resample from it; (d) resample from the counts along with their associated covariate values.*

Model E(y)	AIC	$w_k = \exp(-\mathrm{AIC}/2) \div \Sigma \exp(-\mathrm{AIC}/2)$	Proportion of bootstrap resamples			
			(a)	(b)	(c)	(d)
$e^{\alpha+\beta t}$	42.27	0.16	0.08	0.32	0.27	0.32
$e^{\alpha+\gamma s}$	40.28	0.43	0.69	0.48	0.51	0.48
$e^{\alpha+\beta t_\gamma s}$	40.39	0.41	0.23	0.20	0.22	0.20

that for either model with a single covariate for all four resampling methods. It is wider under methods (b) and (d) and narrower under methods (a) and (c) than the interval obtained from the model with both covariates present. Under methods (b)–(d), the composite interval is very close to the analytic interval above. This is perhaps surprising, given both the different philosophy and the assumption of perfect correlation between estimates obtained from different models that underlie the analytic interval.

4.2 *Line Transect Sampling*

In line transect sampling, an observer walks along a line and records the perpendicular distance x from the line of each animal detected. It can be shown that the usual estimate of animal density $\hat{D}$ is a function of $\hat{f}(0)$, the fitted probability density function of perpendicular distances evaluated at zero distance, i.e.,

$$\hat{D} = \frac{n \cdot \hat{f}(0)}{2L}, \tag{16}$$

where n is the number of animals detected and L is the total length of transect travelled (Buckland et al., 1993).

We use a ruffed grouse data set taken from Gates (1979), in which $n = 218$, as an example of the effect of incorporating model selection uncertainty in inference. The statistical problem is to model the probability density function $f(x)$. The program DISTANCE (Laake et al., 1993) was used. It first fits a user-specified parametric key function to the data, then allows polynomial or cosine

Table 3

Bootstrap confidence intervals for count on day 19 *under each Poisson regression model, excluding the null model. Temperature is denoted by t and wind speed by s. The composite intervals are obtained by pooling across resamples in which different models were selected (see text). Nominal confidence level is* 95%. *Bootstrap methods (a)–(d) are defined in Table* 2. *The number below each interval is the number of resamples from which the interval is calculated using the weighted percentile method.*

Model E(y)	Bootstrap resampling method			
	(a)	(b)	(c)	(d)
$e^{\alpha+\beta t}$	(24.8, 33.1) 82	(25.4, 33.4) 325	(25.9, 34.7) 269	(25.5, 35.7) 317
$e^{\alpha+\gamma s}$	(22.8, 29.2) 689	(22.7, 29.3) 477	(23.1, 30.2) 514	(22.2, 29.8) 483
$e^{\alpha+\beta t+\gamma s}$	(19.8, 31.2) 229	(22.6, 31.8) 198	(20.8, 33.7) 217	(23.4, 31.6) 200
Composite model	(20.9, 31.1) 1000	(22.9, 32.7) 1000	(22.2, 33.7) 1000	(22.9, 33.8) 1000

