esa ECOSPHERE

Is your ad hoc model selection strategy affecting your multimodel inference?

Dana J. Morin,[1,†] Charles B. Yackulic,[2] Jay E. Diffendorfer,[3] Damon B. Lesmeister,[4] Clayton K. Nielsen,[5] Janice Reid,[6] and Eric M. Schauber[7]

[1]*Department of Wildlife, Fisheries and Aquaculture, Mississippi State University, Box 9680, Mississippi State, Mississippi 39762 USA*
[2]*Southwest Biological Science Center, U.S. Geological Survey, 2255 N. Gemini Drive, Flagstaff, Arizona 86001 USA*
[3]*Denver Federal Center, U.S. Geological Survey, Geosciences and Environmental Change Science Center, Denver, Colorado 80225 USA*
[4]*Pacific Northwest Research Station, U.S. Forest Service and Department of Fisheries and Wildlife, Oregon State University, 3200 SW Jefferson Way, Corvallis, Oregon 97331 USA*
[5]*Cooperative Wildlife Research Laboratory and Department of Forestry, Southern Illinois University, 251 Life Science II, Mail Code 6504, Carbondale, Illinois 62901 USA*
[6]*Pacific Northwest Research Station, U.S. Forest Service, 777 NW Garden Valley Blvd, Roseburg, Oregon 97471 USA*
[7]*Illinois Natural History Survey, Prairie Research Institute, University of Illinois Urbana-Champaign, 1816 S. Oak Street, Champaign, Illinois 61820 USA*

Citation: Morin, D. J., C. B. Yackulic, J. E. Diffendorfer, D. B. Lesmeister, C. K. Nielsen, J. Reid, and E. M. Schauber. 2020. Is your ad hoc model selection strategy affecting your multimodel inference? Ecosphere 11(1):e02997. 10.1002/ecs2.2997

Abstract. Ecologists routinely fit complex models with multiple parameters of interest, where hundreds or more competing models are plausible. To limit the number of fitted models, ecologists often define a model selection strategy composed of a series of stages in which certain features of a model are compared while other features are held constant. Defining these multi-stage strategies requires making a series of decisions, which may potentially impact inferences, but have not been critically evaluated. We begin by identifying key features of strategies, introducing descriptive terms when they did not already exist in the literature. Strategies differ in how they define and order model building stages. Sequential-by-sub-model strategies focus on one sub-model (parameter) at a time with modeling of subsequent sub-models dependent on the selected sub-model structures from the previous stages. Secondary candidate set strategies model sub-models independently and combine the top set of models from each sub-model for selection in a final stage. Build-up approaches define stages across sub-models and increase in complexity at each stage. Strategies also differ in how the top set of models is selected in each stage and whether they use null or more complex sub-model structures for non-target sub-models. We tested the performance of different model selection strategies using four data sets and three model types. For each data set, we determined the "true" distribution of AIC weights by fitting all plausible models. Then, we calculated the number of models that would have been fitted and the portion of "true" AIC weight we recovered under different model selection strategies. Sequential-by-sub-model strategies often performed poorly. Based on our results, we recommend using a build-up or secondary candidate sets, which were more reliable and carrying all models within 5–10 AIC of the top model forward to subsequent stages. The structure of non-target sub-models was less important. Multi-stage approaches cannot compensate for a lack of critical thought in selecting covariates and building models to represent competing a priori hypotheses. However, even when competing hypotheses for different sub-models are limited, thousands or more models may be possible so strategies to explore candidate model space reliably and efficiently will be necessary.

Key words: AIC; information criterion; model selection; multimodel inference; occupancy models; parameter estimation; population models.

Received 24 October 2019; accepted 28 October 2019. Corresponding Editor: Debra P. C. Peters.

† **E-mail:** dana.morin@msstate.edu

Introduction

Statistical models allow ecologists to approximate the underlying complexity of ecological systems but can also lead to large sets of plausible models that are difficult for practitioners to fully explore. Fitting all combinations of reasonable, hypothesis-based structures and covariates for each sub-model (i.e., parameters) within a well-defined candidate set is a recommended practice in model selection (Doherty et al. 2012). However, even when practitioners develop a candidate model set based on reasonable hypotheses for each sub-model (Buckland et al. 1997, Burnham and Anderson 2002, Lukacs et al. 2010), they may still find themselves faced with thousands or more plausible models prohibiting adoption of an all combinations model set. For example, an ecologist interested in testing five covariates in each of four sub-models in a dynamic occupancy model (e.g., detection probability, initial occupancy, colonization, and extinction probability) would need to fit over 1 million (2^{20}) models, without even considering interactions or quadratic terms! Such an exercise would be exceedingly difficult to accomplish in computational environments that require each candidate model to be specified individually in a batch file (e.g., program PRESENCE; Hines 2006). Thoughtfully examining and synthesizing the results of huge model sets could be very time-consuming regardless of the software used.

In response, ecologists commonly adopt multi-stage strategies to reduce the total number of models considered to a more sensible number, with the (often untested) assumption or hope that they will reach the same inference as one would if the all combinations practice was adopted (i.e., if the full model space was explored; Doherty et al. 2012). In particular, investigators hope to identify the same top-ranked model(s) as would be identified by the all combinations standard, and also to fit a large percentage of the top set of models, such that model-averaged estimates and/or variable importance weights will be only minimally different than if the all combinations standard had been adopted.

The potential difficulty with multi-stage approaches is that different model components are estimated from a joint-likelihood and are statistically dependent. Thus, the structure of one sub-model (i.e., capture probability or survival) can influence the structure of another (defined as leak in Catchpole et al. 2004 and Bromaghin et al. 2013). While considerable focus has been placed on determining the best approach to scoring alternative models (e.g., Burnham and Anderson 2002), far less emphasis has been placed on understanding how multi-stage strategies affect inference.

The only studies we are aware of that assessed multi-stage strategies (Doherty et al. 2012, Bromaghin et al. 2013) were focused on the Cormack-Jolly-Seber (CJS) model (Cormack 1964, Jolly 1965, Seber 1965). While these studies provided baseline guidance for multi-stage model building, CJS is a relatively simple model with only two sub-models (detection probability, p, and apparent survival, φ). In more complex models, model weights are distributed across more sub-models and more decisions must be made, increasing the chance of sub-model selection errors cascading through stages. Furthermore, past studies have focused on simulated data where true parameter values are known, but often do not capture the multicollinearity, uncertainty, and noise found in empirical data sets from observational studies. Lastly, past studies have only explored a limited set of the multi-stage strategies that practitioners have applied.

Here, we examine the effectiveness of a broader set of potential strategies for multi-stage model selection. We begin by describing the potential multi-stage strategies, focusing on the various decisions that differentiate strategies. Next, we analyzed four empirical data sets under three model types to consider how inferences would differ from the all combinations practice based on the decisions a hypothetical user might make in developing a strategy. Specifically, we determined the "true" distribution of AIC weights for each data set by fitting all plausible models. Then, we calculated the number of models that would have been fitted and the portion of "true" AIC weight we recovered under different model selection strategies to quantify efficiency and reliability.

Materials and Methods

Multi-stage modeling strategies

Existing multi-stage strategies can be distinguished as a series of analysis decisions. We delineate three general strategies describing the way in which stages are defined and ordered.

1. Sequential-by-sub-model strategies focus on one sub-model at a time (parameter such as detection, *p*; survival, ϕ; or occupancy, Ψ) with modeling of subsequent sub-models dependent on the selected model structures from the previous stages. For example, in a CJS model, an analyst may first fit sub-models for detection (*p*), ultimately determining a sub-model with *p* varying by time is the best supported model. Next, they will fit the sub-model for survival while keeping the detection sub-model as time-dependent. The approach continues until all sub-models are fit and a final model achieved.
2. Secondary candidate set strategies fit sub-models independently and combine the top set of models from each sub-model for selection in a final stage (Bromaghin et al. 2013). For example, in a CJS model, a practitioner may choose to fit sub-models for *p* and ϕ independently (holding the other sub-model structure constant) in the first stage, then select the sub-model structures with greatest support in the initial stage to combine in a secondary set for a final selection stage.
3. Build-up strategies (Catchpole et al. 2004) define stages across sub-models and increase in complexity at each stage. Stages may be defined across sub-models describing types of variation, perhaps starting by describing temporal trends across sub-groups, adding structure including life history groups, and then adding individual or habitat covariates. For example, Catchpole et al. (2004) employed a build-up strategy to estimate survival in red deer (*Cervus elaphus*), starting with including age structure across all sub-models simultaneously (*p* and ϕ), then comparing time-dependent sub-model structures, and lastly investigating addition of environmental covariates.

