MODEL AVERAGING IN AGRICULTURE AND NATURAL RESOURCES:
WHAT IS IT? WHEN IS IT USEFUL? WHEN IS IT A DISTRACTION?

Philip M. Dixon
Department of Statistics,
Iowa State University,
Ames IA, 50011

## Abstract:

I use two examples to illustrate three methods for model averaging: using AIC weights, using BIC weights, and fully Bayesian analyses. The first example is a capture-recapture study that estimates the population size by averaging over 4 models for capture probabilities. The second is an analysis of a study of logging impacts on Curculionid weevils using a before-after-control-impact (BACI) study design. The estimated impact is averaged over 4 ecologically relevant models.

Both examples demonstrate the sensitivity of model weights, or posterior model probabilities, to the choice of prior model probabilities and prior distributions for parameters. The model averaged estimates and their confidence intervals are less influenced by those choices. The BACI-design example also demonstrates the need to carefully choose the model parameterization so that the parameter of interest, the interaction, has the same interpretation for all models in the model set. I also briefly discuss three other frequentist approaches to model averaging: bagging, stacking, and model-averaged-tail-area confidence intervals.

Keywords:
AIC, BIC, Bayesian model averaging, capture-recapture, BACI design, model weights, posterior model probabilities, MATA confidence intervals

# 1 Introduction

A statistical analysis relies on a model. Commonly, multiple models are possible. Some examples of model choices include:

- predictive modeling of observational data. It is well-known that using an appropriately chosen subset of the possible "$X$" variables provides more precise predictions. The question is then choosing an appropriate subset.
- using propensity score matching to reduce bias in estimated treatment effects due to confounding (Guo and Fraser 2010). The propensity score model predicts the probability that a unit receives the treatment instead of a control. What set of possible "$X$" variables should be included in that propensity score model?
- including baseline covariates in a randomized experiment. Although randomization guarantees unbiased estimates of treatment means, including baseline covariates may increase the precision of those estimates. Again, the question is which covariates to include.
- estimating population size using mark-recapture methods. The goal is to estimate $N$, but doing that requires modeling the detection process (Otis et al. 1978). The question is which detection model is most appropriate for the population under study.
- modeling variances in a designed experiment. The study design may specify the model for the expected values, but what model should be used for the variances? Homoscedastic? Dependent on the mean? Unstructured heteroscedastic?
- interpolating spatial data. Using kriging requires choosing a model for the covariance between pairs of observations. There are many different possible models. The question is which is the most appropriate.
- choosing a variance-covariance matrix for repeated measurements on the same subjects. There are many possible models for repeated measures data (Diggle et al. 2002). Which one is most appropriate?
- for any of the above situations, what distribution should be assumed for the observations?

Often, choices are made by default. The model for continuous data from a randomized experiments frequently assumes equal variances and normally distributed errors. It is good data analysis practice to then use diagnostic tools to assess the adequacy of these choices.

An alternative is to use the data to select a single model. This can be done formally by using a statistic such as AIC or BIC to choose variables, variance structures, spatial correlation models, or distributions. Or, it can be done informally by using graphical diagnostics such as different types of residual plots to assess preliminary model choices. Once all the choices have been made, the resulting model is treated as if it were known *a-priori*. The fact that the data were used to choose that model is ignored. Breiman (1992) has called this a "quiet scandal in the statistical community". He continues, "it is clear that selecting a sequence of submodels in terms of an optimum or suboptimum fit to the data can produce severe biases in all statistical measures used for the classical linear model." (Breiman 1992, p. 738).

Model averaging (MA) provides a mechanism to avoid the choice of a single model. Multiple models are fit to a data set and the results combined. The benefits of model averaging can be viewed from two complementary perspectives:

- The MA estimator of a parameter is often less biased than a single model estimator. This is because estimators from different models are likely to have different biases. If some biases have opposite signs, the bias of the averaged estimate is the average of the biases. This is often closer to 0.
- MA accounts for the uncertainty in the choice of model. The variance of an MA estimator is usually larger than the variance from a single model, because MA accounts for the heterogeneity of estimates across models.

An analogy due to Ripley (2004) provides a useful comparison between selecting a single model and model averaging. Imagine you have a large panel of experts, each of whom has provided an estimate. What is the best way to use that collection of estimates? You could decide on the expert you trust the most and adopt their estimate and ignore all others. That is selecting a single model. Or, you could seek a consensus estimate that combines all the estimates. That is model averaging. If you could consistently identify the most accurate expert, choosing their single estimate is the best approach. In many empirical studies, it is hard to identify the most accurate estimate. In those situations, a consensus estimate turns out to be more accurate.

Model averaging is not a new idea. One early example is forecasting future observations in a time series of airline passenger counts (Bates and Granger 1969). Applications of model averaging in agriculture and natural resources include yield prediction (Huang et al. 2017), forecasting precipitation (Kleiber et al. 2011), and yield-gap analysis of the factors limiting crop yield (Prost et al., 2008). Modern computing power and software have made MA more feasible.

The MA estimator of some quantity, $\hat{\theta}_{MA}$, is very simple. Consider multiple models, $M_k$, $k = 1, 2, \ldots, K$, where $K$ is the number of models under consideration. All of the models provide an estimate of the quantity of interest, $\theta_k$. Fitting all models to the data gives you $K$ estimates of $\theta$: $\hat{\theta}_k$, $k = 1, 2, \ldots, K$. Associated with each model is a model weight, $w_k$. Model weights are non-negative and sum to 1, i.e. $w_k \geq 0 \, \forall k$ and $\sum_k w_k = 1$. The MA estimator is then

$$\hat{\theta}_{MA} = \sum_k w_k \hat{\theta}_k. \tag{1}$$

Everything else about model averaging is just the details. Three of the important details are:

- How do you choose the model weights, $w_k$?
- How do you estimate the precision of $\hat{\theta}_{MA}$ or construct an interval estimate from $\hat{\theta}_{MA}$?
- Should you take a frequentist or Bayesian approach to model averaging?

The literature on these details is extensive. Fletcher (2018) provides a summary of MA methods and a thorough overview of the literature. Dormann et al. (2018) review the use of model averaging in ecology. I will use two examples to provide an introduction to model averaging methods.

# 2 Examples

I use data from a mark-recapture study of eastern chipmunks and a before-after-control-impact (BACI) study of logging impacts on Chironomid weevils to motivate, describe, and illustrate model averaging. R code, BUGS code and both data files are included in the supplemental material (to be provided soon).

The chipmunk data set is based on Mares et al. (1981) experimental introduction of 85 eastern chipmunks to an island in Pymatuning Reservoir, Pennsylvania. Prior to the introduction, there were neither chipmunks nor predators on the island. A mark-recapture study with 194 traps on a regular grid across the island was set up. Traps were checked once or twice a day for a total of 13 capture occasions over 8 days. Mares et al. used these data to compare the accuracy of Lincoln-Peterson and related estimators of population size. To emphasize the consequences of model averaging, I removed data for 3 capture occasions with especially low capture probability. The result is data from 10 capture occasions; 71 animals are seen at least once. Three animals died or were removed from the island, so the population size during the mark-recapture sampling was 82 animals.

Mares et al. (1981) report the numbers of tagged animals and total animals caught each trapping occasion and the frequency distribution of the number of recaptures. I simulated capture histories consistent with these summaries. One difference is that Mares et al. suggest that there were two subpopulations with different capture probabilities. My simulated data assumed no heterogeneity.

The second example is based on a study of the impact of logging a tract of tropical forest on the abundance of many species of herbivores (Basset et al. 2001). My example is based on the data for Chironomid weevils. Two tracts of tropical forest were delineated. The number of weevils was counted in each tract monthly for 11 months (the "before" data). One tract was randomly chosen to be logged, while the other was left undisturbed. The number of weevils were again counted monthly for 11 months (the "after" data). Basset et al. (2001) provided the means and standard errors for each of the four groups of samples from which I recreated the data set. The monthly counts were independent Poisson samples constrained to match the means provided by Basset et al. (2001). Because there is no replication of the logging treatment, logging status is confounded with the tract. This is typical when BACI designs are used to assess the environmental impact of a facility or point source of pollution. One advantage of the BACI design is that the "before" data helps control for pre-existing differences between tracts. Stewart-Oaten and Bence (2001) and Underwood (1994) provide more background on BACI designs and related approaches to assess environmental impact.

