bio-101 www.bio-protocol.org/exxxx DOI:10.21769/BioProtoc.xxxx

利用种分布模型绘制微生物分布图谱

Illustrating Microbial Distribution Map Utilizing Species Distribution Modeling

李云涛 [1]，褚海燕 [1, 2, *]

[1] 中国科学院南京土壤研究所，土壤与农业可持续发展国家重点实验室，南京，江苏省，210008；

[2] 中国科学院大学，北京，100049

*通讯作者邮箱：hychu@issas.ac.cn

摘要：当以点向面进行变量预测时，常使用插值法来实现。但是，环境微生物是异质性极高的复杂有机群体，使用插值法对其多样性或群落结构的预测效果通常很差。另一方面，环境微生物极易受到环境及气候因子的影响，这一特性使得我们可以通过环境或气候因素来对微生物的地理分布进行高效而准确的预测。本文以中国华北平原麦田土壤细菌的地理分布研究为示例，基于环境因子对微生物群落的驱动作用，利用种分布模型，在 R 软件中绘制微生物的分布图谱，给出了微生物分布图谱的标准化绘制及验证流程，为环境微生物的生物地理学研究提供了新的思路。

关键词：微生物群落，种分布模型，分布图谱

仪器设备

1. 普通 Windows 系统个人电脑，内存 8 G，需求硬盘空间 (含软件)约 500 M

软件

1. R (v3.5.1)，所需依赖包：sp、raster、rgdal、dismo、ggplot2 和 ggthemes

   注：本教程是基于已经在个人电脑上安装好的相关软件和依赖包进行的。如果安装出现问题，请参考以下链接：

   https://cran.r-project.org/web/packages/sp/index.html

   https://cran.r-project.org/web/packages/raster/index.html

   https://cran.r-project.org/web/packages/rgdal/index.html

   https://cran.r-project.org/web/packages/dismo/index.html

   https://cran.r-project.org/web/packages/ggplot2/index.html

https://cran.r-project.org/web/packages/ggthemes/index.html

实验步骤

1. 数据准备

本分析中需要用到三个数据，即微生物群落数据 (如多样性指数)，土壤理化性状数据 (与采样点一一对应)， 以及待分析区域内的背景土壤多边形数据 （ 一般为 shapefile 格式）。

注：背景土壤理化性状多边形数据可能需要联系相关专业单位或课题组获得。本文中所使用的数据来源于国家土壤信息服务平台（http://www.soil.csdb.cn/）。

2. 数据导入及土壤理化性状筛选

2.1 导入观测的微生物多样性数据及土壤理化性状数据（图 1）

```
obs = read.table ("observation_data.txt", header = T)
```

	PD	OTUs	Chao1	Shannon	NMDS1	pH	OM
LX1	344.6568	6320.1	16521.72	11.08131	-0.10449	8.36	18.25027
LX2	331.2003	6137.5	15998.44	11.04205	-0.09119	8.48	21.11475
LX3	347.4073	6399.9	17148.35	11.1504	-0.08111	8.39	18.62106
LX4	341.2902	6317.2	16807.49	11.10065	-0.08311	8.61	15.85528

图 1. 微生物群落指标及土壤理化性状数据集

2.2 使用广义线性模型筛选出与微生物多样性显著相关的土壤理化性状，以 PD 值为例

```
m = glm (PD~., data = obs)
summary (m)
```

3. 读取背景土壤多边形数据

```
shpmap = readOGR (dsn = "7Province", layer = "bg_soil_data")
```

注：其中 dsn 指代包含所有背景地图数据图层文件的文件夹，layer 指代文件夹中图层文件的名字 (不含扩展名)。

如有必要，将地图投影转换为常用的 WGS84 坐标系：

```
shpmap84 = spTransform (shpmap, CRS ("+proj = longlat +ellps = WGS84"))
```

4. 提取需要的土壤背景栅格数据

4.1 创建一个空白栅格对象，其尺度等于背景地图的尺度

```
r1 = raster (extent(shpmap84))
```

4.2 设置空白栅格的分辨率，其单位与背景地图保持一致，数值可自行调整

```
res(r1) = c (1/50, 1/50)
```

4.3 将背景多边形数据的其中一个土壤理化性状提取到新创建的空白栅格对象中（图 2）

```
r1 = rasterize (shpmap84, r1, field = "pH")
plot (r1)
```

