单元 3 SDM 数据

欢迎回到我们“物种分布模型”课程！在前两个单元中，我们更好地了解了如何使用物种分布模型和一些与模型相关的生态学理论。现在你可能想要自己运行一个模型。从哪里开始？对，数据！

随着信息技术的不断发展，世界上数据量增加迅猛。除此之外，我们已经看到了公开数据的大趋势，一些数据应该免费提供给所有人使用和重新发表。这就是说有很多计算平台和数据可以使用，这使构建和运行模型，如物种分布模型，变得更容易。在本单元中，我们将详细介绍运行物种分布模型所需要的不同类型的数据，从何处获取数据，需要注意的事项以及处理数据时一些标准操作。

要运行物种分布模型，你需要两种类型的数据：生物数据，即你感兴趣的物种的分布位点，以及描述位点环境条件的环境数据。

生物数据

我们从生物数据开始：你感兴趣的物种分布在哪里或不分布在哪里。像过去许多探险家一样，你可以去实地进行调查，来记录目标物种是否有分布。这当然是一个很费劲的工作，特别是如果你需要非常大的地区的数据，如大陆尺度。因此，很多物种分布建模的研究人员依赖于其他研究人员或机构收集的数据集。这些数据可以来自博物馆记录、大型研究调查或者公民科学行动如年度鸟类统计。这些数据的在线资源越来越多，你可以浏览或下载研究物种的分布位点数据集。例如澳大利亚的 Atlas of Living Australia 网站，有 194 个数据集记录了超过 10 万个物种的分布记录。在全球尺度上，全球生物多样性信息基金（GBIF）是一个可以免费获取超过 150 万个物种分布记录的宝贵资源。如果你对特定的类群或一组物种感兴趣，则有各种资源，如 Track 是有关于澳大利亚鱼类分布的数据。

所以我鼓励你搜集不同来源的可用数据，但我想指出需要注意的几件事情。掌握正确使用数据的方法很重要。首先，请注意，当你使用别人的数据时，请正确标注数据来源。最好的方法是与数据提供者确认数据的使用条款，以及如何引用数据源。其次，当你使用开放的在线数据时，你有责任核对数据的正确性。尽管许多数据提供者都会控制数据质量，但是数据集总是有可能包含一些不准确的记录。使用物种分布数据，你至少需要检查是否有重复记录或任何异常的物种分布地点，例如异常的分布地点，例如分布物种分布在不应该分布的地点，例如这里红色点所示的陆地物种分布在海洋中。你可以在下载的 Excel 文件中，检查并删除重复数据，然后将数据按经度和纬度排序，以检查是否有异常记录，以便删除异常值。最后，你可以检查数据记录的年份，并根据个人意愿删除早期的数据。此外，请注意物种别名问题，一个物种可能在不同地区有不同叫法，因而以不同的名称出现在不同的数据集里。。在这种情况下，使用该物种的拉丁名是最效的方法，因为这是全球通用的。请记住，要想得到好的模型预测结果 必须基于好的数据。有一个古话是输入垃圾 输出也会是垃圾，这意味着如果你用一个无意义的数据运行模型，你会得到一个无意义的结果。

要记住的另一件事是，物种分布位点有时会集中于易到达的采样区域。离人近的地区的物种更容易被观测到，而偏远地区会缺少该物种的分布记录。这会导致获得的样本不能真正代表其分布的环境条件。例如环境变量（如温度）未必受到人造设施的影响，因此不会受到到达难易程度的影响。但是，如果你研究的物种的分布与土壤类型或土地利用类型等因素有关，那么分布模型很可能受采样点偏差的影响。

根据你要使用的物种分布模型算法，你可能不仅需要物种分布位点的数据，还需要物种没有分

布的位点信息。我们将在下一个单元中 介绍物种分布模型的不同算法的假设和准则，但我想再解释一下，关于真正的分布无数据和伪分布无数据之间的区别。

当你反复观察到一个物种没有分布在一个具体位置时，它是真正的不存在。真正的分布无数据是指该地的环境条件不适合该物种生存。我应该指出，但对于一些物种 如迁徙动物，你在下结论的时候必须谨慎，因为该类物种只有在特定的季节才会无分布。但一般来说，通过适宜的调查方法，经过多次调查没有记录到目标物种，才能确保分布无数据的可靠性。例如，要获取夜行性物种的分布无数据，那你应该在夜间进行调查，而不是在日间调查的结果上得出没有该物种分布的结论。与收集分布有数据类似，这是一个非常耗时的工作，因此对任何物种而言，都难以获取真实的分布无数据。

如果你没有真实的物种分布无数据，但你确实希望比较有分布位点和无分布位点的环境条件，则你必须“构建”分布无数据。例如，如果你无法到达调查地点，则可能需要推断得到分布无数据。这被称为伪分布无数据。