# 3 Model selection using the chipmunk data

The goal of the chipmunk data analysis is to estimate the population size. The population is sampled by 10 capture occasions over 8 days, so it is reasonable to assume closure, i.e., no births, deaths, immigrants, or emigrants. Because of imperfect detection (not all animals are trapped on a capture occasion), the true population size is likely to be larger than the number of animals seen at least once. The classic approach to estimate the size of a closed population is to fit one of the Otis et al. (1978) models. Different models correspond to different assumptions about detection probabilities. Model $M_0$ assumes that every animal has the same capture probability on all sampling occasions. Model $M_t$ assumes that the capture probability varies between sampling occasions (times) but each animal has the same capture probability on a sampling occasion. Model $M_b$ models a behavioral response known as "trap-happiness" or "trap-shyness". That is each animal has one capture probability until the first time they are captured; subsequent times have a different recapture probability. Model $M_{tb}$ includes both time-varying capture probabilities and a behavioral response. In its most general form, this model is overparameterized. A common approach is to assume a logit-additive model for the capture probability. That is, logit $p_{ij} = \mu + t_j + cX_{ij}$, where $p_{ij}$ is the capture probability for animal $i$ on occasion $j$, $\mu$ and $t_i$ model the time-specific probability of first capture, $X_{ij}$ is an indicator variable with the value of 1 if animal $i$ has been captured prior to occasion $j$ and 0 otherwise. The parameter $c$ describes the behavioral change in capture probabilities. Otis et al. (1978) proposed 4 additional models that include heterogeneity of capture probability between individuals. I do not consider any of the heterogeneity models.

All four Otis models can be fit by maximum likelihood. The estimated population sizes, $\hat{N}$, from each model are given in Table 1. Although similar, they are not identical. Which value should be reported? The current standard approach is to use model selection to identify the best model.

| Model | # param | AICc | $\Delta$ AICc | $\hat{N}$ | se |
|---|---|---|---|---|---|
| Mtb | 12 | 395.5 | 0 | 76.7 | 5.3 |
| Mt | 11 | 396.3 | 0.79 | 72.5 | 1.6 |
| Mb | 3 | 396.6 | 1.08 | 78.1 | 4.9 |
| M0 | 2 | 402.1 | 6.57 | 72.7 | 1.6 |

Table 1: Number of parameters, model fit statistics, estimated population size $\hat{N}$ and the standard error of $\hat{N}$ for each of the four Otis models fit to the chipmunk data. Models are sorted from best (smallest AICc value) to worst fit.

The model selection approach (Burnham and Anderson 2002) is to identify a set of biologically reasonable models, fit each to the data, compute a model fit statistic for each model, then choose the best model. Inferences for parameters such as $N$ are then conditional on that choice of model. For my analysis of the chipmunk data, the model set is the four models, $M_0$, $M_t$, $M_b$, and $M_{tb}$. Two frequently used model fit statistics are the Akaike Information Criterion (AIC) and its small-sample corrected modification, AICc. Both are computed from the log likelihood

evaluated at the maximum likelihood estimates of the parameters, $\log L$, and the number of model parameters, $p$. AIC is:

$$\text{AIC} = -2 \log L + 2p. \tag{2}$$

AICc also depends on the number of observations $n$.

$$\text{AICc} = \text{AIC} + \frac{2p(p+1)}{n-p-1}.$$

Both can be viewed as measure of the fit of the model to the data, quantified by the deviance $= -2 \log L$, with a penalty for the complexity of the model. For AIC, that penalty is $2p$. The AICc penalty is slightly larger. When the number of observations is large relative to the number of parameters in the model, the difference between the two statistics is small. AICc values for the four models are given in Table 1.

The best model is the one with the smallest AIC or AICc value. Here, that is model $M_{tb}$ (Table 1, although two other models have very similar AICc values. Since AICc is a statistic computed from the data, it is subject to sampling variability. A new sample of data may indicate a different best model. Burnham and Anderson (2002) suggest that models with AICc values within 2 units (or 4 units, Burnham and Anderson 2004) of the best model are possible alternatives to the best model and models with AICc values more than 10 units from the best model can be ruled out as implausible. Using these guidelines, models $M_t$ and $M_b$ are possible alternatives.

Model selection is often complemented by a model sensitivity analysis. Results from other reasonable models are reported along with those from the best model. Applying this approach to the chipmunk data, you would report an estimated population size (standard error) of 76.7 (5.3) from model $M_{tb}$ along with results from model $M_t$: 72.5 (1.6) and $M_b$: 78.1 (4.9). Each of these estimates assumes that the named model is the model that generated the data. Instead of reporting multiple results, model averaging will provide a single estimate with an uncertainty that accounts for the choice of model.

# 4 Introduction to model averaging, using the chipmunk data

A model averaged estimate is a weighted average of the model-specific estimates from multiple models. The weights for each model are the $w_k$ coefficients in Equation (1). There are three general approaches to determining those model weights (Fletcher 2018):

1. The probability that a model is the "true" model.
   This approach assumes that the "true" model, the one that generated the data, is one of the models in the evaluation set. In practice, the true model can be relaxed to be an approximation to the true model. The model weights quantify the probability that a particular model is the true model or its approximation. This approach is closely related to

the use of Bayesian estimates of posterior model probabilities or the BIC statistic (defined below).

2. The out-of-sample prediction error for a model.
   This approach focuses on estimation and prediction of parameters or responses. The model weights reflect the accuracy of out-of-sample predictions. That out-of-sample prediction error can be estimated either by in-sample prediction error penalized for model complexity or out-of-sample prediction error. The model set need not include the true model. The goal of model averaging is to obtain more accurate estimates by trading off bias and variance. This approach is closely related to the use of the AIC, AICc, or cross-validation.
3. Targeted criteria that focus on some other important aspect of a model.
   Using out-of-sample prediction error or the model probability as the criterion implicitly considers all parameters in a model. An alternative is to focus on specific parameters or linear combinations of them. Model weights are based on a focused information criterion that targets that specific aspect of the model (Claeskens and Hjort (2008, pp 145 et seq.). If multiple aspects are relevant, different model weights will be used for each target. Targeted approaches will not be discussed here. Further information can be found in Claeskens and Hjort (2003) and Claeskens and Hjort (2008); an application is described in Yang et al. (2015).

Each of these approaches can be implemented in a frequentist manner or a Bayesian manner.

## 4.1 Frequentist model averaging

Burnham and Anderson (2002) recommend frequentist model averaging using model weights calculated from model AIC or AICc statistics. At least for now, this approach is the most commonly used model averaging method in wildlife research. I illustrate this approach using the chipmunk study. The AICc estimated weight for model $k$ in a model set of $K$ models is defined as

$$w_k = \frac{\exp(-\Delta\text{AICc}_k/2)}{\sum_{i=1\cdots K}\exp(-\Delta\text{AICc}_i/2)}. \tag{3}$$

The quantity $\Delta\text{AICc}_i$ is the difference in AICc statistics between the best model in the model set, i.e., the one with the lowest AICc statistic, and the AICc statistic for model $i$. The AIC weight is similar, except using $\text{AIC}_k$ instead of $\text{AICc}_k$. Either set of weights sums to 1 because of the denominator. Table 2 shows the AIC weights for the four models fit to the chipmunk data. We see that the largest weight is given to model Mtb that has the smallest AICc statistic. The two models, Mt and Mb, with AICc statistics within 2 units of model Mtb have appreciable model weight. The weight given to a model declines as its AICc statistic is further from that of the best model, so model M0 with a moderately large $\Delta$ AICc has a small model weight. The model averaged estimate of $\hat{N}$ is $\hat{N}_{MA} = \sum_{k=1,2,3,4} w_k\,(\hat{N}_k \mid M_k) = 75.8$. In this case, I am explicitly indicating the dependence of $\hat{N}_k$ on the model, $M_k$.