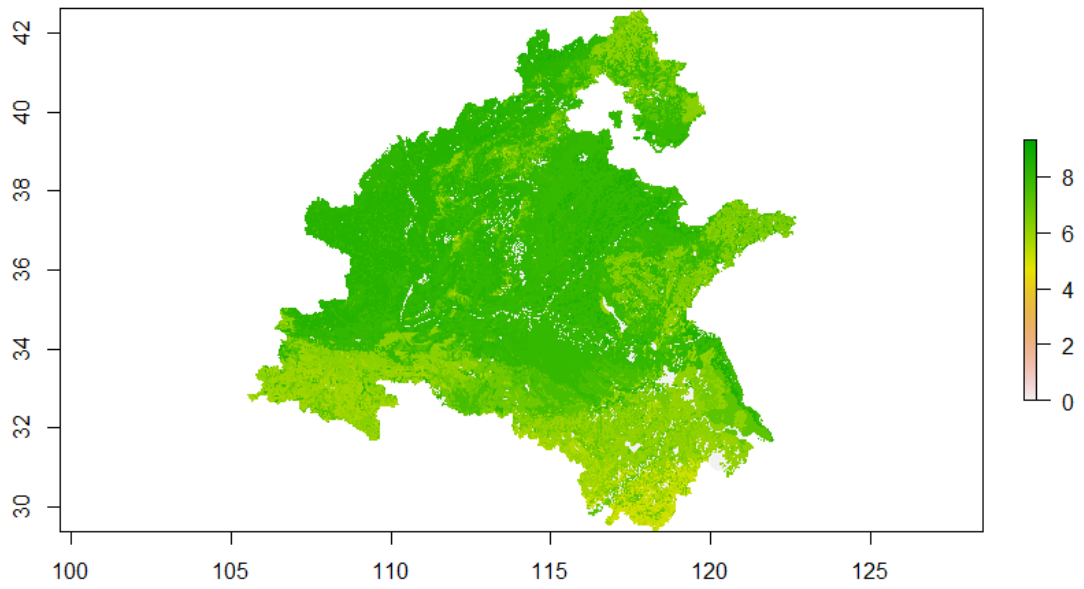

图 2. 栅格对象 r1 中的土壤背景 pH 分布

注：选择的土壤理化性状须与步骤 2 中所筛选的因子完全一致，如有多个性状，则应创建多个空白栅格 (r1, r2, r3…)并重复本步骤。

5. 种分布模型构建

5.1 合并步骤 4 中的所有栅格对象

```
predictors = stack (r1, r2...)
```

5.2 查看合并后每个土壤理化性状对应的名字

Names (predictors)

注：合并后，根据 stack ()函数中每个栅格对象输入的顺序，predictors 对象中土壤理化性状的名字将会变为layer.1、layer.2...，因此还需将观测数据中的对应土壤理化性状也更改为 layer.1、layer.2...，再次建立步骤 2 中的广义线性模型，以保持变量名的统一。

5.3 使用合并后的土壤栅格数据以及广义线性模型对象来进行微生物多样性的种分布模型构建

p = predict (predictors, m)

6. 导出微生物多样性预测值的栅格数据

将步骤 5 中的种分布模型栅格对象输出为文本格式（图 3）

p.xy = as.data.frame (p, xy = TRUE)

write.table (p.xy, file = "PD_prediction.txt", sep = "\t", col.names = TRUE)

注：输出的文本文件包含每个栅格的经纬度坐标以及预测的微生物多样性数值(x 表示经度，y 表示纬度），这里可以保存并过滤少量异常值点。

x	y	PD
117.8316	42.53966	165.0208
117.5016	42.47966	165.0208
117.7716	42.47966	165.0208

图 3. 种分布模型预测结果展示

7. 将种分布模型绘制成图

7.1 输入文本格式的栅格数据

p.xy = read.table ("PD_prediction.txt", header = T)

7.2 使用 ggplot2 包生成图像（图 4）

g = ggplot (p.xy, aes(x, y))

g + geom_tile (aes (fill = PD)) + scale_fill_gradientn (colours = c ("blue", "green", "red")) + theme_few () + xlab (label = "Latitude") + ylab (label = "Longitude")