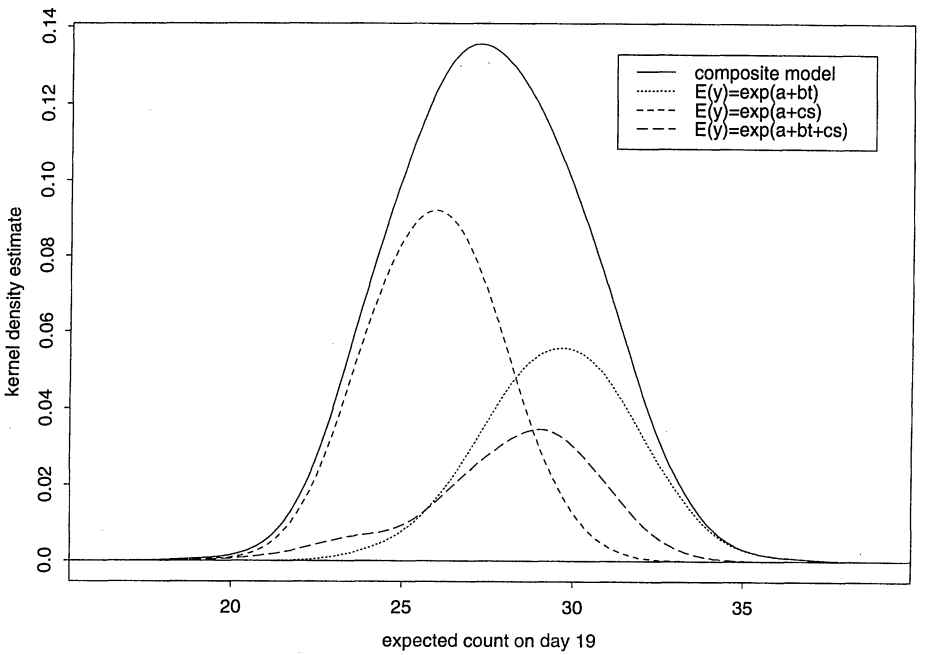

Figure 2. The distribution of predictions for day 19 under each model. The curves were fitted using the kernel algorithm of Silverman (1982), with the default window width. Each curve was scaled so that it integrates to the proportion of times the corresponding model was selected when analysing bootstrap resamples. Thus the sum of the curves (the 'composite model') estimates the probability density function of the estimate of expected count for day 19.

series adjustments to improve the fit (Buckland, 1992). Model selection was restricted to four models here: a half-normal key and cosine adjustment terms; a hazard-rate key and simple polynomial adjustment terms; the Fourier series model (uniform key and cosine adjustment terms); and a negative exponential key and simple polynomial adjustment terms (Buckland et al., 1993). Plots of the perpendicular distance data x, together with the fitted density functions, are shown in Figure 3.

Usually in line transect sampling, the nonparametric bootstrap is implemented by resampling from the transects rather than from individual detections. However, for estimating the precision of $\hat{f}(0)$, resampling from individual detections works well, and we adopt that strategy here, as we only have the data pooled across lines. Each model was fitted both to the original data and to each of $b = 1000$ resamples, and Akaike's Information Criterion was used to select between the four models. In Figure 4, we show the distribution of bootstrap estimates of $f(0)$ under each model, using kernel density estimation to smooth each distribution. Each curve is scaled so that the area under it is equal to the proportion of times the corresponding model was selected in the bootstrap resamples. Also shown is the sum of these four curves; the area under this composite curve is unity. For each curve, the endpoints of a 95% percentile confidence interval are shown. It can be seen that the lower confidence limit under the negative exponential model lies above the upper limit for the hazard-rate model. The two remaining models yield confidence intervals intermediate between these, although the lower limit under the negative exponential model is close to the upper limit under all three of the other models. The limits corresponding to the composite curve better reflect uncertainty in the true value $f(0)$ than do those from any of the individual curves.

In the above, we have allowed the bootstrap to determine weights for the different estimators. These weights are equal to the proportion of times each estimator was selected by AIC when analyzing the resamples. In Table 4, we show these weights in column (a), together with those obtained from equation (14). In this example, both approaches yield very similar results. If we were to replot Figure 4, scaling each individual curve to integrate to the corresponding weight from equation (14), the composite curve and, hence, the weighted percentile interval would change only marginally. In column (b) of Table 4, we show the bootstrap weights obtained if we constrain the number of adjustment terms for each model to that selected in the analysis of the real data.