Calculating a standard error or a confidence interval for frequentist model averaging is difficult for two reasons (Hjort and Claeskens 2003). The sampling distribution of $\hat{N}_{MA}$ is a mixture

| Model | $\Delta$ AICc$_k$ | $w_k$ | $\hat{N}_k$ | se |
|---|---|---|---|---|
| Mtb | 0 | 0.436 | 76.7 | 5.3 |
| Mt | 0.79 | 0.294 | 72.5 | 1.6 |
| Mb | 1.08 | 0.254 | 78.1 | 4.9 |
| M0 | 6.57 | 0.016 | 72.7 | 1.6 |

Table 2: Model weights, AICc statistics, $w_i$, and estimated population size $\hat{N}_i$ and the standard error of $\hat{N}_i$ for each of the four Otis models fit to the chipmunk data. Models are sorted from best (smallest AICc value) to worst fit.

of the model-specific sampling distributions, with mixture proportions given by the estimated model weights. And, the estimates from different models are almost always correlated with a usually unknown correlation structure.

There are two standard error estimators in common use. Both combine within-model uncertainty and between-model heterogeneity of the estimates. The first, Burnham and Anderson (2004)'s "revised estimator" ignores the correlation between estimators and calculates

$$\text{se}_1 = \sqrt{\sum_{k=1}^{K} w_k \left[\widehat{\text{var}}(\hat{N}_k \mid m_k) + (\hat{N}_k \mid m_k - \hat{N}_{MA})^2\right]}.$$

This averages the variances and contributions to heterogeneity, then takes the square root to convert a variance to a standard error. The alternative, proposed by Buckland et al. (1997),

$$\text{se}_2 = \sum_{k=1}^{K} w_k \sqrt{\hat{\text{var}}(\hat{N}_k \mid m_k) + (\hat{N}_k \mid m_k - \hat{N}_{MA})^2}.$$

This averages the standard errors. Buckland et al. motivate their estimator as an ad-hoc correction for the correlation among estimates. For the chipmunk data, the two standard error estimates of $\hat{N}_{MA}$ have very similar values: 4.40 for the revised se estimator and 4.48 for the Buckland estimator. An extensive set of simulations by Burnham and Anderson (2004) suggests the two estimators frequently have very similar values.

Calculating an appropriate confidence interval for a model-averaged estimate is even more troublesome than calculating the standard error. The problem is that the sampling distribution of $\hat{N}_{MA}$ is a mixture distribution. If the model-specific estimates are maximum likelihood estimates, they have asymptotic normal distributions. The model-averaged estimate may have a sampling distribution that is a mixture of normal distributions, which could be skewed or multimodal, depending on the model-specific estimates and model weights. This suggests that the empirical coverage of Wald-style confidence intervals may be far from nominal. The currently best available confidence interval estimator is the model-averaged-tail-area (MATA) estimator (Fletcher and Turek 2011, Fletcher and Turek 2012). Given a cumulative sampling distribution, $F_{\hat{N}}(x)$ for a model-specific estimate $\hat{N}$, the lower bound of a $1-\alpha$ two-sided equal-tailed model-specific

| Type | Interval | 95% CI |
|---|---|---|
| Model averaged | MATA-Wald | (72.4, 100.1) |
| Model averaged | Wald | (72.5, 107.9) |
| Model-specific | Wald Mtb | (74.0, 174.0) |
| Model-specific | Wald Mt | (72.6, 102.1) |
| Model-specific | Wald Mb | (74.0, 172.9) |
| Model-specific | Wald M0 | (72.6, 103.7) |

Table 3: 95% confidence intervals for the model-averaged estimate of number of chipmunks, computed using either the MATA method or Wald intervals, and the four model-specific Wald confidence intervals.

confidence interval is the value, $x_l$, for which $F_{\hat{N}}(x_l) = \alpha/2$. Turek and Fletcher (2011) extend this to a model-averaged estimate by considering the weighted average of the lower tail probabilities. The lower confidence bound of a two-sided equal-tailed $1 - \alpha$ confidence interval is the value, $x_l$, that solves

$$\sum_{k=1}^{K} w_k \, F_{\hat{N}_k}(x_l) = \alpha/2.$$

This approach can be used with any valid cumulative sampling distribution. Turek and Fletcher (2012) give the name "MATA-Wald interval" to MATA intervals computed from normal or T sampling distributions; Fletcher and Turek (2011) give the name "MATA-profile interval" to MATA intervals computed from profile likelihood statistics. Simulation-based comparisons of the coverage of the MATA intervals and various alternatives show that the MATA intervals have coverages closer to nominal than do the alternatives. The full confidence distribution can be estimated by computing model-averaged tail probabilities for a sequence of $x$ values (Fletcher et al. 2019).

For the chipmunk data, the AICc-selected model, Mtb, has a very large upper bound for the 95% Wald confidence interval for $\hat{N}$ (Table 3). The upper bound of the MATA-Wald interval is much lower. An alternative confidence interval is a Wald interval computed from $\log \hat{N}_{MA}$ and its standard error; this has a slightly larger upper bound than does the MATA-Wald interval. The estimated sampling distributions of $\log \hat{N}_{MA}$ and $\hat{N}_{MA}$ (Figure 1) illustrate why the MATA interval differs from the Wald interval. Log $\hat{N}_{MA}$ has a skewed sampling distribution, while the Wald interval method assumes a normal distribution. When back transformed to population sizes, the consequence is that upper quantiles of the empirical distribution are smaller than those assumed by the Wald method.

When AIC is used to determine model weights, the model set needs to be chosen carefully to avoid redundant or near-redundant models. Near-redundant models are two or more models containing highly correlated variables. To illustrate the issue caused by redundant models, consider a set of 6 models. Model 1 includes temperature in degrees Celsius; model 2 includes temperature in degrees Fahrenheit. These are redundant models and will have the same log Likelihood and AIC statistics. The other 4 models have different variables. The 6 models have $\Delta$AIC statistics

