This is a <u>video transcript</u> of Module 4 of the Online Open Course in Species Distribution Modelling. To access the full suite of videos click [here](#).

Online Open Course - Species Distribution Modelling

Powered by the [Biodiversity and Climate Change Virtual Laboratory](#)

Module 4 Design a SDM

Welcome back to this online open course about species distribution modelling.
Now we know more about the theoretical background of species distribution models, and the different types of data that you need to built a model, it is time to design your model. What I mean with that is that you need to think about the question that you are trying to answer, and what kind of data and algorithm you need to find that answer. As I mentioned in the previous module, the saying 'garbage in, garbage out' means that if you just throw some data into a model, the result won't be very meaningful. With the enormous amount of data available online, and tools like virtual labs that make it easier to run a species distribution model, it is good to take a step back and evaluate your input to make sure your results will be reliable.

In order to design your species distribution model, you need to think carefully about the three components of your model: the species data, the environmental data, and the algorithm. When we talk about the algorithm, we mean the actual method that you use to determine the probability of occurrence based on a set of environmental data. Which algorithm is most suitable is dependent on the type of data that you have, but it also goes the other way around: based on your preferred algorithm there are different optimal methods for example to generate the pseudo-absence data.

Let's start with a few general questions that you need to answer about your data: first, what data is available about your species of interest, and is this data accurate? As we mentioned in the previous module of this course, it is important that you check whether there are any anomalies in your occurrence dataset, or whether there is any sampling bias that you need to take into account such as the geographical coverage of your data. And this of course also counts for your environmental data.

You might wonder how large your occurrence dataset needs to be, that means how many occurrence points are necessary for a good performance of the model. Of course you are dependent on the data that is available. For some common species, such as the wedge-tailed eagle, you might find datasets with tens of thousands records. But for conservation purposes you might be interested in a rare species with much less records such as the Richmond frog. The optimal number of occurrence records is related to the geographic range of your species. In general, models tend to be less accurate for species that have a broad geographic range and that are tolerant to a range of environmental conditions compared to species with smaller geographic ranges and limited environmental tolerances. So even if your species is rare and you only have a few occurrence records, if its geographical range is small, it is likely that the suitable environmental conditions for it are accurately sampled with fewer points compared to a species with a larger range. Generally, the minimum necessary number of occurrence records is about 30, and algorithms that use only presence data are less affected by small sample sizes.

Before you start your search for environmental data, you should think about which factors are likely to influence the distribution of your species of interest. Although some algorithms are able to handle a large amount of predictor variables, it is always good to remain critical about which variables you include in your model. This means you have to do a bit of research to get to know your species, and choose predictors that directly affect the distribution of your species. For example, if you know that your species is sensitive to very high or low temperatures, you have to make sure that you include temperature-related variables in your model. If you are not sure which factors influence your species, you can first run a model with lots of predictors. The outcome of the model will show the response curves for each environmental variable which you can use as a guide to select the most important predictors to run a subsequent, more refined model. In this example, the response curves for soil type and radiation show a flat line, which means that they did not influence the probability of occurrence, and you could thus choose to leave these variables out in the next model. You do have to be aware that most algorithms take into account interactions between variables and thus adding or leaving out variables can change the outcome of the model. This again highlights the importance of doing some research when you design your species distribution model.

The third aspect of a species distribution model that you have to select is the algorithm that you will use to associate species occurrences with environmental conditions. There are a lot of different algorithms available to model species distributions. In this course, we focus on four main groups: geographic, profile, statistical regression and machine learning models. This categorization is not set in stone, and can be a bit arbitrary, as many machine learning models are based on regression techniques that are also used in statistical regression models. So other sources might use different categorizations, but here I will use this categorization to give a quick overview of the algorithms that we will discuss in much more detail in module 5, 6 and 7 of this course.

Geographic models only use presence data, and do not use environmental data. They function in geographic space, and can thus be graphically visualized with latitude and longitude on the axes. These models use simple algorithms that predict that a species is present at sites within a certain shape or distance around the occurrence points. So in this example, the model draws a shape around the outermost occurrence points and predicts that a species can be present anywhere within that shape, here indicated in green. Because geographic models do not take into account the environmental conditions of occurrence sites, they are often not considered as true species distribution models. But they provide a good method to get a quick idea of the spatial extent of a species.

Profile models are the most basic true species distribution models. Like geographic models, they also use only occurrence data, but these models do use environmental data as well. Therefore they function in environmental space, and the axes of the graph represent different environmental variables that are used to predict the probability of occurrence. The best known profile model is Bioclim, which is regarded as the first species distribution model. Bioclim constructs a boundary box around the minimum and maximum values of each environmental variable, and it predicts that species can be present in all locations that fall within those boundaries. Profile models have a few limitations as they can only handle continuous environmental variables, and they do not take into account interactions between the variables, but they are very good to explore which factors influence your species if this information is not available beforehand. We will explain geographic models and Bioclim in more detail in module 5.

Statistical regression models need both presence and absence data. As we have learned in module 3, absence data can either be true absence data or be represented by ‘made up’ data, which we call pseudo-absence data. These models also use environmental data, and the algorithms use all the data available to estimate the coefficients of the environmental variables, and they construct a function that best describes the effect of those variables on species occurrence. Statistical regression models can handle both continuous and categorical predictors and also include interactions between those variables. In module 6 we will explain three popular statistical algorithms for species distribution modelling: generalized linear models, generalized additive models and multivariate adaptive regression splines.