After defining stages, the next decision is how to progress through stages, focusing on particular sub-model components while holding sub-model structure for other parameters constant (e.g., by using an intercept-only model structure or maintaining a fixed set of covariates on non-target sub-models). When stages are defined by sub-models, fitting detection first (*p*-first) is common as heterogeneity in detection can bias estimates of sub-models associated with the demographic parameters of interest. However, one could also fit the demographic parameter sub-models first and then fit detection to correct for biases in estimates. The demographic parameter-first progression imposes the fewest constraints on model selection during the initial stages and thus may allow the greatest freedom in modeling the demographic parameter sub-models. However, when there are multiple demographic sub-models, such as survival and recruitment in Jolly-Seber type models (Jolly 1965, Seber 1965), it is still necessary to decide which to fit conditional on another for the sequential-by-sub-model approach. When stages are defined across sub-models in the build-up strategy, the natural progression is to start with temporal or group trends and then add detail with covariates. Order is not a consideration with the secondary candidate set approach as each sub-model is fit independently of the others before combining for final selection.

Strategies also exist for how to proceed within stages including choice of structure for non-target components held constant, and thresholds for within-stage model exclusion. Using the null structure (i.e., not varying through time and space, also referred to as (.) sub-model structures) for non-target sub-models is common as it reduces the number of estimated parameters in the sub-model sets (which in turn makes model fitting quicker and potentially avoids convergence issues), allows for more complex target sub-model structures to be examined, and eliminates decisions about how to define a general sub-model structure. However, using a general sub-model structure for non-targets (allowing for inclusion of covariates or temporal and spatial structure) to avoid constraining the fit of target sub-models has often been advocated (Lebreton et al. 1992, Doherty et al. 2012, Bromaghin et al. 2013). Depending on the study or the data set, a general sub-model structure can range from restricted site and time-dependent differences, include additive or interaction terms, or may use all available covariates and the most complex sub-model structure available (the global sub-model structure), leading to some subjectivity in how the general model is defined.

While some strategies take a multi-stage approach focused only on the best sub-model structure identified in each stage, other strategies carry forward all sub-model structures within some threshold of support (we will refer to AIC throughout, although other metrics could be used) or even incorporate additional stages in which competitive sub-model structures that were eliminated in earlier stages are reconsidered (tracking). Selection of sub-model structures to carry forward into subsequent steps is typically defined by a ΔAIC threshold, where selecting only the top-ranked sub-model structure at each step is equivalent to a threshold of $\Delta\text{AIC} = 0$. Relaxing the support threshold (e.g., using a greater ΔAIC cutoff) will typically mean more first-stage models will be carried forward, complicating model selection in subsequent stages, but also expanding the model space explored. Thus, more liberal support thresholds could produce inference that is more consistent with the all combinations practice.

However, carrying forward models solely on the basis of ΔAIC might omit more general (i.e., with a large number of covariates) sub-model structures early on that would be well-supported in the full all combinations model set. To address these possible omissions in early stages, the plausible combinations strategy (Bromaghin et al. 2013) carries forward not only sub-model structures within a ΔAIC threshold, but also any sub-model structures with a greater likelihood than the models within the ΔAIC threshold, allowing consideration of models with more estimated parameters in subsequent steps. Although Bromaghin et al. (2013) evaluated a plausible combinations approach using a secondary candidate set, it is possible to use plausible combinations criteria in all model building strategies. The plausible combinations approach moderately increases the number of models run compared to the equivalent ΔAIC threshold approach, but may reduce the omission of structures with more estimated parameters in early stages in the progression. Similarly, one may employ tracking, simply keeping track of neighboring or competing sub-model structures at each stage to allow for inclusion of those structures prior to final model selection and inference (Lebreton et al. 1992, Yackulic et al. 2014). For example, one might choose to use a ΔAIC threshold of 0, advancing only the top model at each stage, but keep track of sub-model structures within a much more lenient threshold to substitute in at the final stage in an effort to not exclude any important variables. This final tracking of sub-model structures could provide an efficiency trade-off, relaxing ΔAIC thresholds and dependency of inference on structures selected at previous stages, but with reduced number of models run compared to global application of higher ΔAIC thresholds or creating secondary candidate sets.

Evaluation of ad hoc model selection strategies with empirical data sets

We used four empirical data sets and three model types to compare results and inference from multi-stage model selection strategies with results from the all combinations model sets (Table 1). We selected data sets for two different species from the same camera-trap study (Lesmeister et al. 2015) for a single-species, single-season occupancy (simple occupancy, hereafter) model type (MacKenzie et al. 2002), used a long-term northern spotted owl (*Strix occidentalis caurina*) and barred owl (*Strix varia*) survey data set for a dynamic two-species occupancy (dynamic occupancy, hereafter) model type (MacKenzie et al. 2009, Yackulic et al. 2014), and chose a previously unpublished multi-site and multi-session small mammal capture–mark–recapture data set for a robust design parameterization of Pradel's seniority population (Pradel model, hereafter) model type (Pradel 1996). We evaluated success of multi-stage strategies in four ways. First, we determined if the top-ranked model in the all combinations model set was included in the reduced multi-stage model selection set. Second, we calculated proportion of the "true" model space recovered, or the proportion of the total Akaike weight from the all combinations model set recovered by the ad hoc selection strategies (Σw_i for the all combinations model set conserved in the multi-stage model sets) based on the recalculated Σw_i when models with uninformative parameters were removed from the all combinations model sets (Arnold 2010). Third, we counted the number of models run for each strategy and used that to calculate the proportion of all combinations models run to evaluate the

Table 1. Factors and covariates for candidate sets for the four example data sets.

Model	Parameter	Factors and covariates
Bobcat and gray fox simple occupancy	p	Intercept only
		Mean temperature during a survey week
		Mean precipitation during a survey week
		Year of survey
		Previous detection at the same camera cluster
	Ψ	Intercept only
		Distance to structure
		No. structures/ha
		No. streams/ha
		% urban cover
		% forest cover
		% agricultural cover
		% grassland cover
Owls dynamic occupancy	p_{SO}	Day, night, mtd1, and mtd3
		Effect of barred owl
	γ_{SO} and $\varepsilon_{\mathrm{SO}}$	Intercept only
		Autologistic effect
		Year effect
		Riparian forest
		Older forest
		Effect of barred owl
	γ_{BO} and $\varepsilon_{\mathrm{BO}}$	Autologistic effect and forest type
		Effect of SO
Pocket mice Pradel model	$p \neq c$	Session
		Plot × session
	ϕ and γ	Plot + session
		Plot × session
		No. times a plot burned (time varying)
		Vegetation (PCA1, time varying)
		Vegetation (PCA2, time varying)
		Total precipitation for the prior interval

