This is a video transcript of Module 3 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 3 Data for SDMs

Welcome back to our species distribution modelling course! In the first two modules we gained a better understanding of how you can use species distribution models and some of the ecological theory underpinning these models. So now you might be interested in running one of these models yourself. Where do you start? Right, with data!

With the increasing development of information technology, the amount of data in the world has been exploding. On top of that, we have seen a large movement towards open data, which is the idea that some data should be freely available to everyone to use and republish. That means there is a lot of computing infrastructure and data available, which makes it easier to build and run models such as species distribution models. In this module, we will have a closer look at the different types of data that you need to run a species distribution model, where to get this data from, and things to be aware of and some standard good practices when dealing with data.

To run a species distribution model, you need two types of data: biological data, which are the coordinates of the places where your species of interest occurs, and environmental data, that describe the environmental conditions of those places.

Biological data

Let's start with the biological data: where does your species of interest occur or not occur. Like many explorers have done in the past, you can go out in the field and conduct surveys to discover where your species is present or absent. This is of course quite a big job, especially if you need data for very large areas, for example at the scale of a continent. A lot of people that model species distributions therefore rely on datasets that have been collected by other researchers or institutions. These data can come from museum records, from large research surveys or citizen science initiatives such as annual bird counts. There are an increasing number of resources online where such data is collated and you can visualize or download an occurrence dataset of your species. In Australia for example, there is the Atlas of Living Australia, which has 194 dataset collections with occurrence records for more than 100,000

species. At a global level, the Global Biodiversity Information Facility, GBIF, is a valuable resource with free and open access to occurrence records of more than 1.5 million species. If you are more interested in a particular taxa or group of species, there is a variety of resources such as Track, which has data about Australian fish distributions.

So I encourage you to explore the different sources of data that are available, but I would like to point out a few things that you need to be aware of. Good data practices are important. First, it is important to note that when you use data that you have not collected yourself that you need to correctly acknowledge where your data came from. The best way to do this is to check with the data provider what their terms of use are and how you can cite the data source. Secondly, when you are using data that is provided by an open online source, it is your own responsibility to check whether all the data is accurate. Although a lot of the data providers have their own process of checking the quality of the data, there is always the possibility that a dataset includes a few inaccurate records. With species distribution data, the bare minimum you can do to validate your data is to check whether there are duplicate records, or any strange occurrence points on the map, for example occurrence points in odd places such as points in the ocean for a terrestrial species. You can easily clean the data by downloading the Excel file, check for duplicates and remove them if you wish, and then sort them by longitude and latitude to check whether there are any odd records so you can remove the outliers. Lastly, you can check in which years your data was recorded and remove data from very early years if you wish to do so. Also, be aware of alternative names for your species as you might find different datasets for different names according to the local usage of a species' name. In such cases, it is most useful to use the latin name of the species as this is identical globally. Remember that the output of a model is only as valid as your input data. There is an old saying 'garbage in, garbage out', which means that if you run a model with nonsensical data, you will get a nonsensical outcome.

Another thing to keep in mind is that occurrence data can sometimes be biased towards the accessibility of sampling locations. Species are more likely to be observed near places where people go, and thus occurrence data can be lacking for remote areas. This can lead to a non-representative sample of the environmental conditions, although this is not necessarily the case. For example, an environmental variable such as temperature is not necessarily affected by human-related infrastructures and will thus be generally the same in a large area that has both accessible sites as well as more remote sites. But if your species is related to factors like soil type or land use, then a distribution model is likely to be influenced by a bias in sampling locations.

Depending on the species distribution modelling algorithm that you want to use, you might not only need data of where a species occurs, but also data for where a species does not occur, or absence data. We will look at the assumptions and criteria of the different algorithms for species distributions in the next module, but I would like to explain here a little bit more about the difference between true absence data and pseudo-absence data.

You can say that when you have repeatedly observed that a species is not present in a particular location, that it is truly absent. True absence points refer to locations where the environmental conditions are unsuitable for a species to survive. I should point out that with some species, for example migratory animals, you have to be careful with such conclusions, as a species might be absent only in particular seasons. But in general, comprehensive surveys can supply true absence data when sites have been visited one or more times and people used high quality detection methods suitable for the species. For example, if you want to record true absences of a species that is only active during the night, you should carry out the surveys at night and not draw conclusions about absences if you only conducted daytime surveys. Similar to collecting presence data, this is a time consuming job, and thus true absence data is hardly ever available for any species.

If you do not have true absence data, but you do want to use an algorithm that compares the environmental conditions of occurrence sites with those of absence sites, you have to 'make up' absence data. For example, if you are unable to get to a survey site, you may need to infer your absence data. This is called pseudo-absence data.

There are a few different methods to generate pseudo-absence data. The most simple one is to randomly generate pseudo-absence points in a predefined geographical area, anywhere except for locations where presence has been recorded. A more refined method based on this is to use the same predefined geographical area, but exclude not only the exact locations of presences, but all areas that have similar environmental conditions as those occurrence areas. You then only generate pseudo-absence points in areas that have contrasting environmental conditions to the occurrence areas. Another method is to generate pseudo-absence points in a radius around an occurrence point. You first set a minimum distance from your occurrence point. This ensures that you don't place your pseudo-absence point too close to an occurrence record, as you can assume that the environmental conditions would be too similar. You then set a maximum distance from your occurrence point to ensure that the pseudo-absence points are not in inappropriate locations which may result in overprediction. We call this method the min and max radius method. You might wonder which method is the best for your species, and your question. This depends on factors like how widespread your species is, the number of occurrence records you have, and which algorithm you want to use for your model. We will come back to these types of decisions in the next module of this course.