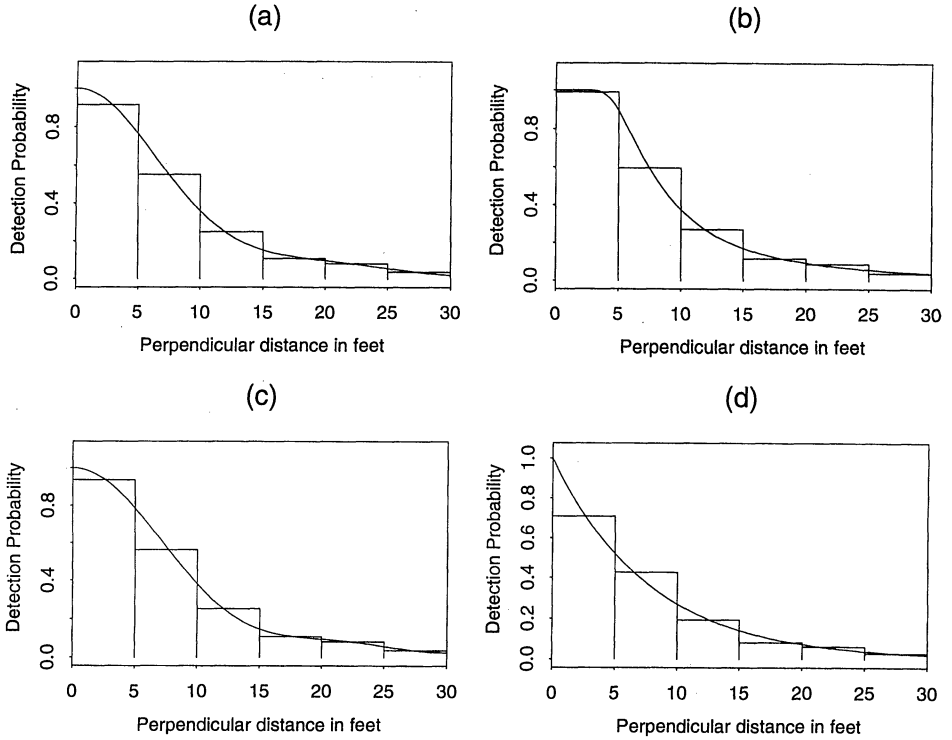

Figure 3. Fit of (**a**) the half-normal, (**b**) the hazard-rate, (**c**) the Fourier series, and (**d**) the negative exponential models to the ruffed grouse perpendicular distance data. No adjustment terms were selected in the cases of the hazard-rate and negative exponential models, whereas one cosine term adjustment was made to the half-normal model, and three cosine terms were selected for the Fourier series model.

Although this method of analysis is less defensible than estimating the number of adjustment terms independently for each resample, it should yield weights closer to those obtained from equation (14). We see from Table 4 that, although agreement is better (the bootstrap weights under methods (a) and (b) differ significantly at the 5% level for the first two models), the differences have little practical significance.

Equation (1) yields $\hat{\theta} = \hat{f}(0) = 0.1123$. The estimate under the model favored by AIC is considerably higher at 0.1333 (Tables 4 and 5). The 95% percentile confidence interval for $f(0)$ obtained from the full set of 1000 bootstrap estimates is (0.0830, 0.1486), and the bootstrap standard error is $\text{se}\{\hat{f}(0)\} = 0.0194$. For comparison, substituting estimates into equation (9) yields $\text{se}(\hat{\theta}) = \text{se}\{\hat{f}(0)\} = 0.0177$. Assuming $\hat{f}(0)$ is log-normally distributed, an approximate 95% confidence interval is then (0.0826, 0.1526). Although the analytic method assumes the weights are known and four models are too few to estimate adequately the contribution of model misspecification to the variance, the differences in precision estimates are again remarkably small. As noted earlier, the methods are based on different philosophies. The bootstrap estimates are based on the assumption that one of the four models is the true model; as the data fail to rule out any of the possible models, the bootstrap interval effectively includes any estimate of $f(0)$ that is plausible under at least one of the models. The analytic method treats each estimate of $f(0)$ as a valid estimate in its own right. If each estimate had been assumed independent of the others, the analytic interval would have been shorter than the bootstrap interval; by assuming perfect correlations between the estimates, we have largely eliminated the difference between the methods. The bootstrap replicates can be used to estimate these correlations, although we then forego the advantage of avoiding simulated inference methods. Instead, we randomly subdivided the data into four groups, fitted one model to each group (thus allowing us to assume that the estimates under each model are independent), and applied equation (10). We obtained $\text{se}\{\hat{f}(0)\} = 0.0238$; assuming a log-normal distribution gives a corresponding confidence interval of (0.0745, 0.1694).

In Table 6, we show the result of replacing AIC by the Bayes Information Criterion (BIC). The AIC penalty of $2p$ becomes $p\log(n) = 5.38p$. Hence, because sample size is relatively large here,