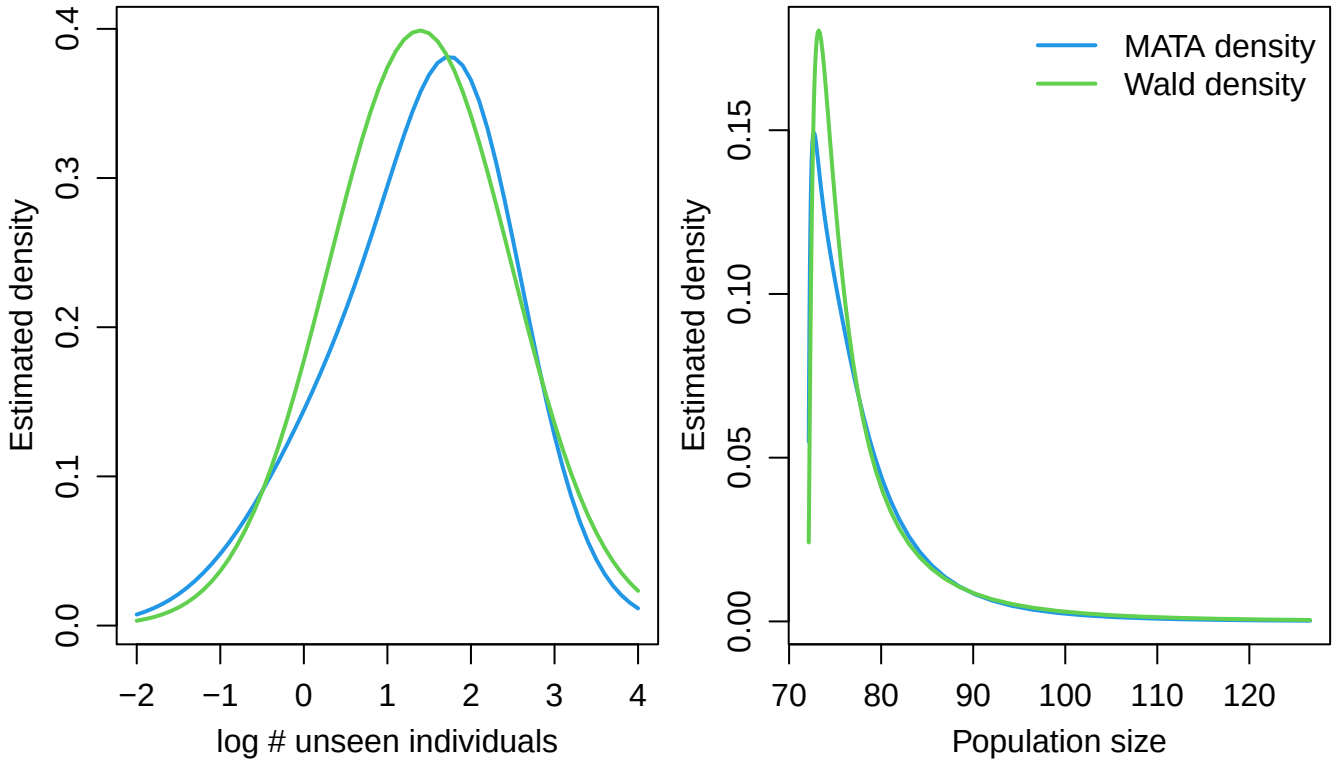


Figure 1: Sampling distributions for $\log \hat{N}_{MA}$ and $\hat{N}_{MA}$ estimated from a sequence of MATA confidence intervals. Each is compared to their sampling distributions assumed by the Wald method, normal for $\log \hat{N}_{MA}$ and log normal for $\hat{N}_{MA}$.

of 0, 1, 1, 3, 5 and 6, where models 1 and 2 both have $\Delta$AIC statistics of 1. The associated model weights are 0.39, 0.24, 0.24, 0.09, 0.03, and 0.02. If only one of the temperature models is included in the model set, there are now only 5 models. The $\Delta$AIC statistics are 0, 1, 3, 5 and 6, with associated model weights of 0.51, 0.31, 0.11, 0.04, and 0.03. The weight given to the single temperature model, 0.31, is neither the weight given to either model 1 or 2, 0.24, when both are in the model set, nor their sum, 0.48.

AIC is just one of the statistics that can be used to determine frequentist model weights (Fletcher 2018, section 3.2). Two promising alternatives are bagging (Buckland et al. 1997) and stacking. I summarize them here; more details are given in Fletcher (2018).

Bagging is the use of a bootstrap samples to estimate model weights. The bootstrap can be a parametric or non-parametric bootstrap of observations or residuals. The choices depends on what is most appropriate for the problem at hand (Buckland et al. 1997). For each of $B$ bootstrap samples, a model selection criterion is computed for every model in the model set; the identity of the "best" model and the estimated parameter (or prediction) for the "best" model, $\hat{\theta}_k$ are recorded. The model weight, $w_k^B$, for model $k$ is the proportion of bootstrap samples for which model $k$ is selected, i.e. $w_k^B = B_k/B$, where $B_k$ is the number of times model $k$ is selected. The model averaged estimate is then:

$$\hat{\theta}_{BAG} = \sum_{k=1}^{K} w_k^B \, \overline{\hat{\theta}}_k,$$

where $\overline{\hat{\theta}}_k$ is the average estimate when model $k$ is selected.

Bagging provides a way around the redundant or nearly-redundant model issue with AIC-based weights. If there are two redundant models, only one will be selected for any bootstrap sample. The total number of times models 1 and 2 are selected will equal, within Monte-Carlo error, the number of times one of the models will be selected when the other is not included in the model set.

Stacking derives model weights from the ability of a model to make out-of-sample predictions, using cross-validation (Stone 1974). Define $\hat{\theta}_{k[-i]}$ as the prediction of observation $i$ applying model $k$ to the leave-one-out cross-validation sample, i.e., omitting observation $i$. Given a set of model weights, $w_k$, the stacking predictor of $\theta_i$ is

$$\hat{\theta}_i = \sum_{k=1}^{K} w_k \, \hat{\theta}_{k[-i]}.$$

The model weights are those that maximize $\sum_{i=1}^{N} \log L(\hat{\theta}_i \mid y_i)$. Here, $\log L(\hat{\theta}_i \mid y_i)$ is the contribution of observation $i$ to the log likelihood using $\hat{\theta}_i$ based on the data omitting observation $i$. Although AIC and cross-validation are derived from different principles, they are asymptotically equivalent (Stone 1977). That suggests that when $N$ is large, stacking model weights and AIC-based model weights will be similar.

## 4.2 Bayesian model averaging

Bayesian approaches to model averaging have advantages and disadvantages. The primary advantage is that model choice can be included as a random variable in the analysis. This provides both a posterior probability for each model and the posterior distribution of the model-averaged estimate or prediction. That posterior distribution can be interpreted directly or summarized as the posterior mean, the posterior median, or a credible interval. The disadvantage is that posterior model probabilities depend on the prior distributions assigned to both the model probability and the parameters, so care is required in specifying and justifying the choice of prior distributions. I use the chipmunk data to illustrate two approaches for Bayesian model averaging.

The simplest approach is to construct model weights from the (Schwartz) Bayesian Information Criterion:

$$\text{BIC} = -2 \log L + p \, \log(N).$$

Like the closely related AIC statistic, equation (2), the BIC statistic combines the fit of the data to the model and a penalty for model complexity, $p \, \log(N)$. For any reasonable sample size, the BIC penalty is larger than the AIC penalty ($2p$) so when used as model selection statistics, BIC will tend to select models with fewer parameters than does AIC. Differences in BIC statistics are converted to posterior model probabilities in exactly the same way as are differences in AIC or AICc statistics:

$$w_k = \frac{\exp(-\Delta \text{BIC}_k/2)}{\sum_{j=1}^{K} \exp(-\Delta \text{BIC}_j/2}.$$

Applied to the four capture probability models for the chipmunk data, using BIC selects a simpler model, Mb, (Table 4) than does AICc (Table 1). BIC gives essentially all the model weight to models Mb and M0 (Table 4). The BIC-based model-averaged estimate of population size is 76.2.

| Model | $\Delta$ BIC$_i$ | $w_i$ | $\hat{N}_i$ |
|---|---|---|---|
| Mb | 0 | 0.615 | 78.1 |
| M0 | 0.94 | 0.385 | 72.7 |
| Mt | 35.89 | 0.0 | 72.5 |
| Mtb | 39.59 | 0.0 | 76.7 |

Table 4: BIC statistics, as difference from the best model, model weights, $w_i$, and estimated population sizes $\hat{N}_i$ for each of the four Otis models fit to the chipmunk data. Models are sorted from best (smallest BIC value) to worst fit.

A Bayesian justification for using BIC is that BIC provides an approximation to the Bayes factor comparing two models, when a unit-information prior is used for the model parameters (Raftery 1999). Conceptually, the unit information prior is a prior distribution that provides the same information about a parameter as does a single typical observation (Raftery 1999). If the model set includes the data generating model, or an approximation to it, the model weights computed

using BIC can be interpreted as posterior probabilities that a model is the data generating model. AIC-based model weights can also be interpreted as posterior model probabilities; the difference is the choice of prior model probabilities. BIC implicitly puts equal prior probabilities on each model; AIC corresponds to a prior model probability that increases with the number of parameters in the model in a specific way (Burnham and Anderson 2004, section 4).

A second approach to Bayesian model averaging is to specify explicit prior distributions for parameters and explicit prior model probabilities. Then, Bayes rule can be used to obtain posterior distributions of parameters and posterior model probabilities. This is most commonly implemented numerically using MCMC methods. Model averaging requires sampling across multiple models. This can be done in various ways (O'Hara and Sillanpää 2009), including adding 0/1 indicator variables to the model (Kuo and Mallick 1998) or by using a reversible jump MCMC algorithm (Green 1995).