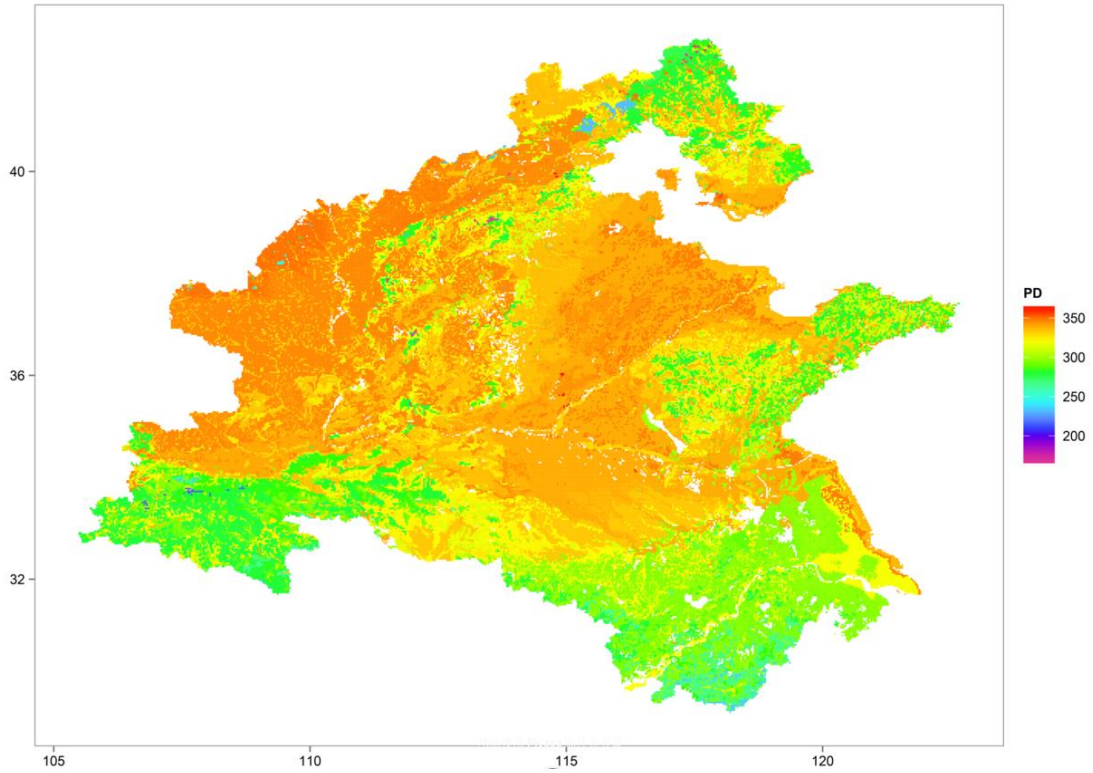

图 4. ggolot2 包绘制微生物群落多样性的分布图谱

8. 交叉验证种分布模型的效力

注：在交叉验证时，首先将观测数据 (示例中为243个)随机分为两部分，其中2/3 (162个)用于广义线性模型构建，称为模型数据集，另外1/3 (81个)用于验证模型的效力，称为验证数据集。

8.1 对观测数据随机取样并建立模型数据集及验证数据集

```
w = sample(1:243,size = 162)
mod_p = obs[which(rownames(obs) %in% w),]
write.table (mod_p, file = "modeling_part.txt ", sep = "\t")
val_p = obs[-which(rownames(obs) %in% w),]
write.table (val_p, file = "validation_part.txt", sep = "\t")
```

8.2 读取模型数据集并建立广义线性模型

```
mod = read.table ("modeling_part.txt", header = T)
m1 = glm (PD ~ pH + ..., data = mod)
```

注：这里选择的土壤理化性状须与步骤2 中保持一致。

8.3 将验证数据集中的土壤理化性状数据代入上述方程，生成对应的多样性指数预测值 *(*该步骤不再做展示*)*然后将验证数据集中的预测值和观测值一同读取

```
crsval = read.table ("PD_obs_pre.txt", header = T)
```

8.4 使用线性回归方程考察预测值与观测值的回归关系，R^2 即为交叉验证的效力，方程的斜率和截距可用于画图

```
m.val = lm (obs_PD ~ pre_PD, data = crsval)
summary (m.val)
```