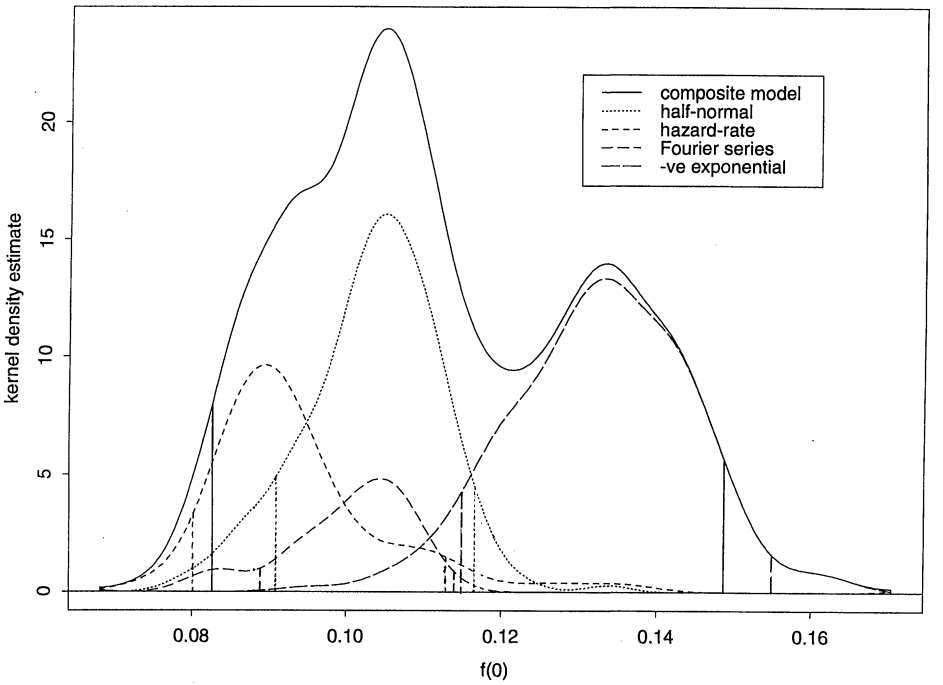

Figure 4. The distribution of bootstrap estimates under each model from Figure 3. The curves were fitted using the kernel algorithm of Silverman (1982), with the default window width. Each curve was scaled so that it integrates to the proportion of times the corresponding model was selected when analysing bootstrap resamples. Thus the sum of the curves (the 'composite model') estimates the probability density function of the estimate $\hat{f}(0)$. The vertical lines indicate 95% percentile confidence limits for $f(0)$ under the respective models.

BIC severely penalizes models with more parameters. Thus, the weights w_k are appreciably different than for AIC. However, as for AIC, the bootstrap again yields weights closely comparable with those obtained from equation (14). Only in the case of the model that fits least well is the difference in weights noteworthy. Equation (14) gives $w_k = 0.008$ for this model, whereas the two bootstrap proportions are 0.000 and 0.001. That is, the poorest fitting model was selected in just one of the analyses of 2000 bootstrap resamples.

Table 4

Relative penalized likelihoods for four line transect models fitted to ruffed grouse data. Likelihoods were penalized using Akaike's Information Criterion (AIC). Also given is the proportion of times each model was selected by AIC when analyzing 1000 bootstrap resamples (a) allowing the number of adjustment terms used for a particular model to vary between resamples and (b) fixing the number of adjustment terms to equal that determined in the analysis of the original data. (Standard errors in parentheses.)

k	Estimator key	Series	AIC	$w_k = \exp(-\text{AIC}/2) \div \Sigma \exp(-\text{AIC}/2)$	Proportion of bootstrap resamples (a)	(b)
1	Half-normal	Cosine	588.50	0.283	0.326 (0.015)	0.273 (0.014)
2	Hazard-rate	Polynomial	588.86	0.236	0.196 (0.013)	0.234 (0.013)
3	Uniform	Cosine	590.40	0.109	0.087 (0.009)	0.098 (0.009)
4	Negative exp.	Polynomial	587.95	0.372	0.391 (0.015)	0.395 (0.015)

Table 5

Components of estimation using simulation-free approach. Equation (1) *yields* $\hat{\theta} = 0.1123$, *and substituting estimates into equation* (9) *yields* $se(\hat{\theta}) = 0.0177$. *Assuming* $\hat{\theta}$ *is log-normally distributed, an approximate* 95% *confidence interval is* $(0.0826, 0.1526)$.

Estimator	Weight w_k	$\hat{\theta}_k = \hat{f}(0)$	$se(\hat{\theta}_k \mid \theta_k)$	$\hat{\beta}_k = \hat{\theta}_k - \hat{\theta}$
1	0.283	0.1032	0.0066	−0.0091
2	0.236	0.0955	0.0081	−0.0168
3	0.109	0.1010	0.0063	−0.0113
4	0.372	0.1333	0.0103	0.0210

4.3 *A Time-Dependent Survival Model*

In the above two examples, truth is unknown, making comparison of different approaches problematic. To explore model selection under known truth, we implemented in SAS Monte Carlo data generation for a simple cohort survival process such as could be done with radio-tracked animals or such as arises from a certain type of life-table data on animals. These interpretations are not needed, but we will use the terminology of an animal survival cohort process followed over years.

Let n_1 same-age, homogeneous animals initiate the cohort at year 1, of which n_2 survive for 1 year. In general, it is known that n_i animals are alive at the start of year i. All n_i are counted without error. A chain binomial model is used: for each i, n_{i+1} given n_i is an independent binomial(n_i, S_i) random variable, for $i = 1, \ldots, m$. An equivalent, alternate representation is that $n_2, \ldots, n_m$ are multinomial given n_1 with cell probabilities as functions of the S_i. Because of this equivalence, we take the sample size to be n_1.

