DOM manuscript fragments

# Key citations

Merow 2014 – What do we gain from simplicity versus complexity in species distribution models?

* Comprehensive discussion of model ‘complexity,’ including number of parameters, number of features, and linearity of feature responses.

Araujo 2020 – Standards for distribution models in biodiversity assessments

* Covers ‘best practices’ for species distribution models and quantifies how often those are met.

Araújo and Guisan 2006 – Five (or so) challenges for species distribution modelling

* Two relevant challenges here – selection of predictors and model evaluation

Garcia-Callejas and M.B. Araújo 2016 – The effects of model and data complexity on predictions from species distribution models

* Looked at model complexity and predictive performance – found that more complex models did not perform better **or** worse, but simpler data was easier to model

Austin 2002 – Spatial prediction of species distribution: an interface between ecological theory and statistical modelling

* Compares ‘Proximal’ vs ‘Distal’ covariates; can link to directly measured/remotely observed
* Talks about how most ecological responses are non-linear (!)

Austin 2007 – Species distribution models and ecological theory: a critical assessment and some possible new approaches

* Further discusses the connection between ecological theory and models, with notes on model evaluation and reporting.
* Commentary on the different needs for prediction vs ecological explanation

Pearson 2006 – Model-based uncertainty in species range prediction

Lahoz-Monfort 2014 – Imperfect detection impacts the performance of species distribution models

* Good reference on importance of accounting for imperfect detection

Letten 2013 – the importance of temporal climate variability for spatial patterns in plant diversity

Barry and Elith 2006 – error and uncertainty in habitat models

* Discusses ‘missing covariates’ that may contribute to variation; these can introduce bias
* Discusses the nature of the ‘response surface,’ linearity etc.

Tobalske 2002 – Effects of spatial scale on the predictive ability of habitat models for the green woodpecker in Switzerland

Chave 2013 – The problem of pattern and scale in ecology - what have we learned in 20 years?

* Comprehensive discussion on the matter of scale – can be linked to local/large-scale occupancy, as well as short/long-term occupancy concepts.

Evans 2013 – Do simple models lead to generality in ecology?

* Argues in favour of more complex models, and criticises the adage that ‘simple is general is best.’

Lonergan 2014 – Data availability constrains model complexity, generality, and utility: a response to Evans et al.

* Criticises the strength of Evans 2013’s conclusions, says that complexity has its costs

Guisan 2002 – Generalised linear and generalised additive models in studies of species distributions: setting the scene

* More general glm paper; shortcomings with stepwise regressions

Maggini 2006 – Improving generalized regression analysis for spatial predictions of forest communities

* Tested different selection methods for SDMs; found no differences in performance

Johnson and Omland 2004 – Model selection in ecology and evolution

* Discusses basics of model selection with direct reference to mark-recapture literature, where they say model selection is more popular.
* Interesting note – mostly used to select structural elements of the model, not environmental covariates. This has carried over to DOMs

Huston 2002 – Introductory essay: critical issues for improving predictions (Predicting species occurrences: issues of accuracy and scale)

Zurell 2020 – A standard protocol for reporting species distribution models

* Provides framework for reporting on aspects of the modelling process; outlines key steps in modelling.

Guisan and Thuiller 2005 – Predicting species distribution: offering more than simple habitat models

* Early review of SDMS; hit key points on selection, evaluation, complexity, etc.

Randin 2006 – Are niche-based species distribution models transferable in space?

* Not often?

Olden and Jackson 2000 – Torturing data for the sake of generality – how valid are our regression models?

* Early critique of model selection for regression models. Show that many methods are quite flawed in different scenarios.
* Also states importance of validation

Araujo 2005 – validation of species-climate impact models under climate change

* Describes paradigms of evaluation/validation: same-data fit, data splitting, independent validation sets (gold standard)

Mackenzie 2004 – Assessing the fit of site-occupancy models

* Goodness of fit test (via parametric bootstrapping) for the single season occupancy model; does not mention the multi-season version.

Broms 2016 – Model selection and assessment for multi-species occupancy models

* Good discussion of model selection and evaluation in the Bayesian context – specific to multi-species single-season models, but has some relevance.

Stevens and Conway 2019 – Predicting species distributions: unifying model selection and scale optimization for multi-scale occupancy models

* Discusses the concept of scale in occupancy
* Shows that model selection based on predictive performance outperforms information-criterion based methods

Hooten and Hobbs 2015 – A guide to Bayesian model selection for ecologists

* Strong discussion of need for evaluation; better standards in ecology

Wright 2016 – A goodness-of-fit test for occupancy models with correlated within-season revisits

* Another single-season goodness of fit test

Rushing 2019

* Uses a GAM based extension to occupancy model; no explicit colonisation and extinction
* Good note on improved complexity

**Cases where GOF was used**

* Desrochers 2012, Basile 2017
* different method used for Mortelli 2009 following Moore and Swihart 2005
* Croose 2018 uses GOF from AICcModelAvg by Mazzerolle 2017
* Walkup 2018 uses GOF via parboot – says its debated, but not how…
* ‘unmarked’ documentation calls parametric bootstraps experimental.