When the models in the model set differ only the set of parameters included in each model, the simplest multi-model inference analysis adds a 0/1 indicator variable for each parameter that may or may not be included. This approach was developed by Kuo and Mallick (1998) and is described well by Link and Barker (2010). This approach is closely related to the Stochastic Search Variable Selection method of George and McCulloch (1993). To illustrate the approach, consider a linear regression model with two potential variables. The model, augmented with 0/1 indicator variables, is:

$$\begin{aligned} Y_i &= \beta_0 + Z_1\beta_1 X_{1i} + Z_2\beta_2 X_{2i} + \varepsilon_i \\ Z_1 &\sim Bernoulli(\pi_1) \\ Z_2 &\sim Bernoulli(\pi_2) \end{aligned}$$

When $Z_1 = 0$, $X_1$ is excluded from the model; when $Z_2 = 0$, $X_2$ is excluded from the model. The posterior estimates of $\pi_1$ or $\pi_2$ are the marginal probabilities that $X_1$ or $X_2$ are included in the model. The joint distribution of $Z_1$ and $Z_2$ gives the posterior probabilities for all four combinations of $X_1$ and $X_2$.

A Baysian model averaging of capture-recapture data can be constructed by combining the data augmentation strategy of Royle, Dorazio and Link (2007) and the Kuo-Mallick indicator variable parameterization. The Royle, Dorazio and Link (2007) data augmentation strategy is to consider a superpopulation of $M$ individuals, where $M > C$, the number of individuals captured at least once. Some of the $M - C$ individuals are in the population but never captured. An indicator variable, $Z_{0i}$, is defined for each of the $M$ individuals. Individual $i$ is in the population when $Z_{0i} = 1$ and not when $Z_{0i} = 0$. The estimated population size, $\hat{N}$, is then $\hat{N} = \sum_{i=1}^{M} Z_{0i}$.

The four Otis capture probability models can be written as a single equation for the capture probability, $p_{ij}$, for individual $i$ on occasion $j$, as

$$\text{logit}\, p_{ij} = \beta_0 + \alpha c_{ij} + \sum_j \beta_j t_{ij}, \tag{4}$$

where $t_{ij}$ is a set of indicator variables identifying the capture occasion. They have the value 1 on occasion $t_j$ and 0 otherwise. The $c_{ij}$ indicate whether an individual has been previously

captured. Each $c_{ij}$ has the value of 1 if individual $i$ has been capture before occasion $j$ and 0 otherwise. The $\alpha$ parameter quantifies the behavioral response and has the value of 0 if there is no behavioral response, so the model is M0 or Mt. The set of $\beta_1 \cdots \beta_T$ quantify the variability across capture occasions. When all are 0, there is no time variation in capture probability, so the model is model M0 or Mb. Using the Kuo-Mallick approach, there are two indicator variables, $Z_b$ and $Z_t$:

$$\text{logit}p_{ij} = \beta_0 + Z_b \alpha c_{ij} + Z_t(\sum_j \beta_j t_{ij}),$$

where a value of $Z_b = 0$ drops the behavioral response term and $Z_t = 0$ drops all the time effects. The model is completed by specifying distributions for the indicator variables, $Z_{0i}$, $Z_b$ and $Z_t$, and prior distributions for all parameters:

$$\begin{aligned}
Z_{0i} &\sim Bernoulli(\psi) \\
Z_b &\sim Bernoulli(\pi_b) \\
Z_t &\sim Bernoulli(\pi_t) \\
\psi &\sim U(0,1) \\
\pi_b &\sim U(0,1) \\
\pi_t &\sim U(0,1) \\
\beta_0 &\sim U(-2,2) \text{ or } U(-3.5,3.5) \\
\alpha &\sim U(-2,2) \text{ or } U(-3.5,3.5) \\
\beta_1 \cdots \beta_T &\sim U(-0.7,0.7)
\end{aligned}$$

The prior distributions for $\pi_b$ and $\pi_t$ were chosen to give equal prior probabilities to the four capture probability models. I consider two choices of prior distributions for model parameters. With prior 1, the intercept and the coefficient for the behavioural effect are given uniform(-2, 2) distributions. With prior 2, those coefficients are given uniform(-3.5, 3.5) distributions. Uniform prior distributions were used so that back-transformed capture probabilities did not venture too close to 0 or 1.

The model was fit using rjags with 3 parallel chains and a burnin of 10000 samples. The posterior distributions were estimated from the next 10000 samples, thinned to 1000 samples. Convergence was assessed by the Gelman-Rubin statistic and visual inspection of the trace plots. Gelman-Rubin statistics for all parameters were less than 1.05. The posterior model probabilities are given in Table 5.

The weakness of Bayesian model averaging is that posterior model probabilities are sensitive to the choice of prior distributions for model parameters (Raftery 1999), even when sample sizes are large. This is in sharp contrast to the relative robustness of posterior parameter distributions. This is illustrated by the difference in posterior model probabilities between Prior 1 and Prior 2 (Table 5). The posterior model probabilities are more different than expected from 3000 Monte-Carlo samples.

Although the posterior model probabilities for the Mtb model depend on the choice of prior distributions, Mtb has the highest posterior probability with either. The estimated population

| Model | P[model \| data] | |
|---|---|---|
| | Prior 1 | Prior 2 |
| Mtb | 0.59 | 0.46 |
| Mt | 0.26 | 0.37 |
| Mb | 0.14 | 0.13 |
| M0 | 0.02 | 0.04 |

Table 5: Posterior model probabilities using the Kuo-Mallick approach to model averaging with uniform (0,1) prior distributions for probabilities. Prior 1 is U(-2, 2) probabilities for logit effects on capture probabilities; prior 2 is U(-3.5, 3.5) probabilities for those parameters.

sizes are very similar for the two prior distributions. The model-averaged posterior estimate of $\hat{N}$ is 78 with a standard error of 8.9 for prior 1 and 79 with a standard error of 8.3 for prior 2. The 90% credible intervals are (72, 92) and (71, 96). Both sets of results for the estimated population size are similar between the BIC-based and Kuo-Mallick based approaches, in large part because models Mb and Mtb have similar model-specific estimates of $\hat{N}$. The Kuo-Mallick approach has the advantage of easily providing standard errors and credible intervals, in spite of a highly skewed posterior distribution for $\hat{N}$ (Figure 2).

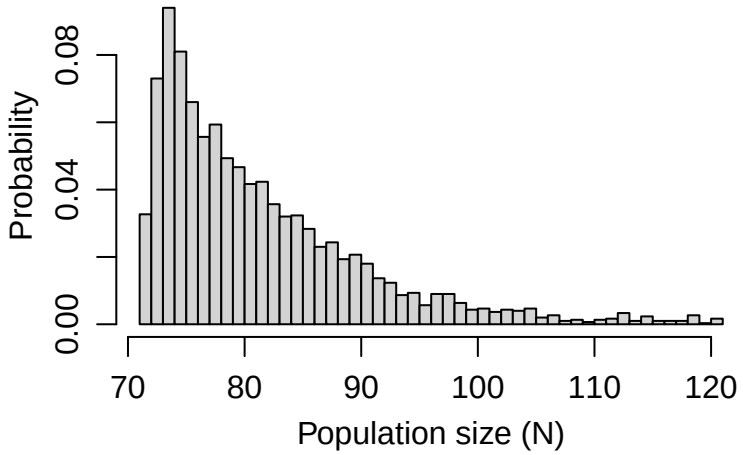


Figure 2: Posterior model-averaged estimate of the population size, $\hat{N}$ for the chipmunk data.

An alternative to the Kuo-Mallick indicator variable approach is a reversible-jump Markov chain Monte-Carlo chain (RJMCMC) algorithm. The reversible jump aspect allows the Markov chain to move between models with different parameters. King et al. (2010) provide an accessible description of the RJMCMC algorithm. Conceptually, the Kuo-Mallick and RJMCMC approaches are similar; the major difference is their behavior when a parameter is being considered to be added to the current model. The Kuo-Mallick considers proposals from the prior distributions (O'Hara and Sillanpää 2009) while the RJMCMC considers proposals that are randomly shifted versions of previous values (King et al. 2010). As a result the RJMCMC algorithm is expected to mix better and converge more quickly. If the prior distributions for the parameters in the

capture probability model are too large, the Kuo-Mallick algorithm will not mix well and may never transition to a model with time effects.