efficiency of strategies that correctly identified the top-ranked model and recovered a large proportion of the total Akaike model weights in the all combinations model set. For example, consider an all combinations model set that consists of five models with Akaike model weights equivalent to 0.50, 0.25, 0.12, 0.07, and 0.06, respectively. If a given multi-stage strategy fit the top four models, but not the last, the proportion of models run would be 0.80, and the total Akaike weight recovered would be 0.94. Whereas if a multi-stage strategy fit the first, fourth, and fifth models in the all combinations set, the proportion of models run would be 0.60 and the total Akaike weight recovered would be 0.63. Note, we did not use a ratio of total Akaike model weight recovered compared to number of models run as a measure of efficiency because the two objectives are not of equivalent importance (i.e., we considered a strategy with a large number of models run that also recovered a large proportion of the Akaike model weights to be more successful than a strategy that required a small number of models but only recovered a moderate proportion of the total Akaike model weights and did not identify the top-ranked model). Finally, we attempted to assess how much the misidentification of selected models might affect final inference. For example, if the identified best model is similar to the actual top-ranked model (such as only mis-specifying the sub-model structure for detection), misidentification may be of little consequence if the primary objective is prediction. However, if practitioners are making inference about selected variables, then omitting an important covariate or identifying an unimportant covariate can bias predictions. We evaluated the risk of these potential consequences by looking at whether the top-ranked all combinations model structure for the demographic sub-models of interest (i.e., occupancy, colonization, extirpation, apparent survival, and seniority, but not detection) were included and ranked highly in the final model set for each multi-stage strategy when the top model was omitted. We used this evaluation to indicate whether the correct sub-model structure of interest would be included in predictions using model averaging and in assessing variable importance. While our assessment of potential multi-stage model selection strategies was not comprehensive, we evaluated a wide range of plausible strategies to draw conclusions about how critical decisions affect outcomes. In addition, we have provided the all combinations data sets and several example selection functions in the supplementary appendices to allow readers to test other possible strategies and scenarios (Data S1–S3).

Single-season single-species occupancy (Simple occupancy, southern Illinois carnivores)

Similar to the CJS model, the simple occupancy model (MacKenzie et al. 2002) is composed of two statistically dependent sub-models, detection (p) and site occupancy Ψ). The simple occupancy model enjoys widespread use for identifying habitat associations and distributions of species while accounting for heterogeneity in detection. Practitioners frequently implement sequential-by-sub-model strategies to accommodate large suites of potential explanatory covariates. We applied simple occupancy modeling to a subset of detection history data for two carnivore species, bobcat (*Lynx rufus*) and gray fox (*Urocyon cinereoargenteus*), from a landscape-scale camera-trap study in southern Illinois, USA (Lesmeister et al. 2015; see Appendices S1, S2 for data collection and covariate details). We used the same all combinations candidate set for both species, with up to four covariates for modeling p and seven for ψ to allow comparison of strategy success between the two species' data sets (Table 1).

We fit the all combinations model set using RPresence (MacKenzie and Hines 2017), an R package that implements the same models as the graphical user interface software PRESENCE (Hines 2006), resulting in 2048 models, and wrote functions in R (R Core Team 2016) to apply multi-stage selection strategies (see Appendices S1, S2 for more details). For the simple occupancy exercise, we compared three general multi-stage strategies (p-first, ψ-first, and fitting both separately and creating a secondary set) using either a null non-target structure (.) or a general non-target structure (all seven covariates for ψ, all four covariates for p). We chose sub-model structures to carry forward at each stage based on a ΔAIC threshold (0, 2, 5, and 10) or using the plausible combinations approach with a ΔAIC threshold of 0. For the p-first and ψ-first approaches, we also tracked sub-model structures within 2–5 ΔAIC when applicable. We evaluated a total of 42 multi-stage model selection strategies for each species data set.

Dynamic two-species occupancy model (dynamic occupancy, northern spotted owl, and barred owl in Oregon)

Dynamic two-species occupancy models are a special case of a dynamic multistate occupancy model (Miller et al. 2012) with four possible states (neither species present; species A present only; species B present only; both species present), which generates estimates of up to 16 unique parameters, each with its own sub-model (three sub-models describing initial occupancy of both species and their initial co-occurrence; five sub-models describing detection, four sub-models describing colonization of either species based on whether the other species is present or absent, and four sub-models describing extinction for either species from co-occupied or solely occupied sites). In practice, related sub-models often share factors or covariates. For example, extinction probability of species A in the presence and absence of species B might share a covariate describing habitat effects, but selection among models may test whether intercepts for the two sub-models are different. Given the large set of potential models implied by a dynamic two-species occupancy model, practitioners are forced to carefully consider which models to test (including how to incorporate past learning) and which multi-stage strategy to use. For example, Dugger et al. (2016) tested a set of hypotheses using dynamic two-species occupancy that implied more than 200 billion plausible models for each of 11 study areas!

Here, we focus on occupancy dynamics of barred and northern spotted owls (SO) territories over 22 yr at a site in western Oregon, USA (hereafter Tyee Study Area; see Appendices S1, S2, and Yackulic et al. 2014 for data collection and covariate details). Prior knowledge of barred owl territory occupancy dynamics derived from a dynamic single-species occupancy analysis in the same study area (Yackulic et al. 2012) was used to constrain the potential model set in the original analysis (Yackulic et al. 2014), and we further constrained the potential model set in our analysis here to 2592 potential models. We created the all combinations model set using a batch script in R as a front end for program PRESENCE (Hines 2006). There was a primary demographic sub-model set describing local colonization and extinction of SO (γ_{SO} and ε_{SO}, respectively), and a secondary demographic sub-model set for barred owls (BO: γ_{BO} and ε_{BO}) that was expected to affect the primary sub-model set, and a nuisance parameter sub-model (p_{SO} detection of SO). Therefore, to implement the sequential-by-sub-model fitting strategy we tried two p-first strategies (SO and BO sub-models fit second or third), and one focal-first strategy (SO-first, BO-second, and p_{SO}-last) in addition to the secondary set approach fitting each of the sub-model sets separately before combining in a final set. We also tried build-up strategies defining stages across sub-models, starting with fitting p_{SO} first, adding baseline temporal structure, then habitat covariates to SO, and lastly testing for species interactions between SO and BO. We compared all strategies using a null or a general non-target sub-model structure, with the general non-target sub-model structure including year effects on γ and ε and species interactions on all sub-models. We chose sub-model structures to carry forward at each step based on a ΔAIC threshold (0, 2, 5, and 10; with or without tracking competing sub-model structures) or using the plausible combinations approach with a ΔAIC = 0. We evaluated a total of 74 multi-stage model selection scenarios for the dynamic occupancy example.

Pradel's temporal symmetry model with robust design (Pradel model, San Diego pocket mice in California)

Pradel's temporal symmetry model (Pradel 1996) is a parameterization of the Jolly-Seber open population model and allows estimation of apparent survival (φ) and recruitment into the population (derived from the seniority parameter γ, estimating the probability of an individual being present in the population for each session before initial detection). The closed-capture robust design implementation allows estimation of a super-population size (f_0) and individual response to initial capture (p is probability of detection, c the probability of recapture following initial capture). Hence, f_0 is estimated for each site and session and there are four sub-models (two demographic and two nuisance) to estimate that may be fit with different sub-model structures (φ, γ, p, and c).

We analyzed capture histories of 352 individual San Diego pocket mice (*Chaetodipus fallax*) sampled for three days during the summer of each year between 2003 and 2008 at 11 sites in southern California, USA (see Appendices S1, S2 for data collection and covariate details). The objective of the study was to identify factors driving demographic responses to wildfires and extreme climatic conditions to inform future management (Table 1). However, unwieldy data structure has hindered analysis without considering multi-stage model selection strategies.

We fit the all combinations model set using RMark (Laake 2013), a wrapper package for the graphical user interface Program MARK (White and Burnham 1999), resulting in a total of 390 models. We tried several sequential-by-sub-model strategies including starting with detection parameters (p and c-first, φ-second or third, and γ-second or third), or starting with the demographic sub-models (φ-first or second, γ-first or second, p and c-last). Alternatively, we fit detection, φ, and γ separately and created a secondary set. For build-up, we defined stages across sub-models, fitting detection first, adding site and temporal structure to φ and γ second, habitat covariates third, and disturbance covariates last (Table 1). We compared all strategies using a null or a general structure for non-target components. Because the data set includes multiple sites and sessions expected to influence demographic parameters, we used plot × time as the general structure for φ and γ. We chose sub-model structures to carry forward at each stage based on a ΔAIC threshold (0, 2, 5, and 10; with or without tracking competing sub-model structures) or using the plausible combinations approach with a ΔAIC = 0. We evaluated a total of 90 multi-stage model selection scenarios for the Pradel model example.