The parameters of the simulation model are $S_1, \ldots, S_m$; they can be all different (for a complex true model) or can be constrained so that a simple true model applies. Conceptually, the model can be parameterized in terms of time-sequential effects. Of special interest is a model with many effects but of diminishing size (e.g., $S_i = 0.8 - 0.3/i$). For data analysis, we approximate truth with simpler models of the form $S_i = S_1, i \geq 1$ (model 1); $S_1, S_i = S_2, i \geq 2$ (model 2); and in general the kth model (having k parameters) has $S_1, \ldots, S_{k-1}, S_i = S_k, i \geq k$. The advantage of these models for data analysis is speed of simulation because the MLEs are closed form.

We define the partial sums $n_{i,+} = n_i + \cdots + n_m$. For model k, $n_{i+1,+}$ given $n_{i,+}$ is binomial$(n_{i,+}, S_k)$. For model k, MLEs are $\hat{S}_i = n_{i+1}/n_i, i = 1, \ldots, k-1$ and $\hat{S}_k = n_{i+1,+}/n_{i,+}$. The log-likelihood for the kth model is the sum of k binomial likelihoods, i.e.,

$$\begin{aligned} \log(L_k) = &\sum_{i=1}^{k-1} \left[n_i \log(\hat{S}_i) + (n_i - n_{i+1}) \log(1 - \hat{S}_i) \right] \\ &+ n_{k,+} \log(\hat{S}_k) + (n_{k,+} - n_{k+1,+}) \log(1 - \hat{S}_k) \end{aligned} \tag{17}$$

(bearing in mind that $0 \log(0) = 0$). It was also convenient to specify a maximal model (i.e., value of k) to consider for data analysis. These are nested models. In data analysis under model k, survival rates for all ages greater than k are estimated as $\hat{S}_k$. We only tabulated and examined survival rate estimators out to age 10.

Age-specific survival rates change appreciably with age (time) at the younger ages, with only small annual changes at older ages. Due to mortality, data are sparse at the older ages, rendering identification of the true model from the data an unachievable goal. Even the most general true model allowed here is simpler than reality, wherein we expect possibly complex age (e.g., senescence) and time effects (nonsmooth time effects and possibly disasters) and individual heterogeneity in survival fitness. Thus, truth in this simulation is simpler than reality is likely to be, but the true generating model here can be sufficiently high dimensional to avoid the pitfall of many simulation studies: (1) there is a simple (low dimension), true model, which (2) is included as one of the contending models and (3) the goal is to select that true model. Studies of this type favor (at least for large sample sizes) BIC over AIC because BIC is designed for this circumstance. For real data analysis in life sciences we contend that no simple true model generated the data and the statistical goal is to select the best approximating model (to conceptual truth) from the set of models considered.

For our simulation, we specified n_1, the S_i, m, and a maximal model (max-k) to consider. A thousand data sets were simulated, and each model (1 to max-k) was fitted by maximum likelihood

Table 6

Relative penalized likelihoods for four line transect models fitted to ruffed grouse data. Likelihoods were penalized using the Bayes Information Criterion (BIC). Also given is the proportion of times each model was selected by BIC when analyzing 1000 *bootstrap resamples (a) allowing the number of adjustment terms used for a particular model to vary between resamples and (b) fixing the number of adjustment terms to equal that determined in the analysis of the original data.*

k	Estimator key	Series	BIC	$w_k = \exp(-\text{BIC}/2) \div \Sigma \exp(-\text{BIC}/2)$	Proportion of bootstrap resamples (a)	Proportion of bootstrap resamples (b)
1	Half-normal	Cosine	595.27	0.111	0.138 (0.011)	0.084 (0.009)
2	Hazard-rate	Polynomial	595.63	0.092	0.093 (0.009)	0.105 (0.010)
3	Uniform	Cosine	600.56	0.008	0.001 (0.001)	0.000 (0.000)
4	Negative exp.	Polynomial	591.34	0.789	0.768 (0.013)	0.811 (0.012)

to each data set. Two model selection criteria were used: BIC and AIC_c. It is beyond the scope of this paper to present extensive simulation results. We have done many runs with different models. However, early on we selected one case to use as an example here. It was selected prior to seeing the results, but turns out to be fairly representative of results under other simulation scenarios. The example used $n_1 = 150$, $m = 30$, max-$k = 10$, and $S_1 = 0.5$, $S_2 = 0.7$, $S_3 = 0.75$, $S_4 = 0.8$, $S_5 = 0.8$, and thereafter S_i declines at 2% a year (corresponding to decreasing survival rates with age). The probability of surviving from time 1 through year 30 is 0.893×10^{-6}.

