在 R 中预测物种的潜在分布区

概念、意义和方法

张金龙

嘉道理农场暨植物园

2021-05-29

潜在分布区分析有哪些应用?

- 生物的入侵风险
- 物种潜在种群的位置
- 物种分布对全球变化的响应和优先保护区的确定
- 物种的生态位进化研究

你还知道哪些?

本讲座的学习的目标

- 掌握**物种潜在分布区分析流程**
- 知道如何获取物种分布数据
- 能绘制物种分布图，并知道分布数据里面常见的问题
- 清洁和稀疏化物种分布数据
- 用 R 驱动 maxent 运行物种潜在分布区预测，知道各参数的意义
- 根据所得结果绘制地图
- 了解**研究的可重复性**

本讲座共三部分，希望能在 2 小时内完成

1. 物种分布模型分析流程和基础（包括物种数据和环境图层数据获取，即本幻灯片）
2. 讲解代码（参见课程资料的三个文件夹）
    1. 分布数据的整理、绘图和空间稀疏化（01mapping/01 mapping.rmd）
    2. 环境图层的筛选和裁切（02 cropping and modelling/01cropping.Rmd）
    3. 用 ENMeval 驱动 maxent 选择合适的模型、转换为 01 数据并求分布区面积（02 cropping and modelling/02 modelling.Rmd）
    4. 用所得结果绘制地图 (03 visualisation/03 visualisation of sdm.Rmd)
3. 开放讨论（discussion 幻灯片）

物种潜在分布区预测难不难?

正方：难。

- 数据难于获取（物种分布数据、气候图层数据、土地利用数据、全球变化数据等）
- 概念众多，且多为一般统计课程里面接触不到的内容（如机器学习）
- 方法复杂，参数众多
- 步骤繁复，常涉及物种分布、GIS、多种建模软件包，每个步骤都有自己的问题

物种潜在分布区预测难不难?

反方：不难

- 物种分布数据越来越容易获取（GBIF）
- 免费和分辨率较高的气候图层数据已经有若干（Bioclim, CHELSA ...）
- Maxent 软件（界面化，简单易学），还有各种培训班
- Maxent 直接生成地图和统计报告

学不学?

物种潜在分布区预测在生物多样性研究中处于十分重要的地位，虽然有一定难度，但是还是要学……

难与不难，它就那里...

- 建模软件众多：BIOMOD, BIOMOD2, dismo, wallace, eSDM, iSDM, jSDN, MinBAR, SDMtune, SDMPlay...
- 数据类型多样：
  - 点数据，物种分布点记录（标本、观测、名录、笔记、植物志、动物志等）
  - 栅格数据：气候图层、地形数据、植被类型、土壤数据
  - 地理底图数据（shape 文件, kml 文件等）
- 不同来源的数据需要整合：GBIF、BIEN、VegBank、iDigBio、NSII、CVH、AVH...

难与不难，它就那里...

- 涉及的概念众多：AIC, Likelihood, AUC, ROC, sensitivity, Kappa, cross validation, n-1 fold validation, jackknife...
- 步骤繁复：往往需要几十步、上百步操作，全手动太辛苦……

学习本课程之前先要掌握什么?

- R 语言的基本操作，以便能看懂代码
- Rstudio 的基本操作，以便能编译 Rmarkdown 文档
- 用到的软件包 Maxent、dismo、tmap、ggplot2 等
- 文件的下载地址 https://github.com/helixcn/sdm_talk