8.5 使用 ggplot2 包生成交叉验证图（图 5）

```
g2=ggplot (crsval, aes (x = obs_PD,y = pre_PD))
g2 + geom_point (size = 3) + theme_few () + geom_abline (intercept = 124.73,
slope = 0.6108) + theme (text = element_text (size = 20)) + labs (x = "Observed
phylogenetic diversity", y = "Predicted phylogenetic diversity")
```

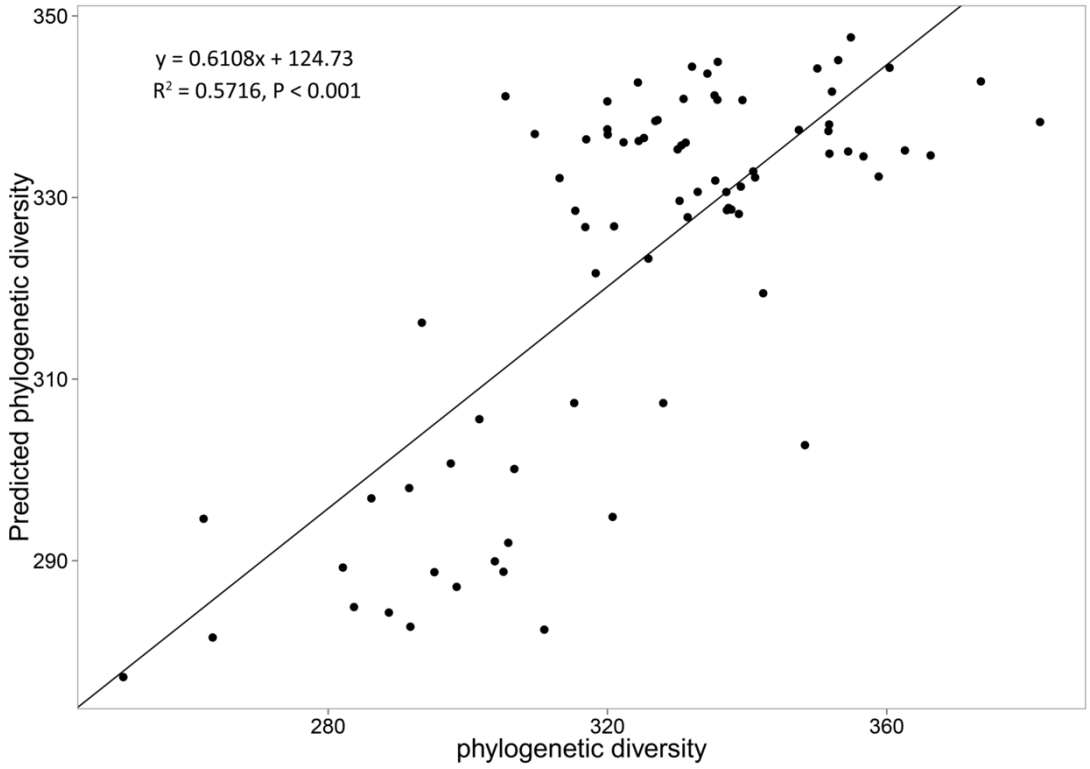

图 5. ggplot2 包绘制种分布模型效力的交叉验证结果

注意事项

种分布模型的优势和局限

优势：只需有限数量的调查样点，就能够高效、准确地预测目标范围内的各项微生物指标，包括多样性、群落结构、相对丰度等。

局限：研究范围中的生境差异性不宜过大，不同生境中的驱动因子存在差异，整体绘制微生物分布图谱时会使预测效力降低。此外，在进行种分布模型预测前，建议先分析微生物群落构建过程。如果随机性过程主导了微生物群落组装过程，那么通过该方法进行微生物地理分布预测的可靠性会降低。

参考文献

1. Hijmans, R., Phillips, S., Leathwick, J. & Elith, J.. (2014). Dismo: species distribution modeling.
2. Shi, Y., Li, Y., Yuan, M., Adams, J. M., & Chu, H.. (2019). A biogeographic map of soil bacterial communities in wheats field of the north china plain. *Soil Ecology Letters*, 1(1), 50-58.