Results

In the four example data sets, we fit 390–2592 models: However, 93 to >99% of the models fit in each example included uninformative parameters (sensu Arnold 2010; see Appendix S2: Tables S1–S3 for models with only informative parameters). Between 0.32 and 0.56 of model weights were contained in the top-ranked model using the all combinations practice (Table 2). There was greater spread in ΔAIC between the null model and best model for the dynamic occupancy and Pradel model example sets (ΔAIC null dynamic occupancy = 95.07, ΔAIC null Pradel model = 60.76) compared to the simple occupancy examples (ΔAIC null bobcat model = 10.48, ΔAIC null gray fox model = 9.22).

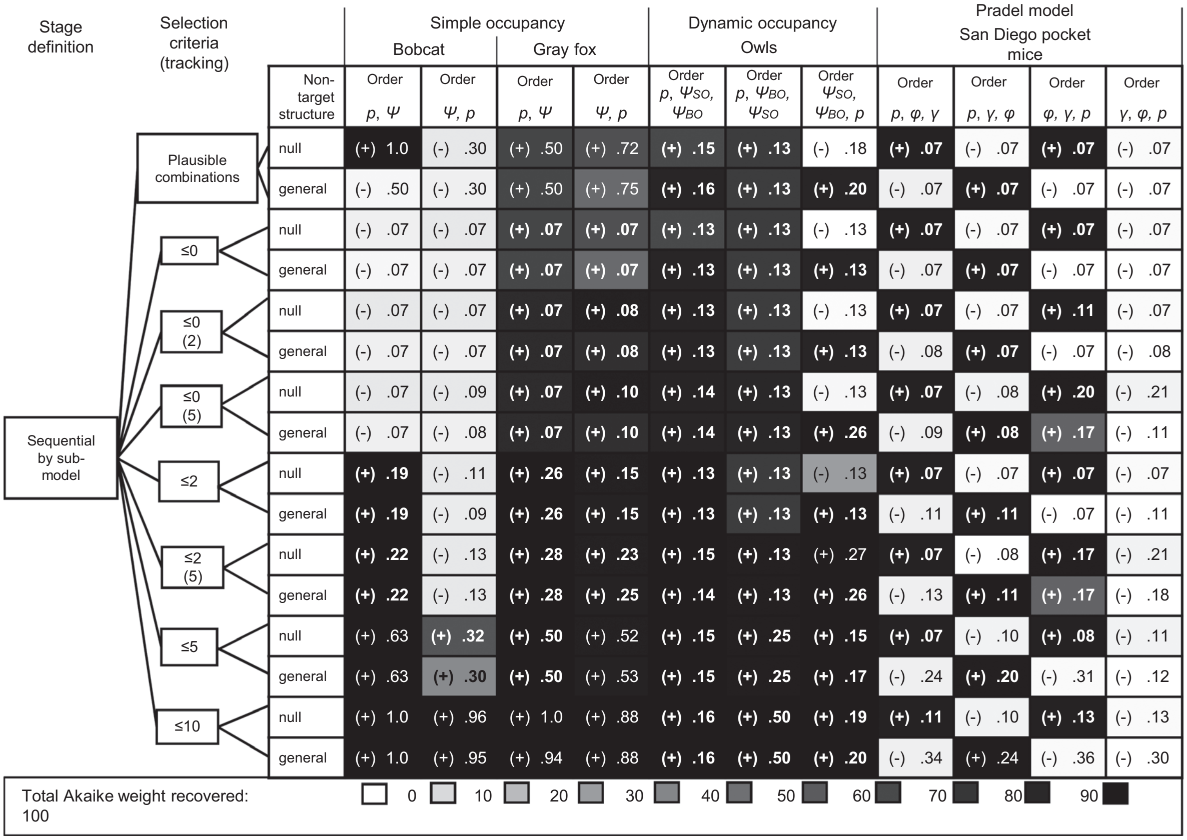
The sequential-by-sub-model approach produced inconsistent results across example sets, identifying top models efficiently in some cases but performing very poorly in others (Fig. 1). For instance, the sequential-by-sub-model approach worked well for the gray fox simple occupancy example set, identifying the top-ranked model and recovering >52% of the total Akaike model weights at all within-stage model selection thresholds. However, results were erratic for the bobcat simple occupancy example set and success depended on which sub-model was modeled first. When p (which had only 1 covariate in the top model) was modeled first, the top model was identified and greater proportion of the total Akaike model weights was recovered even with a conservative threshold of ΔAIC = 2. Conversely, when ψ (which had three covariates in the top model) was modeled first, a ΔAIC = 10 threshold and >95% of the final model set was required to recover the top model. The sequential-by-sub-model approach also yielded inconsistent results when applied to the dynamic occupancy model set, which had >2 sub-model stages. In the dynamic occupancy exercise, the sequential-by-sub-model strategies were more successful when spotted owl detection was fit first (a binary comparison of an effect of BO or not), but the results were less predictable when the demographic sub-models for colonization and local extirpation (γ and ε) for spotted owl were modeled first and detection was fit last. Applying the sequential-by-sub-model approach to the Pradel model example further demonstrated the unpredictable outcomes of this strategy with results changing entirely depending on which of the two demographic sub-models for survival and seniority (φ and γ) was fit first and with the structure of non-target components (null or general).

Symmetrical approaches to model selection using the secondary candidate set (Fig. 2) and build-up (Fig. 3) were more consistent with the all combinations practice and success was independent of whether non-modeled parameters were constrained to a simple or complex sub-model structure. The secondary candidate set yielded similar results to the sequential-by-sub-model approach for the simple occupancy example sets, identifying the top model and recovering >95% of the total Akaike model weights when ΔAIC = 2 for gray fox, but requiring ΔAIC = 10 and >95% of the all combinations bobcat model set to recover >95% of the total Akaike model weights (Fig. 2). However, the secondary set approach identified the top model at all within-stage selection thresholds for the dynamic occupancy and Pradel model example sets, and recovered >95% of the total Akaike model weights when ΔAIC was between 5 and 10 (requiring only 8–16% of the models in the full all combinations sets to be fit). Likewise, the build-up approach consistently identified the top model and recovered >95% of the total Akaike model weights with far fewer models required

Table 2. Summaries of all combinations candidate model sets for the four examples used to evaluate ad hoc model selection strategies (bobcat single-species single-season occupancy, gray fox single-species single-season occupancy, spotted and barred owl two-species dynamic occupancy, and San Diego pocket mouse Pradel's seniority closed population model with robust design).

Model set	Bobcat simple occupancy	Gray fox simple occupancy	Owl dynamic occupancy	Pocket mouse Pradel model
Number of ad hoc strategies tested	42	42	74	90
Number of models in all combinations candidate set	2048	2048	2592	390
Number of models in final model set	14	8	36	27
w_i top model	0.32	0.37	0.55	0.56
Number of covariates in top model	Ψ: 3p: 1	Ψ: 1p: 1	γ_{SO}: 2; ε_{SO}: 2γ_{BO}: 1; ε_{BO}: 1p_{so}: 1	φ: 2γ: 2$p \neq c$: 1
w_i ΔAIC ≤ 2 (number of models)	0.63 (3)	0.84 (3)	0.77 (2)	0.95 (2)
w_i ΔAIC ≤ 5 (number of models)	0.93 (7)	0.98 (5)	0.98 (4)	0.95 (2)
w_i ΔAIC ≤ 10 (number of models)	>0.99 (13)	1.00 (8)	>0.99 (6)	0.99 (7)
w_i ΔAIC > 10 (number of models)	<0.01 (1)	0 (0)	<0.01 (30)	0.01 (20)
ΔAIC null model	10.48	9.22	95.07	60.76

Note: Akaike model weights (w_i) are recalculated based on a final model set and after models with uninformative parameters removed.