工作流程

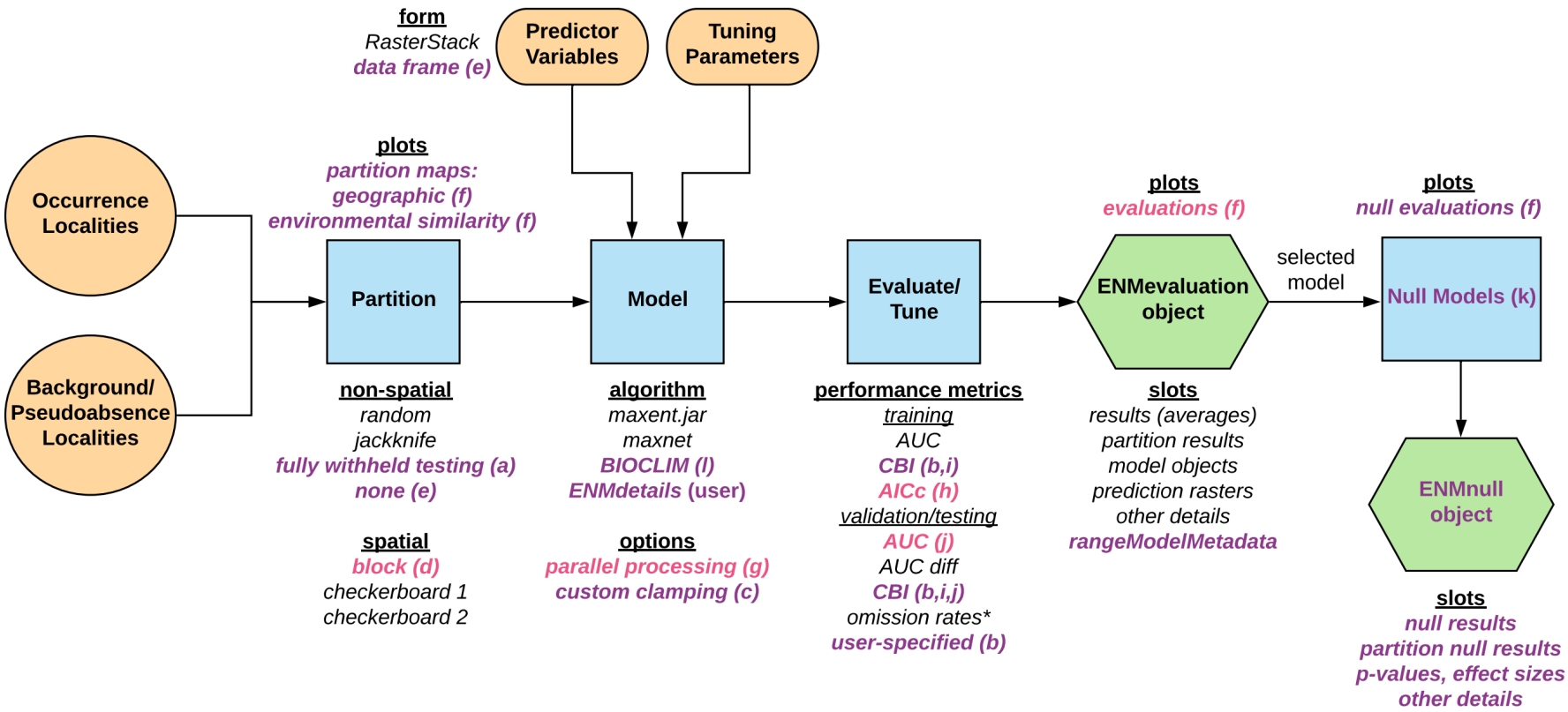


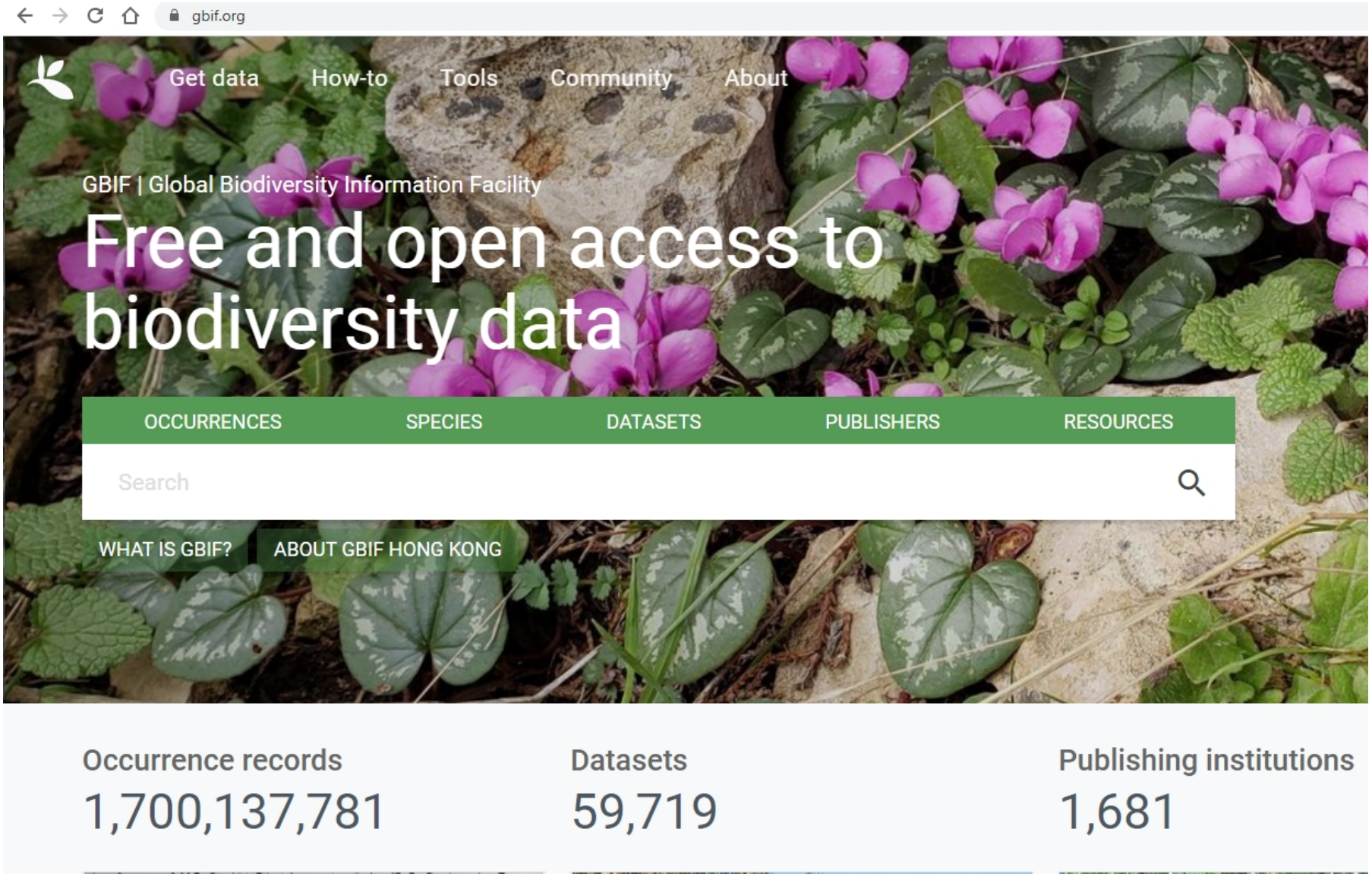
图 1: Jamie M Kass et al., 2021

数据的来源

物种分布数据可能有多种来源

- GBIF
- CVH
- NSII
- 鸟类、爬行类的数据库等
- 植物标本数据
- 植物志记录
- 古代的分布记录

数据来源 GBIF


gbif.org
Get data
How-to
Tools
Community
About
GBIF | Global Biodiversity Information Facility
Free and open access to biodiversity data
OCCURRENCES
SPECIES
DATASETS
PUBLISHERS
RESOURCES
Search
WHAT IS GBIF?
ABOUT GBIF HONG KONG
Occurrence records
1,700,137,781
Datasets
59,719
Publishing institutions
1,681

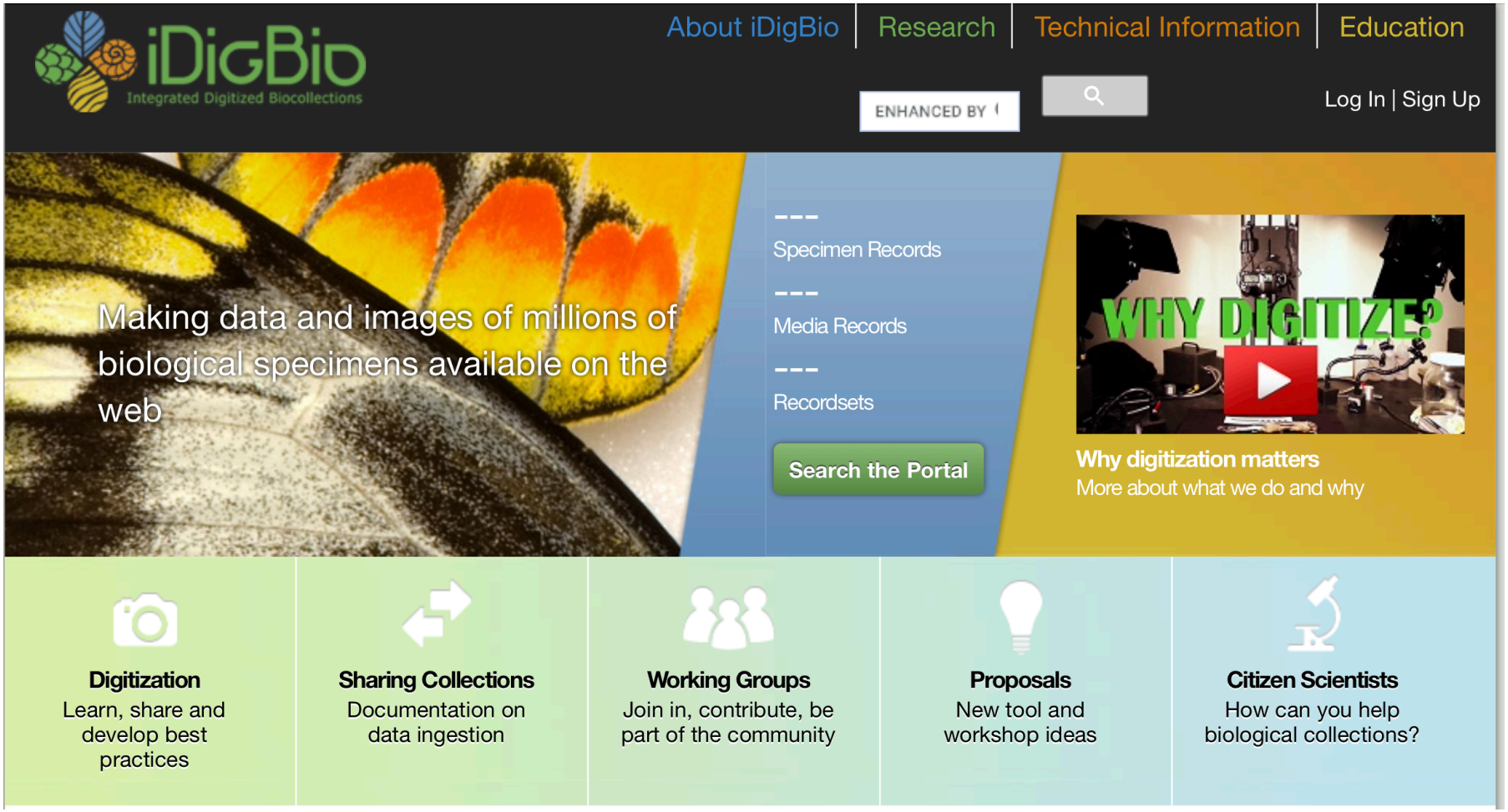
数据来源 NSII


NSII
国家标本平台
National Specimen Information Infrastructure
首页
数据
工具
网站
关于
English
登录
注册
日访问量 1435
请键入学名、中名或英文名搜索标本
16,143,146
标本记录
6,562,210
标本图片
14,917,880
彩色照片
102,658
文献
2,884
视频
植物资源
动物资源
地质资源
极地资源
最新推荐
专题资源

数据来源 CVH


中国数字植物标本馆
Chinese Virtual Herbarium
登录
首页
数据资源
分类树
新闻公告
规章制度
技术支持
实体馆
关于我们
搜索植物标本
中文名或学名
高级检索
全部标本
7,919,389条
模式标本
63,881条
彩色照片
66,113幅

数据来源 iDigBio



不同的数据库可能有不同的字段，往往不能互相兼容

- Darwin Core 格式
- NSII 格式
- CVH 格式
- herblabel

Darwin Core 是国际生物多样性数据的交换标准

TDWG Home Terms Guides Namespace policy

Darwin Core quick reference guide

This document is intended to be an easy-to-read reference of the currently recommended terms maintained as part of the Darwin Core standard. This page itself is not part of the standard. It draws on the term names and definitions from the normative part of the standard and combines them with comments and examples that are not normative, but that are meant to help people to use the terms consistently. Categories such as `Occurrence` and `Event` correspond to Darwin Core classes, which are special category terms used to group sets of terms for convenience. Comprehensive metadata for current and obsolete terms in human readable form are found in a list of terms document. Files with lists of these terms and their full history can be found in the Darwin Core repository.