The following conclusions can be drawn from the results of Tables 7 and 8. In this example, model averaging using equation (1) and AIC_c weights provides estimates of parameters with marginally lower bias than does the best approximating model identified by AIC_c. Using BIC to identify the best model, bias is appreciably higher. Coverage is well below the nominal level if model selection uncertainty is ignored. Using AIC_c, this coverage was 82%, and BIC, which tended to select overly simplistic models, had coverage of 78%. By contrast, equation (9), with weights calculated from AIC_c according to equation (14), achieves the nominal coverage of 95% to a good approximation (94.9%).

Table 7

Number of simulations out of 1000 *in which each survival model was selected by AIC_c and by BIC. Model* 1 *assumes a constant survival rate across the full* 10 *years, model* 2 *assumes a different survival rate in year* 1 *from subsequent years, through to model* 10 *(the true model), in which survival is different for each year. AIC_c weights were calculated from each simulation using equation* (14) *and averaged.*

Model	Frequency that model was selected: AIC_c	Frequency that model was selected: BIC	Average AIC_c weights
1	0	3	0.001
2	496	872	0.291
3	234	111	0.234
4	63	9	0.136
5	60	2	0.097
6	52	3	0.078
7	29	0	0.058
8	32	0	0.045
9	18	0	0.033
10	16	0	0.027

Table 8

Mean survival estimates from 1000 *simulations corresponding to model selected by* AIC_c ($\bar{s}_a$) *and by BIC* ($\bar{s}_b$), *and the mean of the model-averaged estimates obtained from equation* (1) ($\bar{s}_w$). *Also shown is the proportion of confidence intervals* (*calculated as estimate* $\pm$ 1.96 *standard errors*) *that cover the true parameter value for* (*a*) *the model selected by* AIC_c (p_a), (*b*) *the interval obtained from the variance of equation* (9), *using* AIC_c *weights* (p_w), *and* (*c*) *the model selected by BIC* (p_b).

Year (age)	True survival rate	$\bar{s}_a$	$\bar{s}_b$	$\bar{s}_w$	p_a	p_w	p_b
1	0.500	0.499	0.499	0.501	0.955	0.955	0.955
2	0.700	0.711	0.733	0.700	0.729	0.957	0.565
3	0.750	0.754	0.752	0.751	0.908	0.957	0.944
4	0.800	0.775	0.752	0.777	0.663	0.926	0.539
5	0.800	0.771	0.752	0.773	0.655	0.905	0.540
6	0.784	0.763	0.751	0.762	0.817	0.960	0.772
7	0.768	0.756	0.751	0.754	0.907	0.977	0.921
8	0.753	0.749	0.751	0.748	0.928	0.982	0.953
9	0.738	0.748	0.751	0.743	0.862	0.964	0.893
10	0.723	0.743	0.751	0.739	0.753	0.905	0.763
All					0.818	0.949	0.784

5. Discussion

Raftery, Madigan, and Hoeting (1993) and Draper (1995) note that traditional methods of model selection can lead to models that appear to have strong predictive power even for randomly generated data for which there is no true relationship. They show that allowance for model selection uncertainty using a Bayesian framework can resolve this difficulty. We have presented ideas for likelihood-based solutions to the problem that avoid the complexities of Bayesian methods and allow philosophies other than the usual philosophy of Bayesian researchers—that one of the fitted models is the true model. We hope that these ideas will stimulate others to address the issue of model selection uncertainty. The widespread practice of using sophisticated model selection methods, followed by inference that ignores the uncertainty in those methods, is inconsistent and too often leads to overoptimistic estimates of precision.

Given the potential impact on conclusions of allowing for model selection in inference, it seems surprising that more authors have not addressed this issue. In some fields, it would seem essential that the issue be addressed. One example is the use of mark–recapture models in epidemiology, where numbers of people suffering from a disease are estimated from data from lists of sufferers. The lists are treated as captures, with many people appearing on more than one list, corresponding to multiple recaptures. The problem is to estimate the number of sufferers appearing on no list. Heterogeneity and lack of independence are fundamental difficulties; the presence of sufferers on a list may depend heavily on where they live, on income, or on whether they appear on another list. Different plausible models may fit the data equally well, but give rise to confidence intervals on total number of sufferers that do not overlap. Hook and Regal (1995) review the methods and their limitations, and Madigan and York (1995) use Bayesian methods to average across models in this circumstance.