Stage definition	Selection criteria (tracking)		Simple occupancy				Dynamic occupancy			Pradel model			
			Bobcat		Gray fox		Owls			San Diego pocket mice			
		Non-target structure	Order p, Ψ	Order Ψ, p	Order p, Ψ	Order Ψ, p	Order p, Ψ_{SO}, Ψ_{BO}	Order p, Ψ_{BO}, Ψ_{SO}	Order Ψ_{SO}, Ψ_{BO}, p	Order p, φ, γ	Order p, γ, φ	Order φ, γ, p	Order γ, φ, p
Sequential by sub-model	Plausible combinations	null	(+) 1.0	(-) .30	(+) .50	(+) .72	**(+) .15**	**(+) .13**	(-) .18	**(+) .07**	(-) .07	**(+) .07**	(-) .07
		general	(-) .50	(-) .30	(+) .50	(+) .75	**(+) .16**	**(+) .13**	**(+) .20**	(-) .07	**(+) .07**	(-) .07	(-) .07
	≤0	null	(-) .07	(-) .07	**(+) .07**	**(+) .07**	**(+) .13**	**(+) .13**	(-) .13	**(+) .07**	(-) .07	**(+) .07**	(-) .07
		general	(-) .07	(-) .07	**(+) .07**	**(+) .07**	**(+) .13**	**(+) .13**	**(+) .13**	(-) .07	**(+) .07**	(-) .07	(-) .07
	≤0 (2)	null	(-) .07	(-) .07	**(+) .07**	**(+) .08**	**(+) .13**	**(+) .13**	(-) .13	**(+) .07**	(-) .07	**(+) .11**	(-) .07
		general	(-) .07	(-) .07	**(+) .07**	**(+) .08**	**(+) .13**	**(+) .13**	**(+) .13**	(-) .08	**(+) .07**	(-) .07	(-) .08
	≤0 (5)	null	(-) .07	(-) .09	**(+) .07**	**(+) .10**	**(+) .14**	**(+) .13**	(-) .13	**(+) .07**	(-) .08	**(+) .20**	(-) .21
		general	(-) .07	(-) .08	**(+) .07**	**(+) .10**	**(+) .14**	**(+) .13**	**(+) .26**	(-) .09	**(+) .08**	**(+) .17**	(-) .11
	≤2	null	**(+) .19**	(-) .11	**(+) .26**	**(+) .15**	**(+) .13**	**(+) .13**	(-) .13	**(+) .07**	(-) .07	**(+) .07**	(-) .07
		general	**(+) .19**	(-) .09	**(+) .26**	**(+) .15**	**(+) .13**	**(+) .13**	**(+) .13**	(-) .11	**(+) .11**	(-) .07	(-) .11
	≤2 (5)	null	**(+) .22**	(-) .13	**(+) .28**	**(+) .23**	**(+) .15**	**(+) .13**	(+) .27	**(+) .07**	(-) .08	**(+) .17**	(-) .21
		general	**(+) .22**	(-) .13	**(+) .28**	**(+) .25**	**(+) .14**	**(+) .13**	**(+) .26**	(-) .13	**(+) .11**	**(+) .17**	(-) .18
	≤5	null	(+) .63	**(+) .32**	**(+) .50**	(+) .52	**(+) .15**	**(+) .25**	**(+) .15**	**(+) .07**	(-) .10	**(+) .08**	(-) .11
		general	(+) .63	**(+) .30**	**(+) .50**	(+) .53	**(+) .15**	**(+) .25**	**(+) .17**	(-) .24	**(+) .20**	(-) .31	(-) .12
	≤10	null	(+) 1.0	(+) .96	(+) 1.0	(+) .88	**(+) .16**	**(+) .50**	**(+) .19**	**(+) .11**	(-) .10	**(+) .13**	(-) .13
		general	(+) 1.0	(+) .95	(+) .94	(+) .88	**(+) .16**	**(+) .50**	**(+) .20**	(-) .34	(+) .24	(-) .36	(-) .30

Total Akaike weight recovered: 100

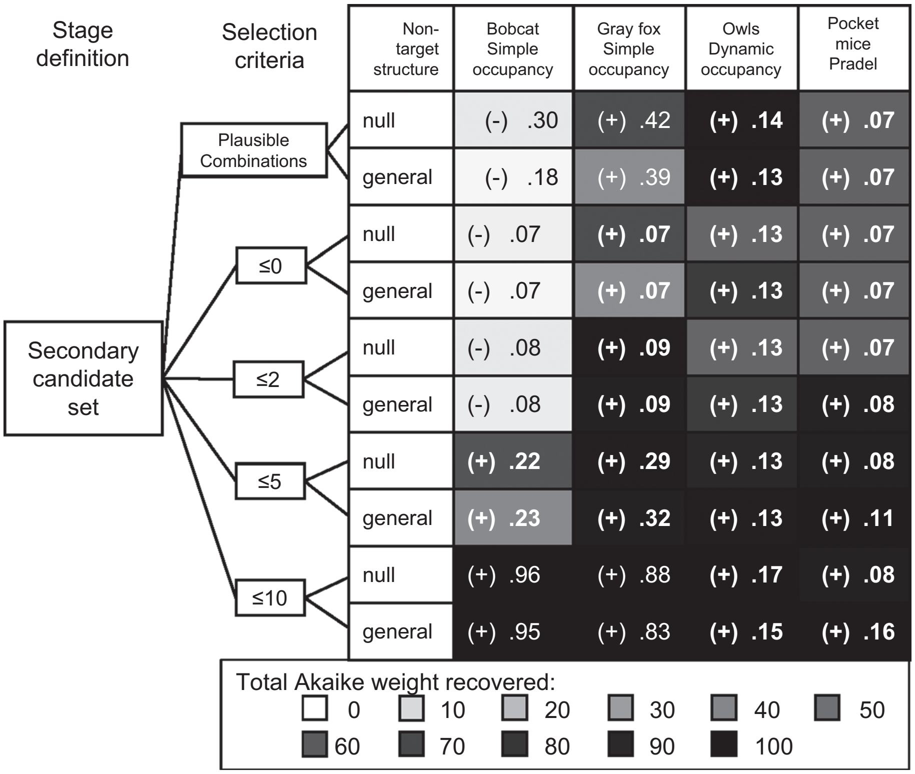
□ 0 ▢ 10 ▢ 20 ▢ 30 ▢ 40 ▢ 50 ▢ 60 ▢ 70 ▢ 80 ▢ 90 ■

Fig. 1. Multi-stage model selection strategy results when stage is defined by sub-model (sequential-by-sub-model). Selection criteria for inclusion of sub-model structures at each stage included ΔAICc thresholds (thresholds for additional tracking of structures in parentheses), and plausible combinations where ΔAIC ≤ 0, but sub-model structures with a greater likelihood than the selected structure are also carried forward in subsequent stages. Gray scale represents proportion of the total Akaike model weight recovered by a strategy. A (+) indicates the top-ranked model was recovered, while a (−) indicates it was not. The value in each cell is the proportion of the all combinations candidate set that was run. Strategies that recovered >50% of the total Akaike weight have white text. Strategies that selected the top-ranked model and required <50% of the model set are in bold.

when ΔAIC threshold = 5 (requiring 7–16% of the full model sets).

Tracking competing sub-model structures through stages and the plausible combinations approach did not routinely improve effectiveness of multi-stage strategies. The plausible combinations threshold identified the top model in 62% of the strategies and recovered >95% of the total Akaike model weights in only 21% of the strategies. Although the plausible combinations approach was intended for use with a secondary candidate set (Bromaghin et al. 2013), the dynamic occupancy was the only example for which plausible combinations with a secondary candidate set was entirely successful. Tracking reduced the number of models run compared to the equivalent ΔAIC threshold, but was also inconsistent, allowing for the omission of models with important covariates, especially in strategies with >2 stages.