To cite the standard upon which this page is built, use the following:

> Darwin Core Maintenance Group. 2020. List of Darwin Core terms. Biodiversity Information Standards (TDWG). http://rs.tdwg.org/dwc/doc/list/

To cite Darwin Core in general, use the peer-reviewed article on Darwin Core:

> Wieczorek J, Bloom D, Guralnick R, Blum S, Döring M, et al. (2012) Darwin Core: An Evolving Community-Developed Biodiversity Data Standard. PLoS ONE 7(1): e29715. https://doi.org/10.1371/journal.pone.0029715

Record-level

type modified language license rightsHolder accessRights bibliographicCitation references institutionID

collectionID datasetID institutionCode collectionCode datasetName ownerInstitutionCode basisOfRecord

informationWithheld dataGeneralizations dynamicProperties

CVH 中国数字植物标本馆的标准字段

序号	字段名（*为必填）	说明
1	provider *	子平台或新增资源所在的域名
2	institutionCode *	馆代码(规范化名称)
3	catalogNumber *	资源号/条形码
4	kingdom *	生物界或大类
5	basisOfRecord *	标本类型
6	family *	科名（生物）或资源归类编码（非生物）
7	scientificName *	学名（生物类，含命名人、年份等信息）或英文名（非生物资源）
8	commonName *	中文名
9	collector *	采集人
10	filedNumber *	采集号
11	eventDate *	采集日期（格式：YYYY-MM-DD）
12	country *	国家
13	province *	省份
14	county *	县名
15	locality	小地点
16	source *	数据页面网址
17	mediaurl	图片资源网址
18	geotime	地质时代
19	latitude	十进制 WGS84 坐标系的纬度
20	longitude	十进制 WGS84 坐标系的经度
21	Copyright	数据使用说明

为什么要有这么多格式?

- 以前，不同单位相对独立开展工作，开发出的数据库字段不同（例如：不同单位分别做标本数字化）
- 当前，主要是因为工作过程中，对数据处理的要求和侧重点不同，例如：
  - 植物标本
  - 矿物标本
  - 昆虫标本
  - 图片记录
  - 声音记录

为什么要有这么多格式?

- GBIF 为的是交换和展示物种分布数据（英文）
- CVH 为展示中文的植物标本数据
- herblabel 为了打印标签
- BG-BASE 为了管理植物园

环境图层数据 (WorldClim)

WorldClim

Below you can download the standard (19) WorldClim Bioclimatic variables for WorldClim version 2. They are the average for the years 1970-2000. Each download is a "zip" file containing 19 GeoTiff (.tif) files, one for each month of the variables.

variable	10 minutes	5 minutes	2.5 minutes	30 seconds
Bioclimatic variables	bio 10m	bio 5m	bio 2.5m	bio 30s

For reference, here is the elevation data that was used to produce WorldClim 2.1. These were derived from the SRTM elevation data.

variable	10 minutes	5 minutes	2.5 minutes	30 seconds
Elevation	elev 10m	elev 5m	elev 2.5m	elev 30s

19 个生物分布相关的重要因子

WorldClim

Bioclimatic variables

Bioclimatic variables are derived from the monthly temperature and rainfall values in order to generate more biologically meaningful variables. These are often used in species distribution modeling and related ecological modeling techniques. The bioclimatic variables represent annual trends (e.g., mean annual temperature, annual precipitation) seasonality (e.g., annual range in temperature and precipitation) and extreme or limiting environmental factors (e.g., temperature of the coldest and warmest month, and precipitation of the wet and dry quarters). A quarter is a period of three months (1/4 of the year).

They are coded as follows:

> BIO1 = Annual Mean Temperature
>
> BIO2 = Mean Diurnal Range (Mean of monthly (max temp - min temp))
>
> BIO3 = Isothermality (BIO2/BIO7) (×100)
>
> BIO4 = Temperature Seasonality (standard deviation ×100)
>
> BIO5 = Max Temperature of Warmest Month
>
> BIO6 = Min Temperature of Coldest Month
>
> BIO7 = Temperature Annual Range (BIO5-BIO6)
>
> BIO8 = Mean Temperature of Wettest Quarter

CHELSA 数据，提供了更多图层

Home Downloads Daily timeseries Monthly timeseries Bioclimate Future Paleo Climate

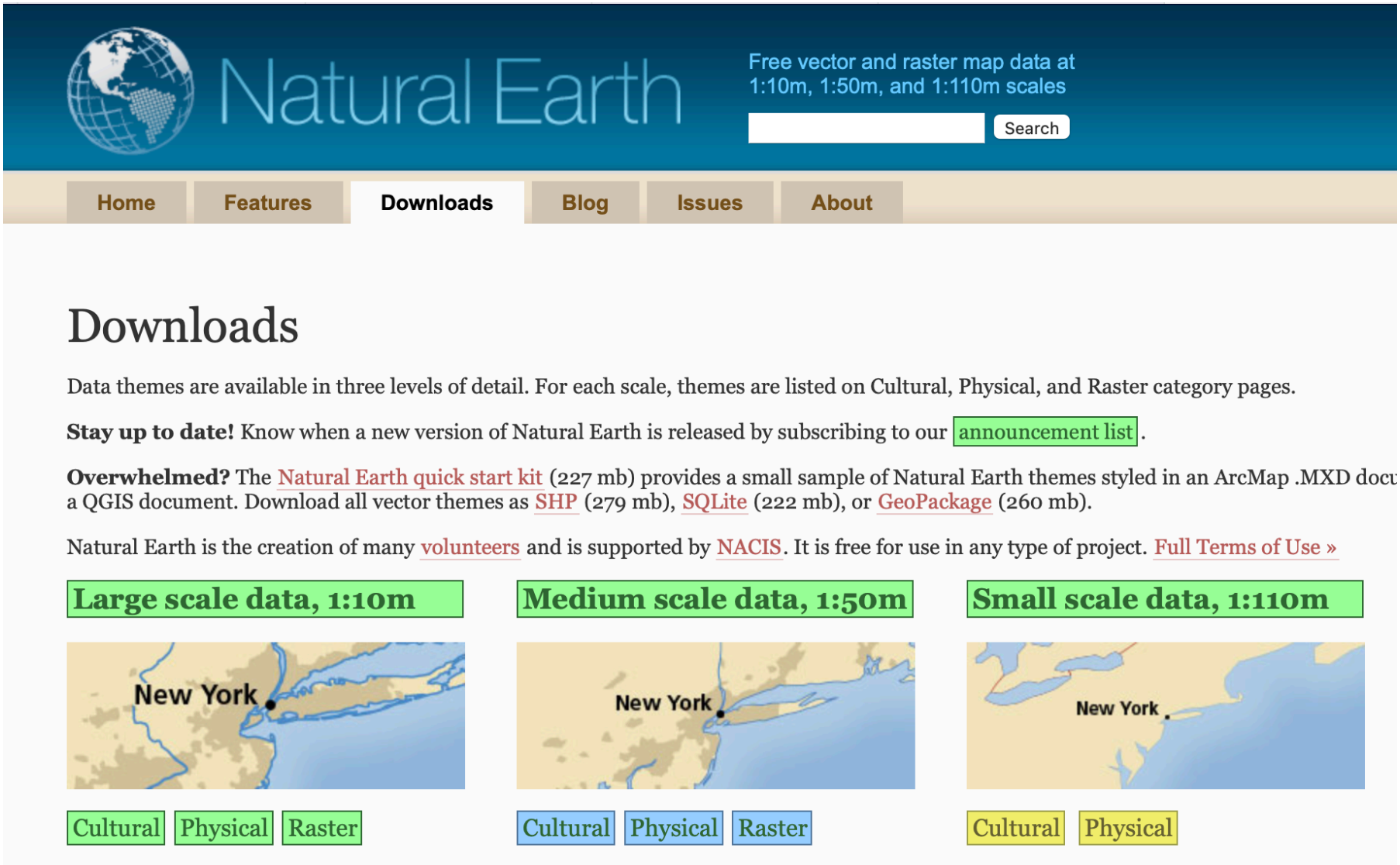
Climate diagrams

Climatologies at high resolution for the earth's land surface areas

CHELSA – Free climate data at high resolution

CHELSA (Climatologies at high resolution for the earth's land surface areas) is a very high resolution (30 arc sec, ~1km) global climate data set currently hosted by the Swiss Federal Institute for Forest, Snow and Landscape Research WSL. It is build to provide free access to high resolution climate data for research and application, and is constantly updated and refined.