# 5 Model averaging the intervention effect in a BACI study

The second example of model averaging evaluates the impact of forest cutting on the abundance of insect herbivores in Guyana (Basset et al. 2001). The study design was a simple example of a Before-After-Control-Impact (BACI) design. Two large forest tracts were delineated. Herbivorous insects were measured monthly for 11 months in both tracts. One tract was randomly chosen to be logged; insect sampling continued monthly for another 11 months. Basset et al. (2001) report total counts for many insect groups. The data used here are based on the reported totals of Curculionid weevils. Monthly counts were simulated from Poisson distributions constrained so that the sum over 11 months matched the reported total count for that period and tract. The mean counts in each combination of period (before/after) and tract (control/impacted by logging) are shown in Figure 3. The control and logged tracts appear to have different temporal trends (after - before).

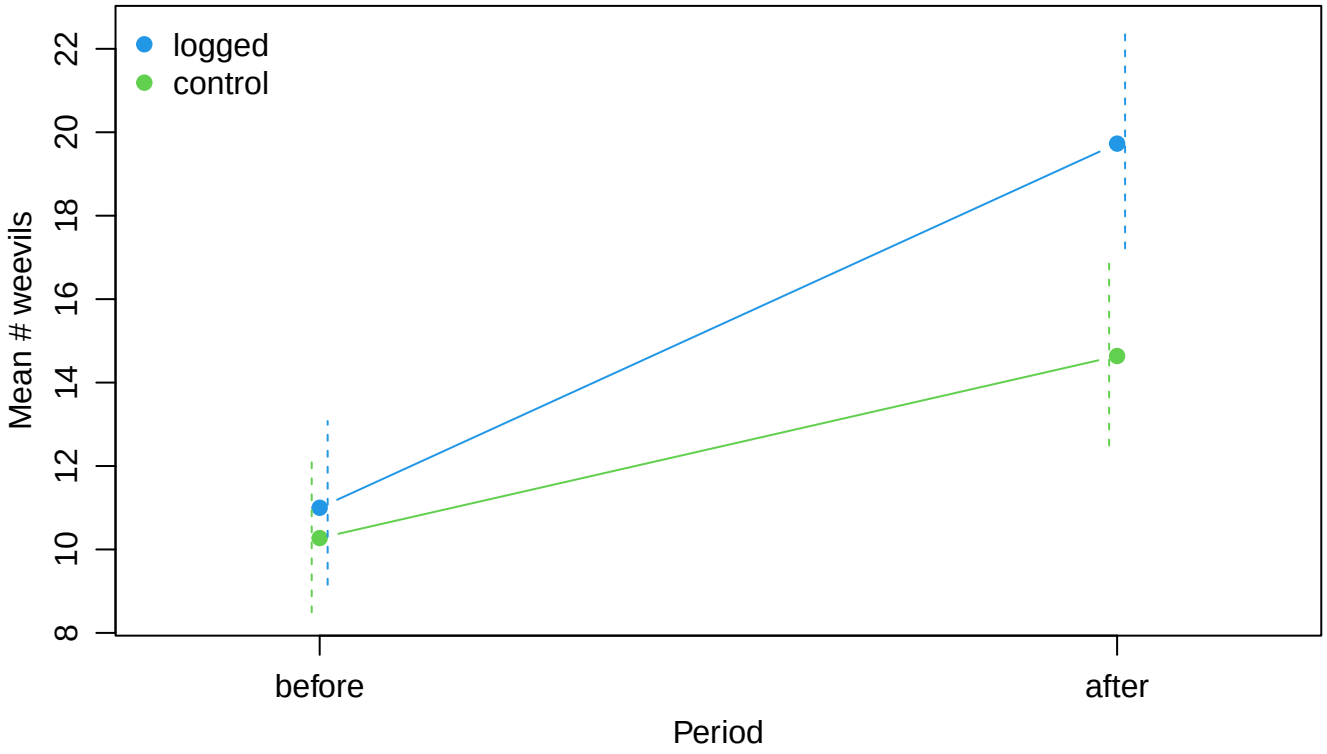


Figure 3: Mean numbers of weevils, with 95% confidence intervals, before and after logging in the control and the logged forest tract. 95% confidence bars are jittered for clarity.

The difference in trend can be evaluated by fitting a model that estimates the interaction effect, (before - after in the logged tract) - (before - after in the control tract, after accounting for preexisting differences between the two tracts (i.e., in the before period) and common temporal trends (after - before) in both sites. The 11 sample quadrats for each combination of period and tract are subsamples, not true replicates. This is not uncommon in environmental BACI studies where there is only one impact site and the object of inference is these two specific tracts (Stewart-Oaten and Bence 2001).

The mean number of weevils in site $i$, $i = c, i$ and period $j, j = b, a$ is denoted by $\theta_{ij}$. The interaction effect is then $(\theta_{ib} - \theta_{ia}) - (\theta_{cb} - \theta_{ca})$. A linear link will be used for interpretability. This doesn't cause any issues with negative estimates of $\hat{\theta}_{ij}$ because all mean counts are larger than 10.

The standard model for the 2x2 BACI design is $\theta_{ij} = \mu + \alpha_i + \beta_j + \gamma_{ij}$. This can be fit as a generalized linear model by defining indicator variables for site effects, $X_{site}$, for period effects, $X_{period}$, and the interaction effect, $X_{interaction}$:

$$\begin{aligned} Y_{ij} &\sim \text{Poisson}(\theta_{ij}) \\ \theta_{ij} &= \mu + \alpha\, X_{site} + \beta\, X_{period} + \gamma\, X_{interaction}. \end{aligned}$$

The estimated interaction effect is 4.36 individuals with a 95% confidence interval of (-0.04, 8.79).

The hierarchy principle requires that when a model includes an interaction term, it also includes all component terms (Nelder 1977). For a 2x2 BACI design, the only model set that respects hierarchy has two models, one with and the other without the interaction term. Factorial designs with 3 or more factors have many more submodels that respect hierarchy (Fletcher and Dillingham 2001).

When model averaging is näively applied to all three terms in equation (5), the results illustrate the importance of respecting hierarchy. The indicator variables in equation (5) are commonly defined one of three ways (Table 6). The interaction effect in the full model, equation (5), is the same for all 3 parameterizations, except perhaps for a sign change or constant multiplier. This is not the case for reduced models that do not respect hierarchy, e.g., $\theta_{ij} = \mu + \gamma\, X_{interaction}$. Under this model with "set first to 0" indicator variables, $\gamma$ is the mean difference between the logged, before cell and the average of the other three cells. With "set last to zero" indicator variables, $\gamma$ is the mean difference between the control, after cell and the average of the other three cells. With "sum to 0" indicator variables, $\gamma$ is still the interaction effect in the full model. Because model averaging only makes sense when all models estimate the same population quantity, a model-averaged estimate of the interaction can not be done with "set first to 0" or "set last to 0" indicator variables. It can be done with "sum to 0" indicator variables, which produce orthogonal columns of the $\boldsymbol{X}$ matrix.

A better way to choose models to be averaged is to specify simpler models that are ecologically relevant. For a study using a 2x2 BACI design, two ecologically relevant simplifications are:

- No difference between impact and control sites during the before period

- No change at the control site, i.e., no difference between before and after at the control site.

We will define $\beta_{before} = \theta_{cb} - \theta_{ib}$ and $\beta_{control} = \theta_{ca} - \theta_{cb}$. We want to estimate the interaction, $\beta_{interaction} = (\theta_{ib} - \theta_{ia}) - (\theta_{cb} - \theta_{ca})$ and will also include the overall intercept, $\beta_0 = (\theta_{ib} + \theta_{ia} + \theta_{cb} + \theta_{ca})/4$. This set of four parameters can be written as a matrix of linear combinations of the four $\theta_{ij}$:

$$\boldsymbol{\beta} = \boldsymbol{C}'\boldsymbol{\theta} = \begin{bmatrix} \beta_0 \\ \beta_{before} \\ \beta_{control} \\ \beta_{interaction} \end{bmatrix} = \begin{bmatrix} 0.25 & 0.25 & 0.25 & 0.25 \\ 0 & 1 & 0 & -1 \\ -1 & 1 & 0 & 0 \\ 1 & -1 & -1 & 1 \end{bmatrix} \begin{bmatrix} \theta_{ca} \\ \theta_{cb} \\ \theta_{ia} \\ \theta_{ib} \end{bmatrix}$$

With this parameterization, the ecologically relevant simplifications correspond to setting $\beta_{before}$ or $\beta_{control}$ to 0.

| | | "set first to 0" | | | "set last to 0" | | | "sum to 0" | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Site | Period | $X_{site}$ | $X_{period}$ | $X_{int.}$ | $X_{site}$ | $X_{period}$ | $X_{int.}$ | $X_{site}$ | $X_{period}$ | $X_{int.}$ |
| control | after | 0 | 0 | 0 | 1 | 1 | 1 | $-1$ | $-1$ | $+1$ |
| control | before | 0 | 1 | 0 | 1 | 0 | 0 | $-1$ | $+1$ | $-1$ |
| impact | after | 1 | 0 | 0 | 0 | 1 | 0 | $+1$ | $-1$ | $-1$ |
| impact | before | 1 | 1 | 1 | 0 | 0 | 0 | $+1$ | $+1$ | $+1$ |

Table 6: Values of indicator variables for the site effect, the period effect and the site*period interaction under three schemes, "set first level to 0", "set last level to 0", and "sum to 0".