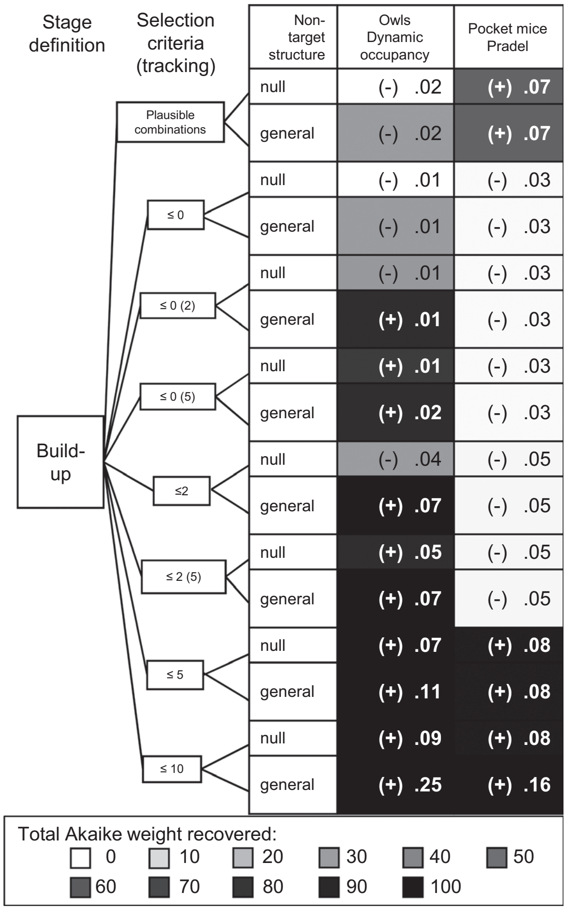
In all instances where a strategy failed to recover the overall all combinations top-ranked model, the misidentified top model lacked important structure in the sub-models describing demographic parameters of most interest (i.e., misidentification was not because of a failure to describe the nuisance parameters accurately). Failure to identify the covariates describing state and vital rates (i.e.,



Selection criteria	Non-target structure	Bobcat Simple occupancy	Gray fox Simple occupancy	Owls Dynamic occupancy	Pocket mice Pradel
Plausible Combinations	null	(-) .30	(+) .42	**(+) .14**	**(+) .07**
	general	(-) .18	(+) .39	**(+) .13**	**(+) .07**
≤0	null	(-) .07	**(+) .07**	**(+) .13**	**(+) .07**
	general	(-) .07	**(+) .07**	**(+) .13**	**(+) .07**
≤2	null	(-) .08	**(+) .09**	**(+) .13**	**(+) .07**
	general	(-) .08	**(+) .09**	**(+) .13**	**(+) .08**
≤5	null	**(+) .22**	**(+) .29**	**(+) .13**	**(+) .08**
	general	**(+) .23**	**(+) .32**	**(+) .13**	**(+) .11**
≤10	null	(+) .96	(+) .88	**(+) .17**	**(+) .08**
	general	(+) .95	(+) .83	**(+) .15**	**(+) .16**

Fig. 2. Multi-stage model selection strategy results when stage is defined by sub-model and sub-model sets are initially fit independent of selection of other sub-model structures (secondary candidate set). Selection criteria for inclusion of sub-model structures at each stage included ΔAICc thresholds, and plausible combinations where ΔAIC ≤ 0, but sub-model structures with a greater likelihood than the selected structure are also carried forward in subsequent stages. Gray scale represents total Akaike model weight recovered by a strategy. A (+) indicates the top-ranked model was recovered, while a (−) indicates it was not. The value in each cell is the proportion of the all combinations candidate set that was run. Strategies that recovered >50% of the total Akaike weight have white text. Strategies that selected the top-ranked model and required <50% of the model set are in bold.

occupancy or population dynamics) is problematic whether the goal of a study is variable selection or accurate prediction. In the bobcat example, the same sub-model structure for occupancy, Ψ(% urban + % agriculture), was consistently misidentified as the top model, ΔAIC = 5.48 compared to the all combinations set top-ranked model; Ψ(distance to structures + distance to streams + % agriculture), omitting two covariates and inappropriately including another. Top-ranked demographic sub-model structures resulting from multi-stage selection strategies for the dynamic occupancy and Pradel model examples could vary widely from the all combinations model set (Tables 3 and 4). In the owl dynamic occupancy example, the original analysis (Yackulic et al. 2014) and subsequent analyses (Dugger et al. 2016) in other study areas have consistently identified that both barred and spotted owls have higher territorial extinction rates in co-occupied territories; however, three of the multi-stage selection strategies selected a top model that only included an impact of competition on spotted owls. In the pocket mouse Pradel model example, there was evidence of leak (i.e., where the structure of one sub-model is inappropriately assigned to another sub-model), such as when the plot × time sub-model structure was assigned to seniority instead of apparent



Stage definition	Selection criteria (tracking)	Non-target structure	Owls Dynamic occupancy	Pocket mice Pradel
Build-up	Plausible combinations	null	(-) .02	**(+) .07**
		general	(-) .02	**(+) .07**
	≤ 0	null	(-) .01	(-) .03
		general	(-) .01	(-) .03
	≤ 0 (2)	null	(-) .01	(-) .03
		general	**(+) .01**	(-) .03
	≤ 0 (5)	null	**(+) .01**	(-) .03
		general	**(+) .02**	(-) .03
	≤2	null	(-) .04	(-) .05
		general	**(+) .07**	(-) .05
	≤ 2 (5)	null	**(+) .05**	(-) .05
		general	**(+) .07**	(-) .05
	≤ 5	null	**(+) .07**	**(+) .08**
		general	**(+) .11**	**(+) .08**
	≤ 10	null	**(+) .09**	**(+) .08**
		general	**(+) .25**	**(+) .16**

Fig. 3. Multi-stage model selection strategy results when stage is defined by adding model complexity (build-up strategy) across sub-models. Selection criteria for inclusion of sub-model structures at each stage included ΔAICc thresholds (thresholds for additional tracking of structures in parentheses), and plausible combinations where ΔAIC ≤ 0, but sub-model structures with a greater likelihood than the selected structure are also carried forward in subsequent stages. Gray scale represents total Akaike model weight recovered by a strategy. A (+) indicates the top-ranked model was recovered, while a (−) indicates it was not. The value in each cell is the proportion of the all combinations candidate set that was run. Strategies that recovered >50% of the total Akaike model weights have white text. Strategies that selected the top-ranked model and required <50% of the model set are in bold.

survival, or when the PC2 habitat covariate describing to seniority was attributed to apparent survival (Table 4).

Discussion

Although there is a recognized need for efficient multi-stage model selection strategies, the potential effects and biases of strategies are rarely considered and there is not general consensus on best practices. Our results indicate that not all multi-stage model selection strategies are created equal, and the resulting inference may deviate greatly from the respective all combinations (Doherty et al. 2012) candidate sets depending on the strategy chosen. The inconsistency in variable importance and omission of important demographic sub-model structures in the dynamic occupancy and Pradel model examples suggests that multi-stage model selection strategies with >2 stages are especially susceptible to mis-interpretation when the top-ranked model is not identified. Symmetrical approaches (i.e., where fit of one sub-model is not conditional on selection of structures for other sub-models) including build-up (Lebreton et al. 1992, Catchpole et al. 2004) and constructing a secondary set (Bromaghin et al. 2013) were more dependable, but only when relatively liberal model selection criteria were employed at each stage (retaining all models within ΔAIC ≥ 5–10). Using a symmetrical approach with a lenient selection threshold would be particularly important when the same variable such a habitat condition could apply to multiple sub-models (e.g., detection and occupancy or abundance) and leak could result in the variable being fit to the wrong sub-model (Catchpole et al. 2004).

Following our recommendations for general approach and thresholds, we found sub-model structure for non-target components (those components held constant while structures of other components are fit) was relatively unimportant, and tracking (keeping track of sub-model structures within a more lenient ΔAIC threshold at each stage for potential inclusion in the final stage) did not consistently improve success. In addition, no multi-stage strategy could overcome covariate combination dredging (using all possible combinations of all covariates for each sub-model instead of constructing candidate sets consisting of specific combinations of covariates representing distinct hypotheses). For example, when combinations of every covariate were used in the simple occupancy example for the all

Table 3. Demographic sub-model structures (not including nuisance sub-model structures for detection) of misidentified top-ranked models resulting from multi-stage model selection strategies for the owl dynamic occupancy example.