The group of machine learning models consist of a lot of different approaches that all use environmental data. Most algorithms use both presence and absence data, except for the popular Maxent technique that uses presence data in combination with background data. We will explain this further in module 5. A variety of machine learning models are based on decision trees. In module 7, we will explain how classification trees work, and also look at more complex tree-based models: random forests and boosted regression trees. Another type of machine learning models that we will explain in detail in module 7 of this course are the Artificial Neural Networks.

Now, the main question is of course how to choose which algorithm to use in your species distribution model. There is no real straightforward answer to this question as it depends on a lot of different things. Although it is almost impossible to recommend one method over another, I will give a short overview of some limitations and assumptions of the models, that might guide you in the design of your species distribution model.

Firstly, the data that you have available or want to use might limit some of your options. If you don’t have any environmental data available, you are limited to a geographic model, which mostly gives just an indication of the range of a species. If you do have data on environmental conditions, then you can design a true species distribution model. The next step is to look at the availability of your species data. If you only have presence data, you can choose to run a simple profile model, such as Bioclim. An alternative if you only have presence data, is Maxent, which is a presence-background model that contrast the environmental conditions of presence locations with all available locations. Alternatives to presence only and presence-background models are presence-absence models with either true absence or pseudo-absence data. These can either be statistical regression models or machine learning models. Each of the algorithms have their own assumptions and limitations with regards to the input data. As I mentioned earlier, profile models are not able to include categorical predictors or interactions. They generally show poorer performance compared to presence-absence or presence-background models. Statistical models tend to be more sensitive to outliers and missing data compared to the machine learning models. But machine learning models are more sensitive to overfitting the data. As we will see in module 5, Maxent has an inbuilt process to avoid overfitting. An advantage of machine learning models, though, is that they are able to handle large datasets. However, if you don’t have many occurrence points available, Maxent or a statistical model might work better. Keep in mind that this guide is not exclusive, and in general it is advised to run multiple models and compare their outcomes.

Besides the influence of your input data on the suitability of a model, the choice for a model also depends a lot on what you want as a user. First of all, the interpretation of the output differs between the models. Maxent and Bioclim work from an environment perspective, and they test the suitability of the environment for presence of a species. Statistical and machine learning models take a species perspective, and test the probability of occurrence in locations with particular environmental conditions. Another concern is the expertise of the user. Although some tools might make it easier to design species distribution models, it is important that you understand what you are modelling. Some models might perform very well, but are more complex to understand and interpret. Additionally, some models need to be tuned by setting the configuration options to specific values depending on your datasets. In those cases, just running an algorithm with the default configuration options might not give an optimal result. Not everyone has the time or resources available to learn new techniques, and thus you have to think about what you are capable of doing. While I would like to encourage you to explore the range of opportunities that species distribution models offer, you have to keep in mind that this is a complex topic that needs some time and investment to fully comprehend. Lastly, there are also some practical things to keep in mind, such as whether the modelling tools are freely available or not. And whether you have access to the computational infrastructure that you sometimes need for running large models, and visualizing the output.

On top of all these choices, there is one last thing I would like to mention with regards to pseudo-absence data. Because this data is generated by the model, it doesn’t represent true observations in the field. This means it will likely introduce some kind of error into the model.

It is therefore important to think carefully about how to generate this data with regards to two aspects: the number of points that you generate, and the method that you use. Researchers have provided general guidelines with a recommendation of 10,000 pseudo-absence points, randomly generated in the study area for statistical models, and an equal number of pseudo-absence points as there are occurrence points, generated in locations with contrasting environmental conditions to those occurrence points for machine learning models. Again, I would like to advise you to treat this guideline with caution, and do some research on the recent developments and recommendations for the algorithms of your interest.

So maybe it is disappointing to learn that there is not one perfect algorithm for all of your research questions. But all these options give you the opportunity to design a species distribution model suitable for your species and study area. Just make sure that you think about the criteria and assumptions and justify why you choose a particular algorithm. Additionally, in a tool such as the BCCVL, you can easily run multiple algorithms and compare their output, so if you're not sure which one fits your data best, you can select more than one. If you use multiple models, you might get slightly different results, and it is always good to take all of these results into account before you draw your final conclusions about the distribution of your species of interest.

I hope that this module has given you a better overview of the different aspects of a species distribution model that you need to take into account when designing your species distribution model. In the next three modules, we will dive into the details of the particular algorithms. I hope to see you there!

Attribution

Please cite this video as follows:

Huijbers CM, Richmond SJ, Low-Choy SJ, Laffan SW, Hallgren W, Holewa H (2016) SDM Online Open Course, module 4: design a species distribution model. Biodiversity and Climate Change Virtual Laboratory, http://www.bccvl.org.au/training/. DDMMYY of access.

Acknowledgements

Lead partners: Griffith University, James Cook University

Thanks to: University of New South Wales, Macquarie University, University of Canberra

Funded by: National eResearch Tools and Resources Project (NeCTAR)

The BCCVL is supported by the National eResearch Tools and Resources Project (NeCTAR), an initiative of the Commonwealth being conducted as part of the Super Science Initiative and financed from the Education Investment Fund, Department of Education. The University of Melbourne is the lead agent for the delivery of the NeCTAR project and Griffith University is the sub-contractor.