To estimate the four $\beta$ parameters using a (generalized) linear model, we need to find coefficients for $X$ variables so that $\boldsymbol{\beta} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\theta} = \boldsymbol{C}'\boldsymbol{\theta}$. When working with the cell means, $\boldsymbol{\theta}$, both the $\boldsymbol{C}$ and $\boldsymbol{X}$ matrices are full rank and invertible. The desired $\boldsymbol{X}$ is given by $\boldsymbol{C}'^{-1}$, which can be verified by substitution into $\boldsymbol{\beta} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\theta}$. For the $\boldsymbol{C}$ matrix given above, we get:

| | | $\boldsymbol{C}$ matrix | | | | $\boldsymbol{X}$ matrix coding | | | |
|---|---|---|---|---|---|---|---|---|---|
| C | B | 0.25 | 1 | -1 | 1 | 1 | 0.5 | 0.5 | -0.25 |
| | A | 0.25 | 0 | 1 | -1 | 1 | 0.5 | -0.5 | -0.25 |
| I | B | 0.25 | -1 | 0 | -1 | 1 | -0.5 | 0.5 | -0.25 |
| | A | 0.25 | 0 | 0 | 1 | 1 | -0.5 | -0.5 | 0.75 |

All the models in the model set to be considered include $\beta_0$ and $\beta_{interaction}$, because we are interested in the interaction effect. The model set is the four models with different combinations of $\beta_{site}$ and $\beta_{period}$. When the Xperiod variable is removed from the full model, equation (5), the estimated coefficient for $X_{interaction}$, $\hat{\gamma}$, equals $\theta_{ca} - \theta_{cb}$, which is the estimate of the interaction under the restriction $\beta_{before} = \theta_{cb} - \theta_{ib} = 0$. When the Xsite variable is dropped, $\hat{\gamma}$ equals $\theta_{ia} - \theta_{ib}$, again what it should be under the restriction $\beta_{control} = \theta_{ca} - \theta_{cb} = 0$. When both the Xperiod and Xsite variables are removed, $\hat{\gamma} = 0.75[\theta_{ia} - (\theta_{ib} + \theta_{ca} + \theta_{cb})/3]$.

For each of these models, table 7 shows the estimated interaction effect, its standard error, and AICc and BIC statistics. The standard error of the interaction effect is smallest for when the model includes only the interaction effect and increases as more terms are added. The model with period and interaction effects has the smallest AICc and BIC statistics. The full model, with period, site and interaction effects is 2.2 AICc units and 3.5 BIC units larger. A model selection approach would make conclusions about the interaction effect using the period + interaction model. A model averaging approach will combine information from all models.

| Model | $\hat{\beta}_{int}$ | se | AICc | BIC |
|---|---|---|---|---|
| Period | 5.09 | 1.77 | 268.0 | 272.8 |
| Period + Site | 4.36 | 2.25 | 270.2 | 276.3 |
| Neither | 7.76 | 1.47 | 275.2 | 278.5 |
| Site | 8.73 | 1.67 | 276.2 | 281.0 |

Table 7: Estimated interaction effect, its standard error, and AICc and BIC statistics for four possible models for the number of Curculionid weevils.

I consider model averaging using AIC weights, BIC weights, and Kuo-Mallick when the prior distributions for $\beta_{site}$, $\beta_{period}$ and $\beta_{interaction}$ are normal with mean 0 and sd of either 10 or 100. The same two models (Period and Period+Site) have appreciable probability, using AICc-derived weights, BIC-derived weights, or Kuo-Mallick with the prior sd = 10. (Table 8). The model with period and interaction effects has the largest weights and posterior model probability, but the full model also has appreciable probability, especially using AICc weights. This is consistent with the larger prior model probability given to the full model (more parameters) by AIC and AICc weights. With a very diffuse prior distribution (sd = 100) for $\beta_{site}$, $\beta_{period}$ and $\beta_{interaction}$, the two models with appreciable prior probability are the model with Period and the model with neither Period nor Site.

| Model | AICc weight | BIC weight | post. model prob. | |
|---|---|---|---|---|
| | | | prior sd=10 | prior sd = 100 |
| Period | 0.722 | 0.802 | 0.818 | 0.590 |
| Period, Site | 0.246 | 0.139 | 0.122 | 0.008 |
| Neither | 0.020 | 0.046 | 0.052 | 0.390 |
| Site | 0.012 | 0.013 | 0.008 | 0.011 |

Table 8: Model weights for the four possible models for Curculionid weavils.

The model averaged estimates of the interaction effect (Table 9) account for the uncertainty in the choice of model. The mean estimated interaction effect depends on the prior probabilities given to the different models and the prior probabilities for model parameters (Table 9). However, all the estimates are within approximately 1 standard error of each other.

| Model weights | Estimate | se. | 90% interval |
|---|---|---|---|
| AIC-based | 5.09 | 2.00 | (1.65, 8.37) |
| BIC-based | 5.16 | 1.98 | (1.84, 8.48) |
| Kuo-Mallick, prior sd = 10 | 5.03 | 1.94 | (1.82, 8.20) |
| Kuo-Mallick, prior sd = 100 | 6.22 | 2.13 | (2.73, 9.66) |

Table 9: Estimated interaction effect, its standard error, and 90% credible intervals for four possible model averaging approaches.

# 6 When is model averaging useful?

The primary value of model averaging is to reflect the scientifically honest admission that the model is not known. Model averaging accounts for the uncertainty in the choice of model. As a result, model averaging is likely to decrease the precision of the estimate or prediction. At the same time, model averaging is likely to increase the validity of an estimate or prediction, in the sense that model averaging does not assume that a single model used in an analysis is the data generating model (or a close approximation).

My two examples illustrate two frequent uses of model averaging. The chipmunk capture-recapture study illustrates model averaging over sets of nuisance parameters, in this case, those describing the capture process. The BACI study illustrates model averaging over ecologically relevant hypotheses.

There are many other possible uses of model averaging. One is a randomized study where various covariates are measured before treatment initiation. Regression matching on a relevant set of covariates usually increases the precision of the treatment effect (Cox 1958). The issue is choosing the relevant set, or sets, of covariates. Averaging estimated treatment effects over models with different sets of covariates accounts for the uncertainty in the choice of covariate model.

A second, quite different example, is combining predictions made by different methods, e.g. a random forest, a generalized additive model, and a neural network. If all methods provide a log likelihood, one could model average using AIC or BIC. If not, a prediction-based method, e.g. stacking, can provide model weights.

# 7 When is model averaging a distraction?

The fundamental assumption of model averaging is that the parameter being averaged has the same interpretation in all models (Cade 2015). In the chipmunk capture-recapture analysis, it is clear that $N$, the number of individuals in the population, has the same interpretation for all models of the capture process. Whether model averaging is appropriate for the BACI analysis

depends on the choice of parameterization. When the model parameterization is the default R (set first to 0) or SAS (set last to 0) parameterizations, MA is not appropriate. It is appropriate with an orthogonal (e.g., sum to 0) or an ecologically relevant parameterization.