Sub-model structures	ΔAIC	γ_{SO}	γ_{SO}	γ_{SO}	γ_{BO}	ε_{SO}	ε_{SO}	ε_{SO}	ε_{BO}
All combinations top-ranked model	0	Year	Forest	No BO effect	SO effect	Autologistic	No habitat	BO effect	SO effect
Misidentified top-ranked model 1 (1)	15.13	**Constant**	**Riparian**	No BO effect	**No SO effect**	**Year**	**Riparian**	BO effect	**No SO effect**
Misidentified top-ranked model 2 (3)	3.09	**Autologistic**	Forest	No BO effect	SO effect	**Year**	No habitat	BO effect	SO effect
Misidentified top-ranked model 3 (1)	8.12	Year	Forest	No BO effect	**No SO effect**	**Constant**	No habitat	BO effect	**No SO effect**
Misidentified top-ranked model 4 (1)	7.87	Year	Forest	No BO effect	**No SO effect**	Autologistic	No habitat	BO effect	**No SO effect**
Misidentified top-ranked model 5 (3)	1.78	Year	Forest	No BO effect	SO effect	**Constant**	No habitat	BO effect	SO effect
Misidentified top-ranked model 6 (1)	1.78	Year	Forest	No BO effect	SO effect	**Year**	No habitat	BO effect	SO effect

Notes: Variables that were misidentified are in bold. The difference in AIC (ΔAIC) compared to the all combinations model set top-ranked model is shown. Sub-models of interest for the owl dynamic occupancy example include probability of a northern spotted owl colonizing (γ_{SO}) or becoming locally extirpated from a site (ε_{SO}), and probability of a barred owl colonizing (γ_{BO}) or becoming locally extirpated from a site (ε_{BO}).

combinations sets, multimodel inference reflected greater uncertainty (lower model weights for top models compared to dynamic occupancy and Pradel model sets with covariates representing structured hypotheses) and multi-stage strategy results were inconsistent. However, when multi-stage approaches were based on well-thought-out candidate sets and incorporated lenient thresholds in selection at each stage, inference was the same as if all models in the candidate set had been fit and number of models fit was substantially reduced.

The ΔAIC threshold was important and we found the threshold often applied for final model selection inference (ΔAIC ≤ 2) was insufficient for within-stage model selection, often excluding sub-model structures with strong support in the final all combinations model sets. Tracking sub-model structures decreased the number of models run, but did not consistently increase total Akaike model weights recovered. The plausible combinations selection criteria also commonly failed to recover ≥95% of the total Akaike model weights or identify the top model. However, we only implemented the plausible combinations approach using the top-ranked model at each stage as was described in the previous simulation study (Bromaghin et al. 2013). Increasing the ΔAIC threshold would likely improve performance, especially for data sets with support for more complex sub-model structures. However, in our examples increasing the threshold would have also increased the number of models run to at least that of the equivalent ΔAIC selection criteria.

The sequential-by-sub-model approach was unreliable for all example data sets. Doherty et al. (2012) found parameter estimates generated using a *p*-first or φ-first sequential-by-sub-model strategy for a CJS data type were not substantially biased, especially with inclusion of model averaging. However, both strategies created unbalanced model sets and overestimated importance of covariates simulated to have low correlation with either parameter (Doherty et al. 2012), similar to the results of the bobcat simple

Table 4. Demographic sub-model structures (not including nuisance sub-model structures for detection) of misidentified top-ranked models resulting from multi-stage model selection strategies for the pocket mouse Pradel model example.

Sub-model structures	ΔAIC	ϕ	ϕ	ϕ	ϕ	γ	γ	γ	γ
All combinations top-ranked model	0	Plot × time	No habitat effect	No burn effect	No rain effect	No plot or time effect	Habitat component PC2	No burn effect	Rain effect
Misidentified top-ranked model 1 (5)	21.44	**No plot or time effect**	No habitat effect	No burn effect	No rain effect	No plot or time effect	**No habitat effect**	No burn effect	Rain effect
Misidentified top-ranked model 2 (1)	10.24	**No plot or time effect**	No habitat effect	No burn effect	No rain effect	***Plot x time***	**No habitat effect**	No burn effect	**No rain effect**
Misidentified top-ranked model 3 (13)	6.85	**No plot or time effect**	No habitat effect	**1 fire effect**	No rain effect	No plot or time effect	**No habitat effect**	**1 fire effect**	Rain effect
Misidentified top-ranked model 4 (2)	9.06	**No plot or time effect**	No habitat effect	**1 fire effect**	No rain effect	**Time effect**	**No habitat effect**	No burn effect	**No rain effect**
Misidentified top-ranked model 5 (12)	8.62	**No plot or time effect**	No habitat effect	**1 fire effect**	No rain effect	No plot or time effect	**No habitat effect**	No burn effect	**No rain effect**
Misidentified top-ranked model 6 (1)	21.04	**No plot or time effect**	***Habitat component PC2***	No burn effect	No rain effect	No plot or time effect	**No habitat effect**	No burn effect	Rain effect
Misidentified top-ranked model 7 (1)	10.71	**No plot or time effect**	***Habitat component PC2***	No burn effect	No rain effect	***Plot x time***	**No habitat effect**	No burn effect	**No rain effect**
Misidentified top-ranked model 8 (1)	21.56	**Time effect**	No habitat effect	No burn effect	No rain effect	No plot or time effect	**No habitat effect**	No burn effect	Rain effect
Misidentified top-ranked model 9 (2)	22.17	**Time effect**	No habitat effect	No burn effect	No rain effect	No plot or time effect	Habitat component PC2	No burn effect	Rain effect
Misidentified top-ranked model 10 (4)	2.77	Plot × time	No habitat effect	No burn effect	No rain effect	No plot or time effect	**No habitat effect**	**1 fire effect**	Rain effect
Misidentified top-ranked model 11 (6)	9.68	Plot × time	No habitat effect	No burn effect	No rain effect	**Time effect**	**No habitat effect**	No burn effect	**No rain effect**

Notes: Variables that were misidentified are in bold. Those that were misidentified as a product of leak (Catchpole et al. 2004) where the structure of one sub-model is inappropriately assigned to another sub-model are italicized. The difference in AIC (ΔAIC) compared to the all combinations model set top-ranked model is shown. Sub-models of interest for the pocket mouse Pradel model example include apparent survival (ϕ) and seniority (γ).

occupancy example set in this study. Mistaken inclusion of unimportant covariates increased in sequential-by-sub-model strategies with >2 sub-models, especially for the Pradel model example with a balanced set of possible covariates and structures for both φ and γ (a result of leak; Catchpole et al. 2004). The build-up and secondary set approaches were both superior to the sequential-by-sub-model approach at recovering the total Akaike model weights and identifying the top model, and did so efficiently by fitting a small proportion of the all combinations model set.

The build-up strategy is not an option in some cases such as the single-season occupancy examples we examined. In these situations, constructing a secondary set is preferential to a sequential-by-sub-model approach, especially if the study objectives include inferences about drivers of the demographic process (i.e., covariates on occupancy) in addition to producing accurate estimates for prediction (Bromaghin et al. 2013). We strongly advocate thoughtful consideration for covariate combinations and discourage using all combinations of any possible covariate that might have some effect on a sub-model

(Burnham and Anderson 2002). For example, in the simple occupancy examples, all seven single covariates for ψ can be explained in an ecological context and reasonably expected to influence bobcat or gray fox occupancy. However, fewer covariate combinations can be construed as reasonable formal hypotheses (see Lesmeister et al. 2015 for examples of thoughtfully constructed candidate sets of covariate combinations for each species). In the simple occupancy bobcat example, multi-covariate models of ψ received greatest support, but support was similar among multi-covariate models and there was not consistent support for any covariate or combination for occupancy (Table 2). However, in the gray fox example there was a strong signal of two correlated covariates with gray fox occupancy increasing with forest cover ($w_i = 0.53$) and decreasing with agriculture ($w_i = 0.43$), as would be expected for the species (Cooper et al. 2012). In this case, the top model was identified in all strategies and loss of total Akaike model weights recovered only resulted when the second-best supported covariate (correlated with the most supported covariate) was removed with strict within-stage model selection criteria.