有几种不同的方法来生成伪分布无数据。最简单的是在物种有分布的区域之外，设定一定的区域（图中的灰色区域），随机生成伪分布无位点。基于此的改进方法是使用相同的预先指定的区域，不仅排除有该物种存在的确切位置，还同时排除与目标物种分布区环境条件相似的地区，这里用红色表示。然后，只在与物种分布地区环境条件有明显差异的地区生成伪分布无数据点。另一种方法是在物种出现点周围一定的半径中生成伪分布无数据点。你首先设定其与分布点的最小距离。这样可以确保你的伪分布无位点不会太靠近物种有分布的位点，避免分布有/无位点的环境条件太相似。然后，你设定其与分布点之间的最大距离，以防止伪分布无位点导致的过度预测。我们称这种方法为最小最大半径法。你可能想知道哪种方法最适合你的物种和你的问题。这取决于你的物种分布的广泛性，你有的物种分布记录个数以及你要用的模型算法。我们将在本课程的下一个单元中回到这些策略上。

我们总结一下生物数据：你需要物种分布位点数据，这些数据是已观测到物种地点的可靠记录。而对于一些算法而言，你需要物种无分布数据，如果你没有真实的无分布数据，就会复杂一些，因为你必须根据一些假设生成伪分布无数据。尽管复杂，使用物种分布有/无数据的算法优于只用物种分布有数据的算法。如果你有两种类型的数据，你可以获得更准确的预测结果。

环境数据

你需要的另一种类型的数据是环境数据：这是你的物种分布或无分布位点的环境条件数据。最常见的环境变量包括四类物理环境，被称为主要环境条件：水分，热，辐射和矿质营养。水分主要用降水量和蒸发量表示，热由温度表示。辐射方式通常是指用光合有效辐射（PAR）表示的太阳辐射。这是光合作用生物（如植物和藻类）在光合作用过程中使用的太阳辐射光谱。该光谱范围与可见光基本重合。矿质营养取决于土壤类型。其他因素，如海拔等也可能影响物种的分布，但通常来说，这些因素对物种的影响是间接的，因为它们是影响环境条件的初级因素。例如，海拔影响温度，因此间接影响物种分布。对于物种分布模型，最好使用对其生存有直接影响的环境变量，而不是间接因素。

对于生活在海洋而不是陆地上的物种，海洋的一些变量，如海面温度和海水盐度可用于物种分布模型。

与物种数据一样，环境数据也有大量在线资源。例如，WorldClim 是全球现在和未来气候的集合。还有一个全球土壤数据库，还有更小尺度的国家或地区数据库。首先要考虑哪些环境变量可能影响你的物种分布，然后搜集适合你物种的环境数据集。

最好了解环境数据是如何生成的。你下载的数据通常不是收集的原始数据。原始数据是测量值如每日降雨量或温度。在澳大利亚，全国分布有 1 万个站点每天上午 9 时测量 24 小时内的降雨量。还有 1500 站连续测量温度，并在上午 9 点报告 24 小时期间的最高和最低气温。

原始数据对于物种分布模型并不是非常有用，因为日常测量是急剧变化的，物种不是对每日的环境条件进行响应，而是在更长的时间尺度上。因此，对这些原始数据进行处理以得到年均温等变量或最暖、最冷、最潮湿、最干燥的月份或季节的最大、最小值。这样的最小值或最大值在物种分布模型中更有意义，因为物种在特定地方的分布概率往往受到环境因素阈值的影响。例如，如果一个物种不能忍受高于或低于某一温度阈值，变量的最小值或最大值对描述物种能够存活的环境条件非常有用。

和物种分布数据一样，环境数据仅在测量站所在的特定位置收集。要在模型中使用此数据，需要将其转换为“栅格表面”，其中每个像元都有一个特定环境因素的值，也包括没有环境数值的像元。因为我们不知道每个像元的确切值，所以我们使用空间插值方法利用该像元周围的有限数量的样本数据点来预测该像元的值。假设相近的像元往往具有相似的特征。

所得到的表面展示为二维的等值线图，或者是三维，其中 x 轴和 y 轴表示经度和纬度，z 轴表示测量的环境因子值。

总而言之，环境数据表示你感兴趣的物种分布位点的环境条件。我们可以按照不同体系中划分环境数据，并且你在模型中使用的数据通常是由测量站收集的原始数据插值得到的。

尺度

物种和环境数据的另一个重要方面是空间尺度问题。空间尺度有两个组成部分：一是数据的分辨率。另外一个是范围：即研究区域面积。所以在这个图像中，分辨率是一个像元的大小，是指单个观测值的采样分辨率。换句话说，物种分布数据的尺度或环境因素的采集尺度。范围是指研究的地理区域的面积。例如，栖息地类型可以对应于 1 km^2 或 10 km^2 大小的像元，粗细两个不同的分辨率。

当你为物种分配模型选择环境数据集时，考虑分辨率很重要。理想情况下，以你物种的相关尺度选择数据集的分辨率。你可以想象，对于固定不动的植物物种和每日飞行范围为 20 或 30 公里的鸟类物种而言，需要不同分辨率的温度数据集用于分布预测。如果你的模型中包括不同来源的环境变量（如温度和土壤），那么你的数据集可能会有不同的分辨率，因为气候数据通常是 1 或 5 km^2 的分辨率，而土壤数据集有更精细的分辨率，如小于 100 m^2。

所以，尽管有很多数据可供你运行物种分布模型，但需要记住的是要检查所使用数据集的准确性和空间尺度，并为你的物种和要解决的问题选择适当的数据集。

在下一个单元中，我们将介绍如何结合所有这些信息来设计你的物种分布模型！