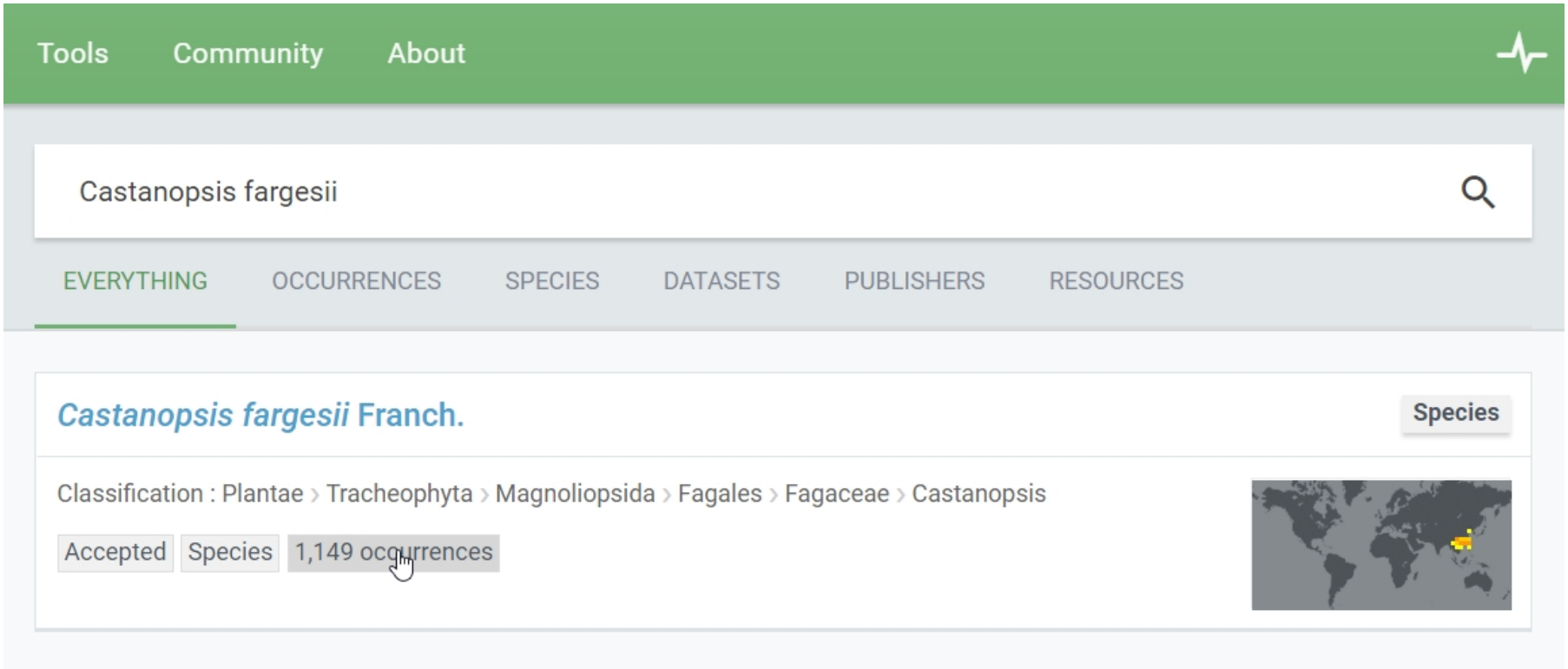
绘制地图的数据 Nature Earth 网站



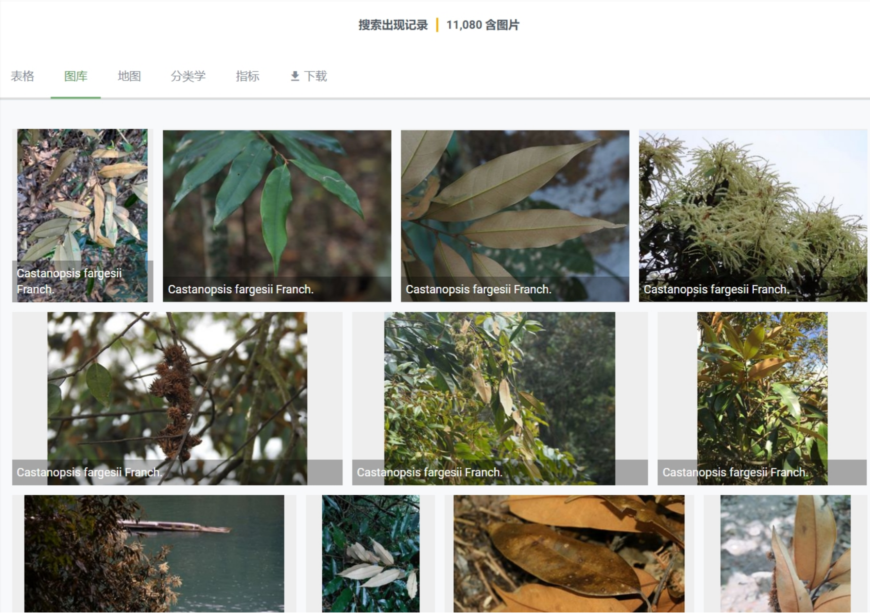
绘制地图的数据

- shape polygons 多边形
- shape polylines 线
- shape polypoints 点
- raster 栅格
- 使用中国地图的正确姿势

GBIF 网站


Tools
Community
About
Castanopsis fargesii
EVERYTHING
OCCURRENCES
SPECIES
DATASETS
PUBLISHERS
RESOURCES
Castanopsis fargesii Franch.
Species
Classification : Plantae › Tracheophyta › Magnoliopsida › Fagales › Fagaceae › Castanopsis
Accepted
Species
1,149 occurrences

练习数据：栲（南方常见树种，鉴定都准确吗?）



下载栲（Castanopsis fargesii）的分布记录

搜索出现记录 | 1,149 结果

表格 图库 地图 分类学 指标 ⬇ 下载

下载选项

	原始数据	Interpreted data	Multimedia	Coordinates	Format	Estimated data size
⬇ CSV	✗	✓	✗	✓ (if available)	Tab-delimited CSV ?	**523 KB** (77 KB 压缩下载)
⬇ 达尔文核心（DARWIN CORE ARCHIVE）	✓	✓	✓ (links)	✓ (if available)	Tab-delimited CSV ?	**1 MB** (195 KB 压缩下载)
⬇ 物种列表	✗	✓	✗	✗	Tab-delimited CSV ?	

