ARC ceed
Centre of Excellence for Environmental Decisions

decision brief A guide to good environmental decision making

Tasmanian Devils (Sarcophilus harrisii) Image. Karen Gillow

Developing species distribution models

A checklist of the basics

By Associate Professor Brendan Wintle, University of Melbourne

Species distribution models (SDMs) combine observations of species occurrence or abundance with information about environmental variables. This provides ecological insights and helps to predict species' distributions across landscapes. SDM outputs usually include:

(i) a document describing the data and statistical modelling process used to generate the SDM,

(ii) a summary of the resulting numerical or statistical model, and

(iii) a map of the probability of a species' occurrence over a region of interest.

Maps of the probability of occurrence generated by SDMs have been used in conservation planning and for the management of habitat at finer scales (eg, for identifying critical habitats to avoid during timber harvesting). They have also been used widely in the development of spatially explicit population viability analysis models (see Decision Brief 3.3). The underlying statistical models of SDMs are useful for ecological inference, allowing analysts to statistically quantify the influence of particular environmental variables (eg, climate and rainfall) or management practices (eg, harvested/ not harvested) on the probability that a species will occupy a given area.

The quality and reliability of SDMs depends on the quality and appropriateness of the field survey and environmental data inputs, the expert opinion and statistical methods underpinning the modelling process, and the recognition and treatment of uncertainty in the model outputs. SDMs may be referred to as habitat models, ecological niche models and a host of other names.

Elements of good design

1. Survey data types
2. Evaluate data quality
3. Environmental data
4. Site-level or visit-level variables in SDMs
5. Match scales of variables
6. Identifying candidate models
7. Model ranking
8. Dependent and independent variables
9. Statistical evaluation & cross-validation
10. Model diagnostics
11. Presenting plausible competing models
12. Simplifying maps

1. Survey Data Types

✓ Is the data type being used justified given data availability and the purpose of the SDM? Will the resulting SDM outputs be informative and meaningful given project objectives?

There are three broad types of species survey data which can be used to build SDMs: presence-only data, presence-absence data (also known as detection/non-detection data), and abundance data. It is important to ensure that the data type being used is appropriate for the proposed use of the SDM. Most SDMs are built using presence-only data, which means that the predictions of the SDM cannot usually be interpreted as probabilities but rather an indication of the relative suitability of different parts of the environment. When imperfect detection of species during biological surveys is an issue (see '*Understanding species distributions*, Decision Brief #3.1), data can be collected in a way which allows simultaneous modelling of detectability and occupancy or abundance. Imperfect detectability is an especially important issue when the aim is to estimate occurrence probabilities or absolute abundance across a region.

2. Evaluate data quality

✓ Has an evaluation of the quality, age, bias and geographic and environmental coverage of the data been undertaken? Have data gaps been filled? Have old and poor quality data been excluded from model development? Is the data of sufficient quality given the likely application of the model, and the accuracy and precision demanded of the SDM output?

Ideally, biological survey data collection would be planned in advance of a distribution modelling exercise to ensure appropriate stratification and replication across key environmental gradients such that the resulting models would provide a faithful representation of species habitat requirements. However, this is seldom the case and more often existing data must be used. Therefore, a review should be undertaken of the quality, age, bias, and geographic and environmental coverage of the existing species data being incorporated in the modelling.

The need to collect new data to fill gaps in coverage/stratification or to provide sufficient replication to develop reliable SDMs should be evaluated. Old and spatially- or environmentally-biased data (eg, data which have been collected close to roads, only in forests on particular tenures, only within a small subset of the potential range of interest) can be misleading in SDMs developed for the purposes of defining critical habitats, total current habitat extent, and estimating overall proportion of the landscape occupied by the species. Using data derived from surveys with insufficient sampling effort to detect species of interest will lead to poorly calibrated models (see '*Understanding species distributions*, Decision Brief #3.1).

3. Environmental data

✓ Has a review of the available mapped GIS environmental data layers in relation to key habitat requirements of the species been undertaken? Have gaps in availability of key habitat variables been filled and/or addressed as necessary?