So, to summarize the biological data: you need occurrence data, which are reliable records of places where a species has been observed. And for some algorithms you need absence data, which can be a bit more complex if you don't have true absence data and you thus have to generate the data based on some assumptions. But despite this complexity, some of the presence-absence algorithms are known to work better than presence-only algorithms, and you can thus get more realistic outcome of the model if you have both types of data.

Environmental data

The other type of data that you need is environmental data: this is data that tells you about the environmental conditions of the places where your species is present or absent. The most common types of environmental variables that are used in species distribution modelling are described by four classes of physical conditions, called the primary environmental regimes: moisture, thermal, radiation, and mineral nutrients. The moisture regime is mostly described by measures of rainfall and evaporation, the thermal regime by temperature measures. The radiation regime refers to solar radiation or sunlight, which is usually measured by the photosynthetically active radiation, or PAR. This is the spectral range of solar radiation that photosynthesising organisms, such as plants and algae, are able to use in the process of photosynthesis. This spectral region corresponds more or less with the range of light that is visible to the human eye. The mineral nutrients regime is determined by the soil type.

Other factors such as altitude can also affect the distribution of a species, but usually these only have an indirect effect on the species, as they are affecting environmental conditions within the primary regimes. For example, altitude has an effect on temperature, and therefore indirectly affects species distributions. For species distribution models it is better to use environmental variables that have a direct effect on survival, rather than indirect factors.

For species living in the ocean instead of on land there are oceanic variables such as sea surface temperature and salinity that can be used in species distribution models.

As with species data, there are a lot of online resources available that provide environmental data. For example, WorldClim is a collection of global climate layers of current and future climate. There is also a global soil database, and again there are also smaller scale national or regional databases. It is important to first think which environmental variables are likely to influence your species and then search for the environmental datasets that suit your species.

It is good to be aware of how environmental data is generated. The data that you download are usually not the raw data collected. Raw data would be measurements such as daily rainfall or temperature. In Australia, there are 10,000 stations around the country that measure the amount of rainfall over 24 hours each morning at 9 am. There are also 1500 stations that continuously measure temperature, and report the maximum and minimum temperature over 24 hours each morning at 9 am.

The raw data is not very useful for a species distribution model as daily measurements are highly variable, and species respond to environmental conditions not so much on a daily basis but rather over longer time scales. Therefore, these raw data are processed to generate variables such as the mean annual temperature or the minimum or maximum of the warmest or coldest, the wettest or the driest month or season. Such minimum or maximum values make much more sense in species distribution modelling as the probability of a species occurring in a particular place is often influenced by a threshold of an environmental factor. For example, if a species cannot tolerate temperatures above or below a certain threshold, variables that

represent minimum or maximum values are very useful to describe the environmental conditions under which a species is able to survive.

Like species occurrence data, environmental data is only collected in particular locations where measuring stations are situated. To use this data in a model, it needs to be converted to create what is known as a 'raster surface' in which each cell has a value for a particular environmental factor, including the cells for which no measurements of environmental factors exist. Because we don't know the exact value for each cell, we use a technique called spatial interpolation that predicts values for unknown cells from a limited number of sample data points around that cell. This interpolation is based on the assumption that cells that are close together tend to have similar characteristics.

The resulting surface can be visually displayed in a two-dimensional graph, which is a contour graph or in a three-dimensional graph in which the x and y axes represent the longitude and latitude and the z-axis the value of the environmental factor measured.

So in summary, environmental data say something about the environmental conditions of the sites where your species of interest occurs. We can divide environmental data in different regimes and the data that you use in a model is usually interpolated data from raw data collected by measuring stations.

Scale

Another important aspect of both species and environmental data is scale. Spatial scale has two components: grain, which is the resolution of your data. The other component is extent, which is the total study area. So in this image, grain or resolution is the size of one individual grid cell, and refers to the sample resolution of a single observation. In other words, at what scale is occurrence data of a particular species or measurements of a particular environmental factor collected? Extent refers to the total geographic area of a study. For example, habitat type can be defined in grid cells of 1 km² or 10 km², which we refer to as fine versus coarser resolution.

It is important to think about resolution when you select an environmental dataset for your species distribution model. Ideally you want to choose the resolution of the dataset at a scale which is relevant to your species. You can imagine that the appropriate resolution of a temperature dataset that is used to model the distribution of a plant species, which always remains in the same place, is different compared to one that you need to model the distribution of a species with a daily home range of 20 or 30 km such as large birds. And if you include environmental variables from different regimes in your model, such as temperature and soil, then you will likely have different resolutions among your datasets, as climate data is often available at resolutions of 1 or 5 km², whereas soil datasets have a much finer resolution such as <100 m².

So, while there is a lot of data available for you to run a species distribution model, you have to keep in mind that you need to check the accuracy and scale of the datasets that you want to use, and choose the appropriate dataset for your species and question.

In the next module we will look at how you can combine all this information to design your species distribution model!

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 3: data for species distribution models. 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.