栲（Castanopsis fargesii）分布记录简要报告

下载报告

总计: 1,149

授权条款: CC BY-NC 4.0

年份范围: 1885–2021

带年份: 92 %

带坐标: 73 %

与分类单元匹配: 100 %

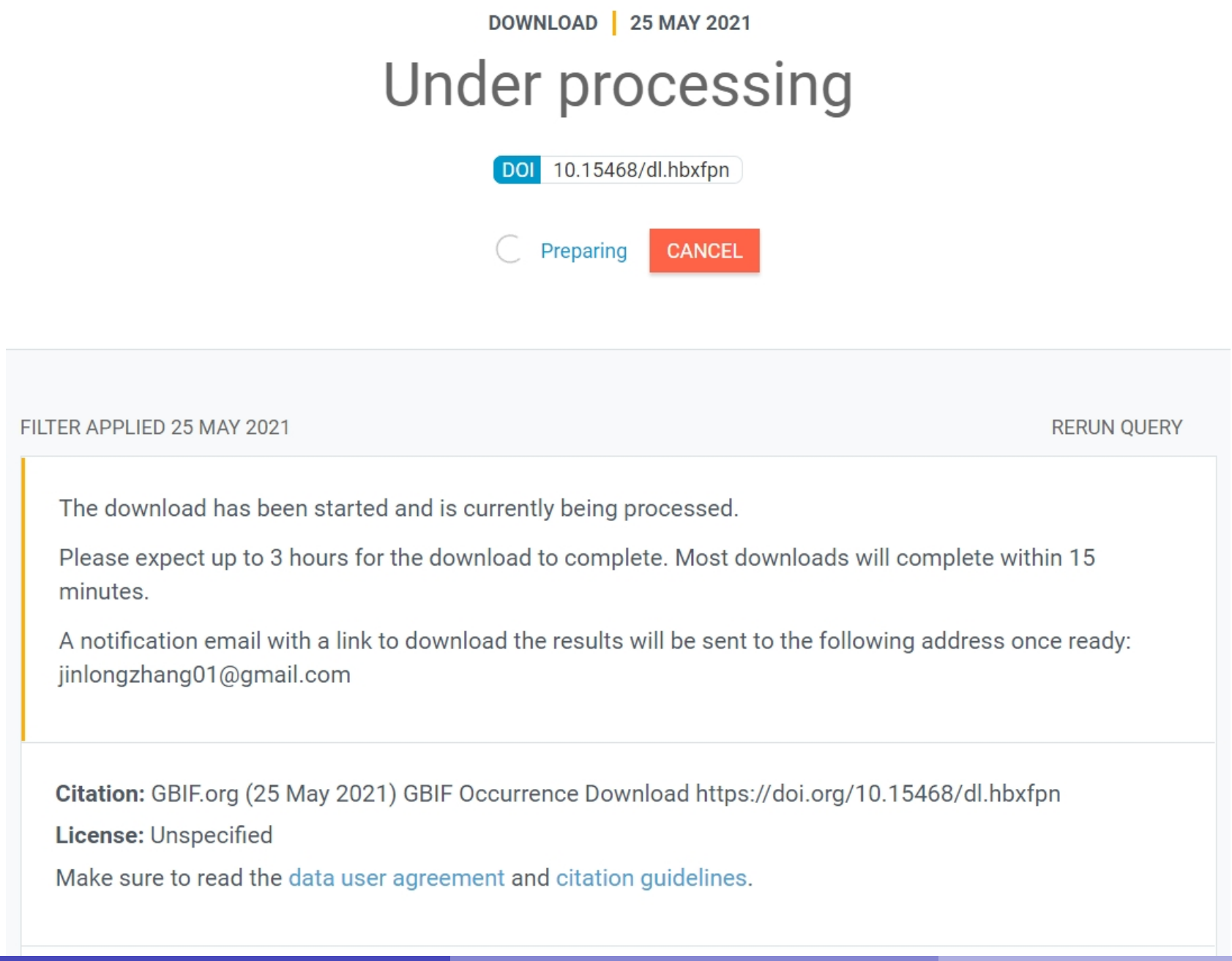
已知问题

GBIF处理的一部分内容是标记上有可疑字段的分布记录。

1,062 Institution match fuzzy
1,053 Collection match none
144 国家无效
101 国家从坐标派生
62 Institution match none
35 Collection match fuzzy
21 大地测量数据无效
18 Ambiguous collection
17 从单个计数推断出的发生状态
16 Institution collection mismatch
11 国家坐标不匹配
9 坐标无效
9 记录日期无效
7 零坐标
4 输入状态无效
2 国家不匹配
2 分类匹配较高等级
1 多媒体日期无效

使用 GBIF 数据需要合理引用

GBIF 数据准备完毕将通过邮件发送链接



GBIF 数据长啥样?