Spatially referenced data layers (generally stored in GIS) are often used as 'independent variables' in the construction of SDMs and in the process of predicting the probability or relative likelihood of occurrence of a species throughout the landscape. The quality and relevance of the environmental data determines the quality of SDM output maps. The environmental data used in SDM development must represent the key habitat requirements of the species if the resulting SDMs are to be a useful guide to the location of critical habitat or the total extent of suitable habitats. If significant components of a species habitat requirements are not mapped (or are not mappable), reliable and useful SDM predictive maps cannot be produced.

4. Site-level or visit-level variables in SDMs

✓ Have relevant site-level variables been included or considered in the model development?

It is common in ecological or impact assessment studies to model independent variables that are measured at survey locations even though they are not mapped or mappable at broad scales. Even when the aim is to model and predict habitat occupancy across large areas, the incorporation of site- or visit-level variables can be used to 'control' for site- or visit-level effects in the modelling process (improving estimation of the importance of mapped variables). However, if the aim is to use models to map suitable habitat across a landscape, region or continent, at least some of the variables in the model must be mapped at the scale of interest.

5. Match scales of variables

✓ Have independent environmental variables included in SDMs been mapped at a resolution and accuracy appropriate for management?

The scale (resolution and extent) of mapped environmental variables must match the scale of use and the species biology. Environmental variables may be mapped at different resolutions depending on the nature of the variable and the data collection method used. For example, digital elevation data may be able to detect subtle variation in elevation over distances of less than a few metres, while variation in soil properties can rarely be mapped with spatial accuracy better than 200 metres over large areas of forest, and variation in climatic variables is seldom relevant over less than 2 kilometres.

If management decisions demand accurate mapping of habitats to within 10 metres and fine scale variation in habitat does matter to a species of interest, then it is crucial that the relevant habitat variables are accurately mapped at an appropriately fine scale. For this reason, SDMs constructed solely on coarsely mapped climate variables are unlikely to be useful for supporting habitat management decisions in which boundaries must be accurately mapped to within 5-10 metres. However, combinations of coarsely and finely mapped habitat variables may be used to develop SDMs capable of informing habitat management decision making with a high degree of spatial precision.

6. Identifying candidate models

✓ Has the approach to model identification been justified with respect to published advice, and the availability of species survey data and expert opinion?

Good practice in SDM development involves using a combination of expert judgement and statistical analysis to identify a set of plausible competing models that do a good job of explaining observed variation in species occupancy data, but which are not too complex ('over-fitted') relative to the amount of species data available.

A few approaches to model selection exist, which emphasise the role of the expert opinion and statistical analysis to a greater or lesser degree. The choice of the most appropriate approach depends on the level of access to experts, the degree to which experts agree with each other, the amount of species survey data available and the degree to which statistical models agree with expert opinion. There is ample published advice on how to identify a set of plausible competing models.

7. Model ranking

✓ Has an information criterion score been presented for each plausible model being considered?

It is unusual that a single model far outperforms all other competing models. This is known as model selection uncertainty because we are generally uncertain about which of a set of plausible models is the best. Therefore, it is good practice to present all of the models which provide plausible representations of how environmental variables influence the probability that a site will be occupied by the species of interest. Each model can be evaluated using an information criterion such as Akaike's Information Criterion which assesses how well the model strikes a balance between explaining variation in the species occurrence or abundance data and minimising complexity. The information criterion score for each model should be presented to enable reviewers to assess the relative plausibility of the competing models.

8. Dependent and independent variables

✓ Have the statistical relationships between environmental variables and species occupancy (or abundance) been provided in a clear graphical form?

The terms (or 'coefficients') in a statistical model indicate how the dependent variable (eg, probability of species occurrence or species abundance) varies with changes in an environmental variable. In many cases, these relationships may be non-linear functions and may also be used to represent the strength of interactions between independent variables that impact on species occupancy or abundance. These relationships should be represented graphically (see Resources) to allow reviewers to confirm that relationships are ecologically plausible.

9. Statistical evaluation and cross-validation

✓ Has model evaluation been undertaken (and model predictive performance reported)?