# Required citations

* Goodness-of-fit tests for occupancy models
* Any other selection papers for occupancy models
* Occupancy model complexity?

# Introduction

For any model estimating occupancy, key environmental drivers must be adequately represented using covariates — the SDM community has long debated how to best approach model selection and evaluation [@guisan2006; @araújo2006; @jarnevich2015; @araújo2019]. The DOM literature has by comparison focused less on these aspects of modelling [@mackenzie2017], although what research does exists indicates that choice of model selection approach can impact model performance [@morin2020].

More generally, identifying which covariates to include in a model is a challenging task made more difficult by the greater number of parameters included in DOMs. The latent nature of the main response variable (that is, occupancy state) also complicates efforts to assess model performance and compare models.

# Methods

# Usage, applications, and potential

DOMs display great variation of scale in all senses. Analyses in our samples range from the hyper-local at individual parks [@zúñiga-vega2019] to continental-scale studies [@zuckerberg2011]. Similar diversity in study duration and site quantity indicates further emphasises the breadth of studies which DOMs may be fit to. Method of species detection also need not be a limitation for the implementation of DOMs, as articles in our sample show that many approaches may be coerced into data suitable for this model class. While studies pre-designed to match hierarchical data inputs may be simplest to align with model assumptions, post-hoc manipulation of other data such as long-term citizen monitoring programs [@zuckerberg2011; @peach2019], disparate agency modelling programs [@mcgowan2020], or camera trapping data [@davis2018] is feasible with careful consideration. Detections from multiple sources, with associated variation in efficiency, have also been incorporated by authors [@warrier2020; @pitman2017].

Reviewed implementations of dynamic occupancy models show widespread representation from mammals, birds, and herptiles. This assemblage contrasts with the related stochastic patch occupancy models (SPOMs, @gutiérrez-arellano), where the majority of models are fit to invertebrates, plants, and fungi. This difference may reflect SPOMs' assumption of perfect detection and discrete habitat patches, limitations which are unlikely to be fulfilled by most vertebrate taxa. Aside from differences in assumptions around detection, SPOMs also differ from DOMs in that they are generally spatially explicit, with probabilities of colonisation and extinction dependent on patch size and the occupancy state of nearby patches. Where desired, these features may also added to dynamic occupancy models — @sutherland2014 presents an implementation which may be considered a fusion of the DOM and SPOM methods, and several other authors incorporate patch characteristics aspects into their models as covariates [@duggan2011; @broms2016].

While objectives vary in our sample, DOMs are rarely used to extrapolate beyond study areas — use of prediction remains the least popular use-case in our sample. Articles which do make predictions provide some of the most directly applicable outputs found in the review sample: @mcgowan2020 provides projections for a threatened species under multiple putative management scenarios, and @pollentier2021 generates maps of distributions more frequently associated with species distribution models (SDMs). DOMs offer advantages over other SDMs for this purpose, particularly where populations are assumed not to be at equilibrium as is the case for many threatened and invasive species. By explicitly accounting for changes in occupancy (i.e., colonisation and extinction) DOMs may better capture future range contraction or expansion — an aspect of prediction where other SDMs may struggle.

This flexibility is reflected in the types of authors who conducted the studies captured in our sample: while 86% of articles had at least one author based at an academic institution, 46% had author(s) at government institutions and 25% had author(s) from non-profit organisations. The high proportion of articles with government and NGO participants suggests a more applied focus for many users of DOMs, underscoring the importance that model outputs are reliable when they may be used to guide management or conservation actions.

# Modelling practices

A diverse set of covariates were identified in the model sample reflecting the diversity of systems and ecological questions described by DOMs. Notably, many models lack covariates on the initial occupancy parameter, raising questions on how the parameterisation of the model in it's first time-step may affect the quality of estimates at later time-steps. The distinction in covariates considered for the ecological parameters vs those considered for detection is also of note: ecological parameters more frequently contained covariates describing environmental features, where detection more frequently contained more structural covariates such as season. Covariate complexity is relatively low across the board, with only a fraction of models implementing either non-linear responses to covariates or interactions between covariates. While this addition complexity does have implications for model selection, it may be necessary to accurately describe many ecological relationships.

Model selection approaches diverge between Bayesian and frequentist implementations of the model. While model selection will naturally be practised differently between these two categories, it is unclear what effect a lack of model selection may have on the outputs produced by Bayesian models, and it is likely that Bayesian model selection approaches should be further explored @hooten2015. Model selection approaches in the Frequentist applications show limited consensus, although most do consider multiple models. Notably, @morin2020 found that the popular procedural approaches to model selection (where one sub-model is assessed at a time) may not provide optimal results. Considerably more research is required on this topic to clarify how authors should best approach covariate selection.