Our results emphasize that multi-stage approaches cannot compensate for indecision regarding suites of covariates and appropriate combinations in candidate set construction. Multimodel inference was never intended to divine some unidentifiable hidden truth in the data, but to make inference about a priori hypotheses conditional on the proposed candidate set (Burnham and Anderson 2002). While excluding important covariates can result in unmodeled heterogeneity in parameters and reduced accuracy in predictions, inclusion of many covariates with weak effects instead sacrifices precision in parameter estimates, resulting in considerable uncertainty in predictions and potentially misconstrued relationships (Freedman 1983, Lukacs et al. 2010). Thus, before ecologists make decisions about what multi-stage strategy to implement, they still must first make hard decisions about the specific hypotheses they care to evaluate (Burnham and Anderson 2002).

Ultimately, choice of model building strategy should be linked to the objectives of the study including consideration of generality, realism, and precision (Levins 1966). We used total Akaike model weight recovered and identification of the top-ranked model in an all combinations model set as measures of success to assess whether inference related to covariates and selected models would be the same. However, other standards may be more appropriate when prediction, or power to detect change in demographic parameters are primary objectives.

When employing a multi-stage strategy instead of constructing the full all combinations model set, we recommend using a build-up or secondary set approach to defining stages, rather than a sequential-by-sub-model approach. Second, we recommend using liberal ΔAIC thresholds for within-stage sub-model structure selection, recognizing that a primary purpose of multimodel inference is to incorporate uncertainty in support among competing hypotheses, and not to select a single best model (or within-stage sub-model structure). We also encourage standard reporting of all decisions in developing and using a multi-stage approach. Often the finer details of the selection strategy are not described in the methods, hindering assessment of results. At minimum, we suggest clearly describing the stage definition, progression, non-target structures, and the within-stage model selection criteria.

Finally, our examples demonstrate that the success of a model building strategy can depend on the data collected, the model set implemented, and the strength of the candidate set. Akin to how we simulate data to ensure appropriate study design, so should we explore the possible impacts of multi-stage model selection strategies using reduced data or candidate sets to ensure consistent inference from results. For example, based on the results of this study we have identified a build-up approach with ΔAIC threshold ≥5 as a promising strategy for further analysis of the full San Diego pocket mouse data set, which will include too many occasions and individuals to feasibly fit all model combinations.

Acknowledgments

Funding to support southern Illinois data collection was provided by Illinois Department of Natural Resources via Federal Aid in Wildlife Restoration Project W-135-R. JEDs contribution was funded by the Land Change Science Program at USGS. DJMs contribution was funded by USDA National Institute of Food and Agriculture, McIntire Stennis project

(1020959). Funding for spotted owl and barred owl data collection was provided by the USDA Forest Service and USDI Bureau of Land Management. Funding for the southern California data set was provided by Joint Fire Science Program (no. 042194), the Blasker Environment Grant Program of the San Diego Foundation, and California Department of Fish and Game provided access to Rancho Jamul Ecological Reserve. We thank the many technicians that contributed to collection of data. We thank Jim Hines for assistance with batch processing scripts for PRESENCE, Scott Tremor, Dana Hogan, Genie Flemming, Jenny Duggan for assistance with the southern California data set, and Suresh Sethi for conversations, and Jim Nichols and two anonymous reviewers for constructive reviews critical to the development and improvement of this manuscript. Any use of trade, product, or firm names is for descriptive purposes only and does not imply endorsement by the U.S. government.

Literature Cited

Arnold, T. W. 2010. Uninformative parameters and model selection using Akaike's Information Criterion. Journal of Wildlife Management 74:1175–1178.

Bromaghin, J. F., T. L. McDonald, and S. C. Amstrup. 2013. Plausible combinations: An improved method to evaluate the covariate structure of Cormack-Jolly-Seber mark-recapture models. Open Journal of Ecology 3:1–11.

Buckland, S. T., K. P. Burnham, and N. H. Augustin. 1997. Model selection: An integral part of inference. Biometrics 53:603–618.

Burnham, K. P., and D. R. Anderson. 2002. Model selection and multimodel inference: a practical information-theoretic approach. Second edition. Springer-Verlag, New York, New York, USA.

Catchpole, E. A., Y. Fan, B. J. Morgan, T. H. Clutton-Brock, and T. Coulson. 2004. Sexual dimorphism, survival and dispersal in red deer. Journal of Agricultural, Biological, and Environmental Statistics 9:1–26.

Cooper, S. E., C. K. Nielsen, and P. T. McDonald. 2012. Landscape factors affecting relative abundance of gray foxes *Urocyon cinereoargenteus* at large scales in Illinois, USA. Wildlife Biology 18:366–373.

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

Doherty, P. F., G. C. White, and K. P. Burnham. 2012. Comparison of model building and selection strategies. Journal of Ornithology 152:317–323.

Dugger, K. M., et al. 2016. The effects of habitat, climate, and barred owls on long-term demography of northern spotted owls. Condor 118:57–116.

Freedman, D. A. 1983. A note on screening regression equations. American Statistician 37:152–155.

Hines, J. E. 2006. PRESENCE2-Software to estimate patch occupancy and related parameters. United States Geological Survey, Patuxent Wildlife Research Center, Laurel, Maryland, USA.

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

Laake, J. L. 2013. RMark: An R interface for analysis of capture-recapture data with MARK. AFSC Processed Report 2013–01. Alaska Fisheries Science Center, National Marine Fisheries Service, Seattle, Washington, USA.

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

Lesmeister, D. B., C. K. Nielsen, E. M. Schauber, and E. C. Hellgren. 2015. Spatial and temporal structure of a mesocarnivore guild in Midwestern North America. Wildlife Monographs 191:1–61.

Levins, R. 1966. The strategy of model building in population biology. American Scientist 54:421–431.

Lukacs, P. M., K. P. Burnham, and D. R. Anderson. 2010. Model selection bias and Freedman's paradox. Annals of the Institute of Statistical Mathematics 62:117–125.

MacKenzie, D. I., and J. E. Hines. 2017. RPresence: R Interface for Program PRESENCE. R package version 2.12.7. https://www.mbr-pwrc.usgs.gov/software/presence.html

MacKenzie, D. I., J. D. Nichols, G. B. Lachman, S. Droege, J. A. Royle, and C. A. Langtimm. 2002. Estimating site occupancy rates when detection probabilities are less than one. Ecology 83:2248–2255.

MacKenzie, D. I., J. D. Nichols, M. E. Seamans, and R. J. Gutiérrez. 2009. Modeling species occurrence dynamics with multiple states and imperfect detection. Ecology 90:823–835.

Miller, D. A. W., C. S. Brehme, J. E. Hines, J. D. Nichols, and R. N. Fisher. 2012. Joint estimation of habitat dynamics and species interactions: Disturbance reduces co-occurrence of non-native predators with an endangered toad. Journal of Animal Ecology 81:1288–1297.

Pradel, R. 1996. Utilization of capture-mark-recapture for the study of recruitment and population growth rate. Biometrics 52:703–709.

R Core Team. 2016. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

Seber, G. A. F.. 1965. A note on the multiple-recapture census. Biometrika 52:249–259.

White, G. C., and K. P. Burnham. 1999. Program MARK: survival estimation from populations of marked animals. Bird Study 46:S120–S139.

Yackulic, C. B., J. Reid, R. Davis, J. E. Hines, J. D. Nichols, and E. Forsman. 2012. Neighborhood and habitat effects on vital rates: expansion of the barred owl in the Oregon Coast Ranges. Ecology 93:1953–1966.

Yackulic, C. B., J. Reid, J. D. Nichols, J. E. Hines, R. Davis, and E. Forsman. 2014. The roles of competition and habitat in the dynamics of populations and species distributions. Ecology 95:265–279.

Supporting Information

Additional Supporting Information may be found online at: http://onlinelibrary.wiley.com/doi/10.1002/ecs2.2997/full