Model evaluation is a critical part of SDM development because it provides an indication of the robustness of predictions of species distribution. Ideally, model evaluation is undertaken by comparing the predictions of the model to new field observations that were not used in the construction of the model. Because it is not always possible to collect a completely new model evaluation field data set, it is common to separate the modelling data into model-fitting and model-evaluation sets within a process called cross-validation. There is ample published advice on practical approaches to model evaluation including cross-validation (see Resources).

10. Model diagnostics

✓ Have model diagnostics been presented and discussed?

Model diagnostics indicate any statistical issues that may impair the ability of the model to provide good inference about key habitat dependencies or good predictions of the spatial distribution of the species and its critical habitats. For example, model diagnostics can indicate whether model residuals (the differences between observed and fitted values) are normally distributed. If they are not, then model coefficient estimates may not be reliable. Presenting and discussing the outcome of model diagnostic tests for the most plausible models in the model set is good practice.

11. Presenting plausible competing models

✓ Have output habitat maps produced by the 'best model' and the 'most competitive' of the non-preferred plausible models been presented and considered? Has the relative statistical support for each model been presented using AIC or equivalent criteria?

It is common practice to present SDM output maps (habitat maps) for only the 'best' of the plausible competing SDMs (assessed using AIC or cross-validation evaluation). This is reasonable because it is not ideal for managers to be dealing with more than one map of habitat for a particular species. However, there is a danger in this practice because plausible alternative models may be capturing something about the environment that is important to the species but which is overlooked by the 'best' model. Therefore, it is important that all areas predicted to be important habitat by plausible competing models are represented so that they can be considered and ground-checked during management decision-making processes.

12. Simplifying maps

✓ Where map outputs must be classified into categories, has this been done in accordance with published advice?

It is common practice to convert maps which provide continuous predictions of occurrence probability or abundance into a binary map representing habitat/not habitat or other categorical presentations such as high/moderate/low quality habitat. This is achieved by identifying a threshold level of occupancy probability (habitat quality or abundance), below which areas will be considered unsuitable. If handled incorrectly, this process can lead to the loss of important habitat areas and is generally not recommended. However, if there is a compelling reason to do so, there is published advice on how to identify defensible thresholds for habitat maps.

More information: Associate Professor Brendan Wintle (Univesity of Melbourne), b.wintle@unimelb.edu.au, (07) 3365 3836

Resources

Araujo MB & AT Peterson (2012) Uses and misuses of bioclimatic envelope modeling. *Ecology* 93: 1527-1539 (See Decision Point #85)

Austin MP (2002) Spatial prediction of species distribution: an interface between ecological theory and statistical modelling. *Ecological modelling* 157: 101-118

Austin MP & KP Van Niel (2011) Improving species distribution models for climate change studies: variable selection and scale. *Journal of Biogeography* 38: 1-8

Elith J, CH Graham, RP Anderson et al. (2006) Novel methods improve prediction of species' distributions from occurrence data. *Ecography* 29: 129-151

Elith J, JR Leathwick, T Hastie (2008) A working guide to boosted regression trees. *Journal of Animal Ecology* 77: 802-813

Elith J & JR Leathwick (2009) Species distribution models: ecological explanation and prediction across space and time. *Annual Review of Ecology, Evolution, and Systematics* 40: 677-697

Franklin J (2009) *Mapping Species Distributions*. Cambridge University Press, Cambridge

Guisan A & NE Zimmerman (2000) Predictive habitat distribution models in ecology. *Ecological modelling* 135: 147-186

Wintle BA, J Elith & J Potts (2005) Fauna habitat modelling and mapping; A review and case study in the Lower Hunter Central Coast region of NSW. *Austral Ecology* 30: 719-738

CEED is an Australian Research Council (ARC) partnership between Australian and international universities and research organisations. Our vision is to be the world's leading research centre for solving environmental management problems and for evaluating the outcomes of environmental actions.

www.ceed.edu.au

ARC Centre of Excellence for Environmental Decisions
The University of Queensland
St Lucia, QLD 4072, Australia
(07) 3365 2527
k.gillow@uq.edu.au