To defend the policy of carrying out inference conditional on the selected model, the analyst must be willing to affirm that the true model has been identified. If the true model were known, then model selection would not be required. However, there are cases where the range of models could, and should, be restricted. In the example of mark–recapture models in epidemiology, we can rule out models that allow for trap response (useful for analyzing data from experiments in which animals are trapped), as the lists cannot usually be ordered chronologically. In our line transect example, we included the negative exponential model, which is spiked at zero distance from the trackline (Figure 3). For surveys that are designed well, such spiked models can be ruled out, and, in the case of our example, this reduces considerably the uncertainty in estimating $f(0)$. The negative exponential model was included for illustrative purposes. Nevertheless, Figure 4 shows that the bootstrap distribution of $f(0)$ estimates differs appreciably among the three remaining models, with the hazard-rate distribution strongly skewed to the left and the other two skewed to

the right. Estimation is appreciably more uncertain than would appear if any single model were selected.

If knowledge of appropriateness of models was extensive, prior weights could be assigned and Bayesian methods could be used to downweight models that are intrinsically less plausible, such as the negative exponential line transect model. By analogy with Bayesian methods, if we denote the prior 'probability' that model M_i is true by $P(M_i)$, then both sides of equation (13) would be multiplied by the prior odds $P(M_i)/P(M_j)$. Equation (14) then becomes

$$w_k = \frac{P(M_k)\exp(-I_k/2)}{\sum_{i=1}^{K} P(M_i)\exp(-I_i/2)}, \qquad k = 1, \ldots, K \tag{18}$$

(see Kishino et al., 1991). Thus, for example, if we chose to assign a prior weight of 0.1 to the negative exponential model in our line transect example and a weight of 0.3 to each of the other three models, the weights for estimators 1–4 of Table 5 would become 0.376, 0.314, 0.145, and 0.165, respectively, so that equation (1) yields a revised estimate of 0.1054, lower than the previous estimate of 0.1123 due to the downweighting of the spiked negative exponential model. Variance and confidence interval estimation would follow as described in the section on simulated inference, using the revised weights.

We advocate the use of an information criterion to allow evaluation of a relative penalized likelihood factor, but we have avoided recommending a single criterion. Many authors advocate use of a criterion such as BIC, which is dimensionally consistent. For such criteria, the penalty increases as sample size increases. However, Burnham, Anderson, and White (1994) argue that dimensional consistency is only useful when there exists a relatively small-dimensional true model, a circumstance that rarely occurs in the real world. They favor use of AIC, in which the penalty, when divided by -2, may be viewed as a bias adjustment term for the expected log likelihood. This interpretation, due to Akaike (1973) and explored in detail by Bozdogan (1987), provides justifcation for use of the relative penalized likelihood factor when AIC is used. Kishino et al. (1991) used this strategy to estimate posterior probabilities of alternative models, given prior probabilities.

We have considered two main strategies of averaging across models that lead to different inference. In one, we assume that there is an infinity of models to choose from and those fitted are a random sample. The bias arising from any single model then becomes the contribution of model selection uncertainty to variance of the weighted estimator. The assumption that models are essentially selected at random from a large population of models is often poor. It is likely to be an acceptable assumption in our line transect example, for which each model had a parametric key that was dissimilar from the others, but the assumption clearly fails in the Poisson regression, as the models with one covariate are both special cases of the model with both covariates and cannot be considered to be independent of it. Another shortcoming of this approach is that estimators of the common parameter of interest under the different models cannot be assumed independent when they are estimated from the same data. As an alternative, we adopt the philosophy that we are seeking to identify the best approximating model from a set of competing models. We recognize that identification of this model is an estimation problem by assigning a weight for each model in proportion to its relative penalized likelihood. The weight for model k is estimated by observing the proportion of bootstrap samples in which that model was identified as the best approximating model. Under this philosophy, we require that the set of contending models includes at least one that approximates well the true model, which might be complex and high dimensional.

Acknowledgements

We are grateful to the referees, whose comments led to a substantially improved manuscript.

Résumé

Nous prétendons que l'incertitude sur le choix de modèle devrait être intégrée à l'inférence statistique chaque fois que l'estimation est sensible au choix du modèle, et que ce choix est fait à partir des données. Nous considérons différentes philosophies pour atteindre ce but, et suggérons des stratégies pour l'analyse des données. Nous illustrons nos méthodes avec trois exemples. Le premier est une régression avec une loi de Poisson sur des comptages d'oiseaux, dans lequel on doit faire un choix entre introduire une ou deux covariables. Le second est un ensemble de points sur un transect pour lequel différents modèles donnent des estimations d'abondance sensiblement différentes. Le troisième est un exemple simulé pour lequel on connaît la vérité.