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W
1	gbifID	datasetKe	occurrenc	kingdom	phylum	class	order	family	genus	species	infraspeci	taxonRanl	scientific	verbatimS	verbatimS	countryCc	locality	stateProv	occurrenc	individual	publishing	decimalLa	decimalLc
71	2.42E+09	3942a8dc-	6905179	Plantae	Tracheoph	Magnolio	Fagales	Fagaceae	Castanops	Castanopsis fargesii		SPECIES	Castanops	Castanopsis fargesii		CN		è´µå·ž	PRESENT		a2fa5b68-	27.69	108.85
72	2.42E+09	3942a8dc-	6905181	Plantae	Tracheoph	Magnolio	Fagales	Fagaceae	Castanops	Castanopsis fargesii		SPECIES	Castanops	Castanopsis fargesii		CN		è´µå·ž	PRESENT		a2fa5b68-	27.69	108.85
73	2.42E+09	3942a8dc-	6451859	Plantae	Tracheoph	Magnolio	Fagales	Fagaceae	Castanops	Castanopsis fargesii		SPECIES	Castanops	Castanopsis fargesii		CN		æ±Ÿè¥¿	PRESENT		a2fa5b68-	26.94	114.24
74	2.42E+09	3942a8dc-	6451631	Plantae	Tracheoph	Magnolio	Fagales	Fagaceae	Castanops	Castanopsis fargesii		SPECIES	Castanops	Castanopsis fargesii		CN		ç¦å»º	PRESENT		a2fa5b68-	27.75	118.03
75	2.42E+09	3942a8dc-	6451853	Plantae	Tracheoph	Magnolio	Fagales	Fagaceae	Castanops	Castanopsis fargesii		SPECIES	Castanops	Castanopsis fargesii		CN		æ±Ÿè¥¿	PRESENT		a2fa5b68-	26.94	114.24
76	2.42E+09	3942a8dc-	6451630	Plantae	Tracheoph	Magnolio	Fagales	Fagaceae	Castanops	Castanopsis fargesii		SPECIES	Castanops	Castanopsis fargesii		CN		ç¦å»º	PRESENT		a2fa5b68-	27.75	118.03
77	2.42E+09	3942a8dc-	6451629	Plantae	Tracheoph	Magnolio	Fagales	Fagaceae	Castanops	Castanopsis fargesii		SPECIES	Castanops	Castanopsis fargesii		CN		ç¦å»º	PRESENT		a2fa5b68-	27.75	118.03
78	2.42E+09	3942a8dc-	5430614	Plantae	Tracheoph	Magnolio	Fagales	Fagaceae	Castanops	Castanopsis fargesii		SPECIES	Castanops	Castanopsis fargesii		CN		æ¹–å—	PRESENT		a2fa5b68-	26.72	110.63
79	2.42E+09	3942a8dc-	5316745	Plantae	Tracheoph	Magnolio	Fagales	Fagaceae	Castanops	Castanopsis fargesii		SPECIES	Castanops	Castanopsis fargesii		CN		ç¦å»º	PRESENT		a2fa5b68-	27.75	118.03
80	2.42E+09	3942a8dc-	5384706	Plantae	Tracheoph	Magnolio	Fagales	Fagaceae	Castanops	Castanopsis fargesii		SPECIES	Castanops	Castanopsis fargesii		CN		æ¹–å—	PRESENT		a2fa5b68-	26.72	110.63
81	2.42E+09	3942a8dc-	6451663	Plantae	Tracheoph	Magnolio	Fagales	Fagaceae	Castanops	Castanopsis fargesii		SPECIES	Castanops	Castanopsis fargesii		CN		è´µå·ž	PRESENT		a2fa5b68-	26	108.45
82	2.42E+09	3942a8dc-	6451668	Plantae	Tracheoph	Magnolio	Fagales	Fagaceae	Castanops	Castanopsis fargesii		SPECIES	Castanops	Castanopsis fargesii		CN		è´µå·ž	PRESENT		a2fa5b68-	26	108.45
83	2.42E+09	3942a8dc-	6905178	Plantae	Tracheoph	Magnolio	Fagales	Fagaceae	Castanops	Castanopsis fargesii		SPECIES	Castanops	Castanopsis fargesii		CN		è´µå·ž	PRESENT		a2fa5b68-	26	108.45
84	2.42E+09	3942a8dc-	6451651	Plantae	Tracheoph	Magnolio	Fagales	Fagaceae	Castanops	Castanopsis fargesii		SPECIES	Castanops	Castanopsis fargesii		CN		è´µå·ž	PRESENT		a2fa5b68-	26	108.45
85	2.42E+09	3942a8dc-	6451578	Plantae	Tracheoph	Magnolio	Fagales	Fagaceae	Castanops	Castanopsis fargesii		SPECIES	Castanops	Castanopsis fargesii		CN		å¹¿ä¸œ	PRESENT		a2fa5b68-	24.35	116.16
86	2.42E+09	3942a8dc-	6451667	Plantae	Tracheoph	Magnolio	Fagales	Fagaceae	Castanops	Castanopsis fargesii		SPECIES	Castanops	Castanopsis fargesii		CN		è´µå·ž	PRESENT		a2fa5b68-	26	108.45
87	2.42E+09	3942a8dc-	6451666	Plantae	Tracheoph	Magnolio	Fagales	Fagaceae	Castanops	Castanopsis fargesii		SPECIES	Castanops	Castanopsis fargesii		CN		è´µå·ž	PRESENT		a2fa5b68-	26	108.45
88	2.42E+09	3942a8dc-	6451762	Plantae	Tracheoph	Magnolio	Fagales	Fagaceae	Castanops	Castanopsis fargesii		SPECIES	Castanops	Castanopsis fargesii		CN		æ¹–åŒ—	PRESENT		a2fa5b68-	29.49	109.4
89	2.42E+09	3942a8dc-	6451650	Plantae	Tracheoph	Magnolio	Fagales	Fagaceae	Castanops	Castanopsis fargesii		SPECIES	Castanops	Castanopsis fargesii		CN		è´µå·ž	PRESENT		a2fa5b68-	25.25	106.17
90	2.42E+09	3942a8dc-	6451873	Plantae	Tracheoph	Magnolio	Fagales	Fagaceae	Castanops	Castanopsis fargesii		SPECIES	Castanops	Castanopsis fargesii		CN		æµ™æ±Ÿ	PRESENT		a2fa5b68-	27.97	119.63
91	2.42E+09	3942a8dc-	5316892	Plantae	Tracheoph	Magnolio	Fagales	Fagaceae	Castanops	Castanopsis fargesii		SPECIES	Castanops	Castanopsis fargesii		CN		æµ™æ±Ÿ	PRESENT		a2fa5b68-	27.97	119.63
92	2.42E+09	3942a8dc-	6451774	Plantae	Tracheoph	Magnolio	Fagales	Fagaceae	Castanops	Castanopsis fargesii		SPECIES	Castanops	Castanopsis fargesii		CN		æ¹–å—	PRESENT		a2fa5b68-	26.43	110.85
93	2.42E+09	3942a8dc-	6451773	Plantae	Tracheoph	Magnolio	Fagales	Fagaceae	Castanops	Castanopsis fargesii		SPECIES	Castanops	Castanopsis fargesii		CN		æ¹–å—	PRESENT		a2fa5b68-	26.43	110.85
94	2.42E+09	3942a8dc-	6451787	Plantae	Tracheoph	Magnolio	Fagales	Fagaceae	Castanops	Castanopsis fargesii		SPECIES	Castanops	Castanopsis fargesii		CN		æ¹–å—	PRESENT		a2fa5b68-	26.43	110.85
95	2.42E+09	3942a8dc-	6211901	Plantae	Tracheoph	Magnolio	Fagales	Fagaceae	Castanops	Castanopsis fargesii		SPECIES	Castanops	Castanopsis fargesii		CN		æ¹–å—	PRESENT		a2fa5b68-	26.43	110.85
96	2.42E+09	3942a8dc-	6451589	Plantae	Tracheoph	Magnolio	Fagales	Fagaceae	Castanops	Castanopsis fargesii		SPECIES	Castanops	Castanopsis fargesii		CN		å¹¿ä¸œ	PRESENT		a2fa5b68-	24.07	114.17
97	2.42E+09	3942a8dc-	5316742	Plantae	Tracheoph	Magnolio	Fagales	Fagaceae	Castanops	Castanopsis fargesii		SPECIES	Castanops	Castanopsis fargesii		CN		æ¹–å—	PRESENT		a2fa5b68-	29.42	111.13
98	2.42E+09	3942a8dc-	5316649	Plantae	Tracheoph	Magnolio	Fagales	Fagaceae	Castanops	Castanopsis fargesii		SPECIES	Castanops	Castanopsis fargesii		CN		æ¹–å—	PRESENT		a2fa5b68-	29.42	111.13
99	2.42E+09	3942a8dc-	5316648	Plantae	Tracheoph	Magnolio	Fagales	Fagaceae	Castanops	Castanopsis fargesii		SPECIES	Castanops	Castanopsis fargesii		CN		æ¹–å—	PRESENT		a2fa5b68-	29.42	111.13
00	2.42E+09	3942a8dc-	5316600	Plantae	Tracheoph	Magnolio	Fagales	Fagaceae	Castanops	Castanopsis fargesii		SPECIES	Castanops	Castanopsis fargesii		CN		æ¹–å—	PRESENT		a2fa5b68-	29.42	111.13
01	2.42E+09	3942a8dc-	6451596	Plantae	Tracheoph	Magnolio	Fagales	Fagaceae	Castanops	Castanopsis fargesii		SPECIES	Castanops	Castanopsis fargesii		CN		å¹¿ä¸œ	PRESENT		a2fa5b68-	23.28	111.9

用 spocc 包也可以下载物种分布数据

spocc (SPecies OCCurrence)

R-check failing | test-sp-sf failing | codecov 61% | CRAN NOTE | downloads 1792/month | CRAN 1.2.0

Docs: https://docs.ropensci.org/spocc/

At rOpenSci, we have been writing R packages to interact with many sources of species occurrence data, including GBIF, Vertnet, BISON, iNaturalist, and eBird. Other databases are out there as well, which we can pull in. `spocc` is an R package to query and collect species occurrence data from many sources. The goal is to to create a seamless search experience across data sources, as well as creating unified outputs across data sources.

`spocc` currently interfaces with nine major biodiversity repositories

1. Global Biodiversity Information Facility (GBIF) (via `rgbif`) GBIF is a government funded open data repository with several partner organizations with the express goal of providing access to data on Earth's biodiversity. The data are made available by a network of member nodes, coordinating information from various participant organizations and government agencies.
2. iNaturalist iNaturalist provides access to crowd sourced citizen science data on species observations.
3. VertNet (via `rvertnet`) Similar to `rgbif` and `rbison` (see below), VertNet provides access to more than 80 million vertebrate records spanning a large number of institutions and museums primarly covering four major disciplines (mammology, herpetology, ornithology, and icthyology).
4. Biodiversity Information Serving Our Nation (via `rbison`) Built by the US Geological Survey's core science analytic team, BISON is a portal that provides access to species occurrence data from several participating institutions.

绘制分布图

以便查看数据可能出现的问题：

- 是否有分布记录出现在不该出现的地方？（漂浮在海上）
- 坐标是否明显错误？
- 是否某些地方特别集中？

数据可能存在的问题以及清洁

1. 学名拼写、异名
2. 经纬度格式的更正
3. 鉴定是否正确?
4. 坐标系不匹配
5. 地名与坐标是否匹配?
6. 分布记录有多完整?
7. 分布记录在空间上是否聚集?

数据清洁 data cleaning

ELSEVIER

Global Ecology and Conservation

Volume 21, March 2020, e00852

Original Research Article

BDcleaner: A workflow for cleaning taxonomic and geographic errors in occurrence data archived in biodiversity databases

Jing Jin [a], Jun Yang [a, b]

Show more

Share Cite

https://doi.org/10.1016/j.gecco.2019.e00852

Get rights and content

有多少没有地理坐标的数据?