“Given the rising popularity of using such techniques in the analysis of ecological data, it is important to realize that they assume that the candidate set contains at least one model that fits the data adequately (Burnham and Anderson 1998, p. 73), and are not a substitute for assessing model fit. The selection of a “best” model(s) does not guarantee the selection of a “good” model.” – From Mackenzie & Bailey 2004

# Conclusions and key considerations

One of the strengths of the dynamic occupancy model is its considerable flexibility — as seen in our review, applications of DOMs over the past two decades have been fit to data for a diverse cohort of species, using a broad range of detection methods and survey designs. This flexibility permits quite a bit of creativity on the part of authors and has allowed for innovative use-cases well beyond simple estimations of occupancy. Despite this promise, questions do remain on the best practices for fitting DOMs. Studies in our sample are discordant on their approaches the modelling process, and there the literature gives little insight on how these approaches may differ in their ability to identify reliable models.

Implementation of dynamic occupancy models requires considerable thought to ensure that model outputs correspond with author expectations and ecological realities. Based upon our review of the existing literature we discuss key questions authors should carefully engage with during the modelling process.

1. **Are seasons ecologically meaningful, and how do they affect the definition of occupancy?** As previously discussed, the duration of the season has important implications for both the reliability and proper interpretation of model results. Seasons which are not ecologically meaningful may run afoul of the closure assumption resulting in associated biases. Also relevant is how the definition of 'occupancy' may shift with season length and life-history of a species — for many species, occupancy within a year and within a week may be considerably different concepts, with shorter intervals more closely relating to the idea of 'use.' With the proliferation of continuous detection methods such as camera traps and acoustic monitors where the definition of 'season' is to some degree arbitrary, it is increasingly important to thoughtfully consider how this element of the DOM is defined.
2. **Are key drivers of occupancy and detection incorporated?** While it is effectively impossible to fully describe heterogeneity across a landscape, authors should make their best efforts to capture the drivers of occupancy and detection most likely to influence their study species. This is true even when another covariate is of principal interest to investigators — failure to include other major determinants of occupancy is likely to bias parameter estimates. One place where this is of particular importance is in determining drivers of detection, which is not only subject to observational conditions but also ecological determinants which may cause the species to spend more time at a location and thus be more frequently available for detection.
3. **Ensure covariates reflect ecological complexity.** Many ecological relationships are non-linear and may require polynomial terms to be accurately represented. This is particularly true at large spatial scales where a larger portion of a species ecological niche is represented in surveyed data. Authors should question whether a linear term is sufficient to capture hypothetical relationships in their study systems, and consider quadratic terms as well as interactions between covariates where appropriate.
4. **Consider model evaluation.** Model evaluation remains rare in ecological research relative to other fields embracing complex modelling, likely due to a reluctance to 'throw away' data for use in testing. This concept should be challenged, particularly for larger scale long-term projects where reserving a fraction of data for testing is unlikely to substantially reduce precision of parameter estimates. For studies where this is genuinely not possible, reporting results of goodness-of-fit tests may still provide valuable information to readers.

### Priorities for future research

Work on dynamic occupancy modelling over the past 20 years has developed this model class into an important tool for describing species distributions and ecological relationships. Historically, methods-development research for dynamic occupancy models has focused on questions of model structure — namely ideas around the closure assumption and imperfect detection. This emphasis likely reflects the DOM's roots in the mark-recapture literature, where these topics have been discussed for several decades. Considerably less emphasis has been placed on the relation of model parameters to environmental factors via covariates and the model selection processes used to do so, despite this being an objective in a considerable fraction of studies fitting DOMs. This stands in stark contrast to related correlative species distribution models, where these topics have been the subject of considerable research over the years. There are opportunities for the DOM literature to learn from species distribution models, and more targeted research on the influence of the model selection process is necessary.

An important opportunity for future research relates to autonomous detection methods, such as camera traps and bio-acoustic monitors — these technologies have permitted collection of enormous quantities of data particularly well-suited to dynamic occupancy models, but have also raised additional questions around their implementation. The definition of seasons is of increased importance for these data types, where division of data into 'seasons' and 'observations' is arbitrary. Further research should explore the implications of this phenomenon and the additional modelling opportunities made possible by this novel data type.

The past two decades have shown that dynamic occupancy models are an excellent tool for addressing a wide range of ecological questions relating to species occupancy throughout space and time. The flexibility afforded by model extensions and the considerable advantages of the class over other species distribution models is likely to ensure that these models continue to increase in popularity. Given their strong use on threatened taxa and excellent potential to support environmental management, it is important that research continues to support model users in creating the most reliable models possible.