References

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

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

Berger, J. O. and Pericchi, L. R. (1996). The intrinsic Bayes factor for model selection and prediction. *Journal of the American Statistical Association* **91,** 109–122.

Berk, R. H. (1966). The limiting behavior of posterior distributions when the model is incorrect. *Annals of Mathematical Statistics* **37,** 51–58.

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

Bravington, M. (1993). The effects of acidification on the population dynamics of brown trout in Norway. Ph.D. thesis, Imperial College, London.

Buckland, S. T. (1982). A note on the Fourier series model for analysing line transect data. *Biometrics* **38,** 469–477.

Buckland, S. T. (1984). Monte Carlo confidence intervals. *Biometrics* **40,** 811–817.

Buckland, S. T. (1992). Fitting density functions using polynomials. *Applied Statistics* **41,** 63–76.

Buckland, S. T., Anderson, D. R., Burnham, K. P., and Laake, J. L. (1993). *Distance Sampling: Estimating Abundance of Biological Populations.* London: Chapman and Hall.

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,* D. R. McCullough and R. H. Barrett (eds), 16–30. London: Elsevier Science Publishers.

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.

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

Draper, D. (1995). Assessment and propagation of model uncertainty (with discussion). *Journal of the Royal Statistical Society, Series B* **57,** 45–97.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *Annals of Statistics* **7,** 1–26.

Efron, B. and Tibshirani, R. J. (1993). *An Introduction to the Bootstrap.* New York: Chapman and Hall.

Gates, C. E. (1979). Line transect and related issues. In *Sampling Biological Populations,* R. M. Cormack, G. P. Patil, and D. S. Robson (eds), 71–154. Fairland, Maryland: International Cooperative Publishing House.

Hjorth, J. S. U. (1994). *Computer Intensive Statistical Methods: Validation, Model Selection and Bootstrap.* London: Chapman and Hall.

Hook, E. B. and Regal, R. R. (1995). Capture–recapture methods in epidemiology—Methods and limitations. *Epidemiologic Reviews* **17,** 243–264.

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

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

Kass, R. E. and Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association* **90,** 773–795.

Kishino, H., Kato, H., Kasamatsu, F., and Fujise, Y. (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.

Laake, J. L., Buckland, S. T., Anderson, D. R., and Burnham, K. P. (1993). *DISTANCE User's Guide,* Version 2.0. Fort Collins: Colorado Cooperative Fish and Wildlife Research Unit.

Madigan, D. M. and York, J. C. (1995). Bayesian graphical models for discrete data. *International Statistical Review* **63,** 215–232.

Madigan, D. M., Raftery, A. E., York, J. C., Bradshaw, J. M., and Almond, R. G. (1994). Strategies for graphical model selection. In *Selecting Models from Data: AI and Statistics IV,* P. Cheeseman and R. W. Oldford (eds), *Lecture Notes in Statistics* **89,** 91–100. New York: Springer-Verlag.

Norris, J. L., III and Pollock, K. H. (1996). Including model uncertainty in estimating variances in multiple capture studies. *Environmental and Ecological Statistics,* **3,** 235–244.

Raftery, A. E. (1996). Approximate Bayes factors and accounting for model uncertainty in generalized linear models. *Biometrika* **83,** 251–266.

Raftery, A. E., Madigan, D. M., and Hoeting, J. (1993). *Model selection and accounting for model uncertainty in linear regression models.* Technical Report 262, Department of Statistics, University of Washington, Seattle.

Rissanan, J. (1989). *Stochastic Complexity in Statistical Inquiry, Series in Computer Science,* Volume 15. London: World Scientific.

Sakamoto, Y., Ishiguro, M., and Kitagawa, G. (1986). *Akaike Information Criterion Statistics.* Tokyo: KTK Scientific Publishers.

Sauerbrei, W. and Schumacher, M. (1992). A bootstrap resampling procedure for model building: Application to the Cox regression model. *Statistics in Medicine* **11,** 2093–2109.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics* **6,** 461–464.

Sclove, S. L. (1987). Application of model selection criteria to some problems in multivariate analysis. *Psychometrika* **52,** 333–343.

Shibata, R. (1989). Statistical aspects of model selection. In *From Data to Model,* J. C. Willems (ed), 215–240. London: Springer-Verlag.

Silverman, B. W. (1982). Algorithm AS 176: Kernel density estimation using the fast Fourier transform. *Applied Statistics* **31,** 93–99.

Received April 1995; *revised September* 1996; *accepted November* 1996.