The importance of this fundamental assumption is often overlooked when model averaging is used with multiple linear regressions. This has two potential consequences. The interpretation of a parameter in a multiple regression is conditional on the other variables in the model unless all the variables in the model are uncorrelated (Cade 2015). Different models condition on different sets of variables. One multiple linear regression quantity that does have the same interpretation in all models is a prediction at a specific set of covariate values. In a linear regression, the predicted value is a linear combination of the regression coefficients, so model averaged coefficients provide a short cut to computing model averaged predictions. This is not the case for a generalized linear model with a non-linear link function, e.g., log or logit, because of that non-linear link. Predictions can still be model averaged, but they are no longer linear functions of the model averaged coefficients.

The second consequence is relevant when multiple regression results are interpreted in terms of the importance of individual variables. For an individual variable, this is quantified by the sum of model weights or posterior model probabilities for models that include that variable. Those posterior model probabilities can depend on the parameterization of the model, again unless the variables are uncorrelated. Those posterior model probabilities, and hence the variable importance measure, are sensitive to the explicit or implicit specification of prior distributions for parameters and prior model probabilities. Unless there are good justifications for a specific choice of prior, I suggest not calculating sums of model probabilities.

# 8 Conclusions

Model averaging provides an alternative to selecting a single model and making conclusions that are conditional on that choice. There are many ways to implement model averaging, including using information criteria (AIC, AICc, or BIC), bootstrapping, cross-validation, and fully Bayesian approaches. It should not be used blindly because the estimated model weights or posterior model probabilities depend on the choice of method. Given those model weights or posterior probabilities, the model averaged estimated or predicted values are weighted averages of the model-specific quantities. With a fully Bayesian approach, estimated standard errors and credible intervals are simple to compute from the posterior distribution of a model-averaged quantity. However, the posterior distribution and especially the posterior model probabilities depend on choices of prior distributions for parameters and models. Whenever possible, these prior distributions should reflect knowledge of the study system; default choices may be inappropriate. Standard errors and confidence intervals are harder to compute in the frequentist paradigm. The current best way to compute frequentist confidence intervals is the model-averaged-tail-area method.

# 9 References


Basset, Y., Charles, E., Hammond, D.S. and Brown, V.E. 2001. Short-term effects of canopy opennes on insect herbivores in a rain forest in Guyana. *Journal of Applied Ecology* 38:1045-1058.

Bates, J.M. and Granger, C.W.J. 1969. The combination of forecasts. *Operations Research* 20(4):451-468.

Breiman, L. 1992. The little bootstrap and other methods for dimensionality selection in regression: X-fixed prediction error. *Journal of the American Statistical Association* 87:738-754.

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. and Anderson, D.R. 2002. *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*, 2nd ed. Springer, New York

Burnham K.P. and Anderson, D.R. 2004. *Sociological Methods and Research* 33:261-304.

Cade, B.S. 2015. Model averaging and muddled multimodel inferences. *Ecology* 96(9):2370-2382.

Claeskens, G. and Hjort, N. 2003. The focused information criterion. *Journal of the American Statistical Association* 98:900–916.

Claeskens, G. and Hjort, N. L. 2008. *Model Selection and Model Averaging.* Cambridge University Press, Cambridge, UK.

Cox, D.R. 1958. *Planning of Experiments.* Wiley, New York.

Diggle, P.J., Heagerty, P.J., Liang, K-Y., and Zeger, S.L. 2002. *Analysis of Longitudinal Data* Oxford University Press, Oxford UK.

Dormann, CF, Calabrese JM, Guillera-Arroita G, Matechou E, Bahn V, Bartoń K, Beale CM, Ciuti S, Elith J, Gerstner K et al. 2018. Model averaging in ecology: a review of bayesian, information-theoretic, and tactical approaches for predictive inference. *Ecological Monographs* 88(4):485–504

Fletcher, D. 2018. *Model Averaging.* Springer, Berlin

Fletcher, D., and Dillingham, D.W. and Zeng, J.X. 2019. Model-averaged confidence distributions. *Environmental and Ecological Statistics* 26(4):367-384.

Fletcher, D. and Turek, D. 2011. Model-averaged profile likelihood intervals. *Journal of Agricul-*

*tural, Biological, and Environmental Statistics* 17(1):38–51

George, E.I. and McCulloch, R.E. 1993. Variable selection via Gibbs sampling. *Journal of the American Statistical Association* 88:881-889

Green, P.J. 1995. Reversible jump Markov chain Monte Carlo computations and Bayesian model determination. *Biometrika* 82:711-732

Guo, S. and Fraser, M.W. 2010. *Propensity Score Analysis: Statistical Methods and Applications.* Sage Publications, Thousand Oaks, CA

Hjort, N.L., Claeskens, G., 2003. Frequentist model average estimators. *Journal of the American Statistical Association* 98:879–899.

Huang, X., Huang, G., Yu, C., Ni, S., and Yu, L. 2017. A multiple crop model ensemble for improving broad-scale yield prediction using Bayesian model averaging. *Field Crops Research* 211:114-124.

King, R., Morgan, B.J.T., Gimenez, O. and Brooks, S.P. 2010. *Bayesian Analysis for Population Ecology.* Chapman and Hall, Boca Raton FL

Kleiber, W., Raftery, A.E., and Gneiting T. 2011. Geostatistical Model Averaging for Locally Calibrated Probabilistic Quantitative Precipitation Forecasting. *Journal of the American Statistical Association* 106:1291-1303

Kuo, L. and Mallick, B. 1998. Variable selection for regression models. *Sankyhā, Series B* 60(1):65-81

Link W.R. and Barker, R.J. 2010. *Bayesian Inference with ecological applications* Academic Press, London

Mares, M.A., Streilein, K.E., and Willig, M.R. 1981. Experimental assessment of several population estimation techniques on an introduced population of eastern chipmunks. *Journal of Mammalogy* 62(2):315-328.

Nelder, J.A. 1977. A Reformulation of Linear Models (with discussion). *Journal of the Royal Statistical Society, Series A*, 140(1):48-76.

O'Hara, R.B. and Sillanpää, M.J. 2009. A review of Bayesian variable selection methods: What, How, Which *Bayesian Analysis* 4:85-115.

Otis, D.L., Burnham, K.P., White, G.C. and Anderson, D.R. 1978. *Statistical inference from capture data on closed animal populations.* Wildlife Monographs 62.

Prost, L., Makowski, D., and Jeuffroy, M-H. 2008. Comparison of stepwise selection and Bayesian

model averaging for yield gap analysis. *Ecological Modelling* 219:66-76

Riley, R.D., Higgins, J.P.T., and Deeks, J.J. 2011. Interpretation of random effects meta-analyses. *British Medical Journal* 342:d549, doi: https://doi.org/10.1136/bmj.d549

Ripley, B.D. 2004. Selecting amongst large classes of models. pp. 155-170 In Hand, D.J., Adams, N.M., Stevens, D., and Crowder, M.J. *Methods and Models in Statistics: In honour of Professor John Nelder, FRS* Imperial College Press, London

Royle, J.A., Dorazio, R.M., and Link, W.A. 2007. Analysis of multinomial models with unknown index using data augmentation. *Journal of Computational and Graphical Statistics* 16(1):67-85

Schumacher, F.X. and Eschmeyer, R.W. 1943. The estimation of fish populations in lakes and ponds. Journal of the Tennessee Academy of Science 18:228-249.

Stewart-Oaten, A. and Bence, J.R. 2001. Temporal and spatial variation in environmental impact assessment. *Ecological Monographs* 71(2):305-339

Stone, M. 1974. Cross-validatory choice and assessment of statistical predictions (with discussion) *Journal of the Royal Statistical Society, Series B* 36(2):111-147

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(1):44-47

Turek, D. and Fletcher, D. 2012. Model-averaged Wald confidence intervals. *Computational Statistics & Data Analysis* 56(9):2809-2815.

Underwood, A.J. 1994. On beyond BACI - sampling designs that might reliable detect environmental disturbances. *Ecological Applications* 4(1):3-15

Yang, H., Liu, Y., and Liang, H. 2015. Focused information criterion on predictive models in personalized medicine. *Biometrical Journal* 57:422-440.