Home | Web Application | Collaborative Georeferencing | Developer Resources | Education

GEOLocate

A Platform for Georeferencing Natural History Collections Data

For Users:

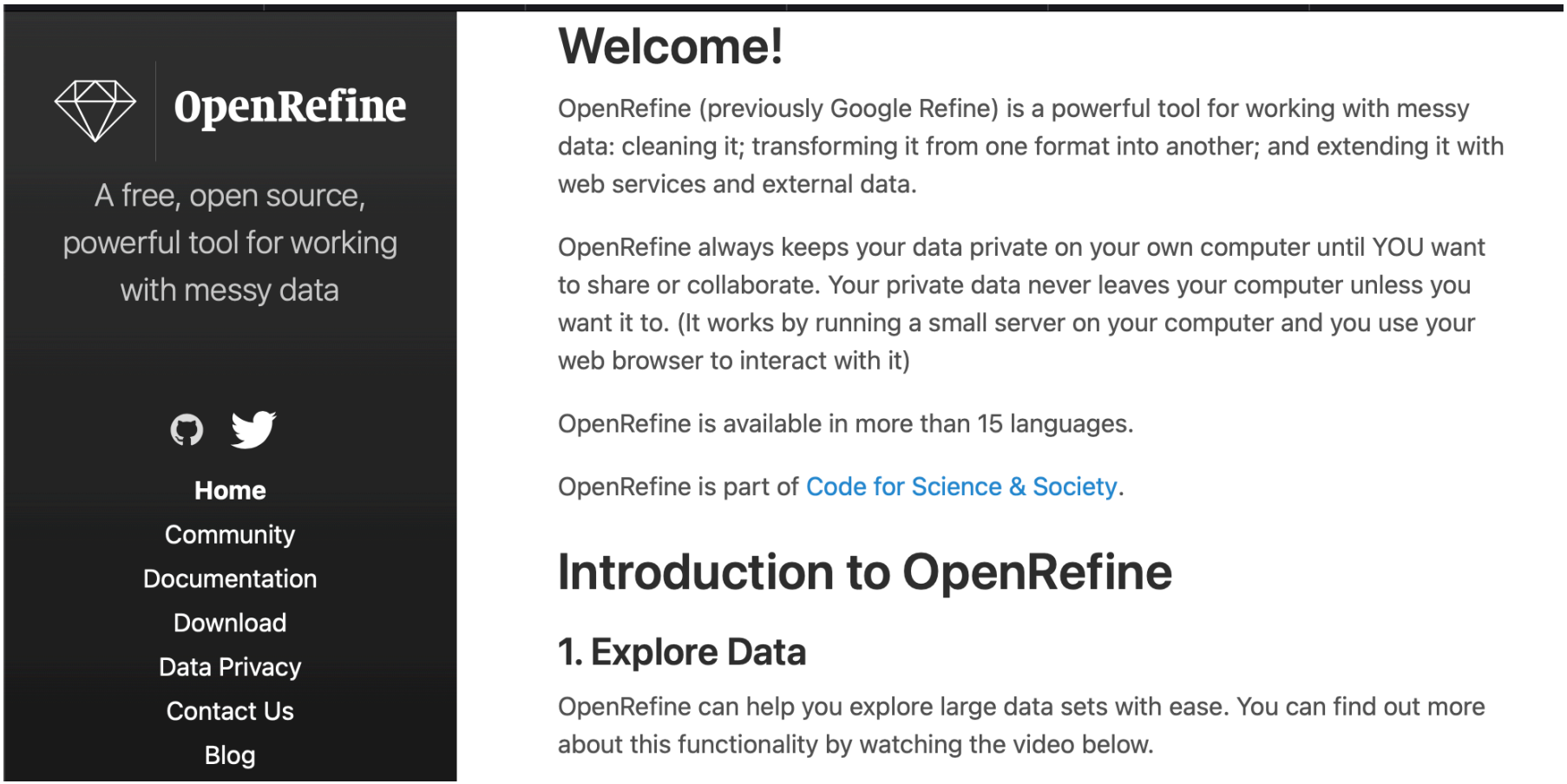
- Overview
- GEOLocate Web Clients
- Collaborative Georeferencing
- Education & Outreach

Brief overview (video) of the GEOLocate Project.

For Developers:

- Web Services
- Embeddable Web Client

有多少数据不统一的？数据的标准化



需要标准化的数据举例


OpenRefine KFBG herbarium
Facet / Filter
Undo / Redo 271 / 271
Refresh
Reset All
Remove All
change invert reset
STATE_PROVINCE
25 choices Sort by: name coun
Cluster
Lantau Island 25
Lautau Island 7
New Tereitories 1
New Teritories 4
New Terriotries 3
New Territeories 1
New Territories 9010
New territories 17
New Territories 57
New Territories, Sai Kung 1

空间上过于聚集的数据需要筛选吗?

ECOGRAPHY

A JOURNAL OF SPACE AND TIME IN ECOLOGY

Software note | Free Access

spThin: an R package for spatial thinning of species occurrence records for use in ecological niche models

Matthew E. Aiello-Lammens, Robert A. Boria, Aleksandar Radosavljevic, Bruno Vilela, Robert P. Anderson

First published: 06 February 2015 | https://doi.org/10.1111/ecog.01132 | Citations: 378

数据需要清理吗?

Received: 14 February 2020 | Revised: 1 October 2020 | Accepted: 25 October 2020

DOI: 10.1111/csp2.311

CONTRIBUTED PAPER

Conservation Science and Practice
A journal of the Society for Conservation Biology
Open Access
WILEY

To clean or not to clean: Cleaning open-source data improves extinction risk assessments for threatened plant species

Connor T. Panter[1,2] | Rosemary L. Clegg[2] | Justin Moat[2] |
Steven P. Bachman[2] | Bente B. Klitgård[2] | Rachel L. White[1]

标本采够了么?

Research Paper | Published: 26 October 2020

Taxonomic bias in occurrence information of angiosperm species in China

Wenjing Yang, Dandan Liu, Qinghui You, Bin Chen, Minfei Jian, Qiwu Hu, Mingyang Cong & Keping Ma

Science China Life Sciences **64**, 584–592 (2021) | Cite this article

129 Accesses | **1** Altmetric | Metrics

环境变量的筛选

- Pearson's r，环境图层两两之间相关系数 >0.80 的，要去掉其中一种
- VIF 膨胀因子 >10 的环境因子不要
- 不筛选

用什么分辨率的环境图层？分辨率越高越好吗?

Article | Open Access | Published: 08 May 2018

Species distribution model transferability and model grain size – finer may not always be better

Syed Amir Manzoor ✉, Geoffrey Griffiths & Martin Lukac

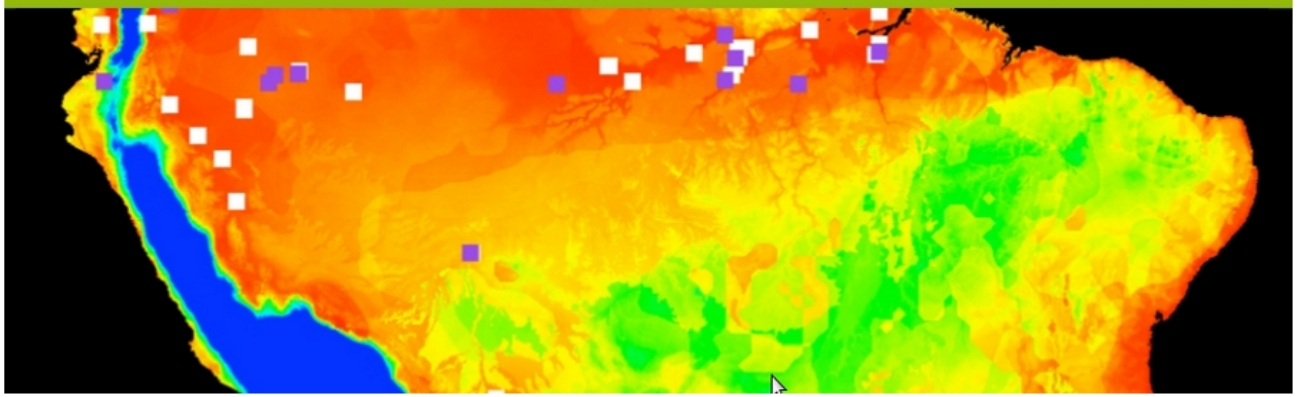
Scientific Reports **8**, Article number: 7168 (2018) | Cite this article

下载 Maxent

AMERICAN MUSEUM OF NATURAL HISTORY

Plan Your Visit | Exhibitions | Learn & Teach | Explore | Our Research | Calendar

Maxent software for modeling species niches and distributions



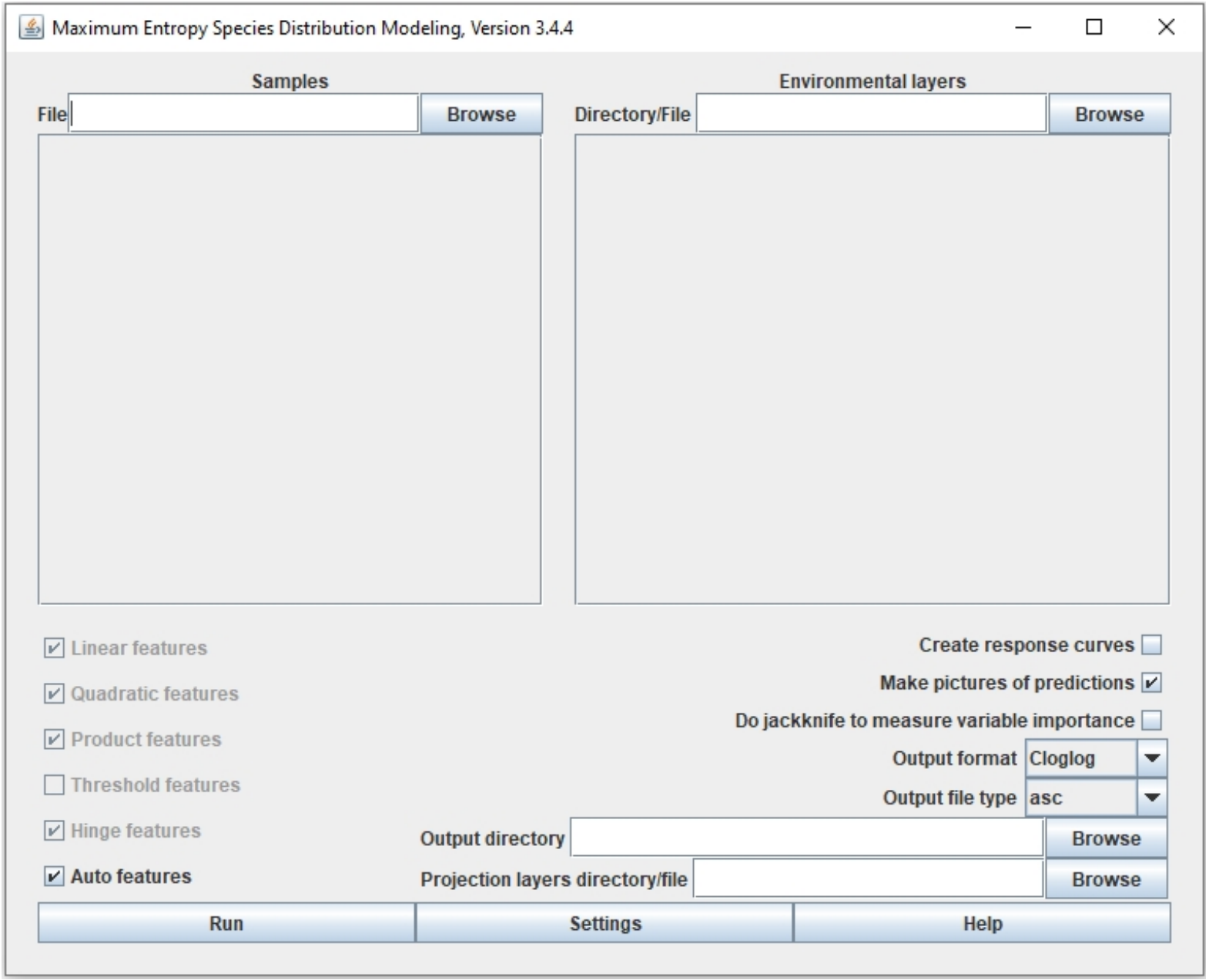
Maxent is now open source!

Use this site to download Maxent software for modeling species niches and distributions by applying a machine-learning technique called maximum entropy modeling. From a set of environmental (e.g., climatic) grids and georeferenced occurrence localities, the model expresses a probability distribution where each grid cell has a predicted suitability of conditions for the species. Under particular assumptions about the input data and biological sampling efforts that led to occurrence records, the output can be interpreted as predicted probability of presence (cloglog transform), or as predicted local abundance (raw exponential output).

运行 Maxent 的两种方法

- 通过界面运行
- 通过命令行运行
- dismo、ENMeval 等程序包都通过命令行方式调用 Maxent

Maxent 的界面


Maximum Entropy Species Distribution Modeling, Version 3.4.4
Samples
File
Browse
Environmental layers
Directory/File
Browse
Linear features
Quadratic features
Product features
Threshold features
Hinge features
Auto features
Create response curves
Make pictures of predictions
Do jackknife to measure variable importance
Output format
Cloglog
Output file type
asc
Output directory
Browse
Projection layers directory/file
Browse
Run
Settings
Help

Maxent 的命令行

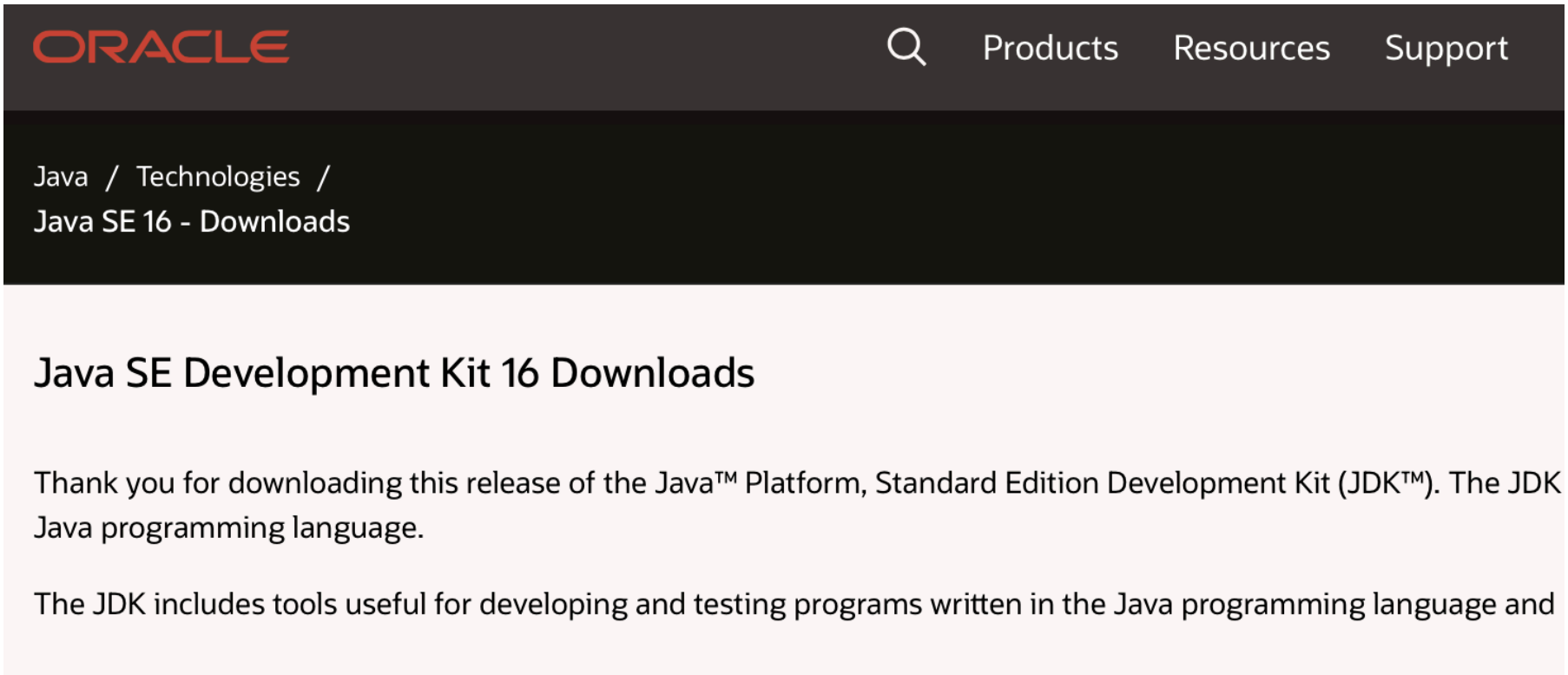
Maxent 命令行举例

```
java -mx512m -jar maxent.jar environmentallayers=layers
togglelayertype=ecoreg samplesfile=samples\bradypus.csv
outputdirectory=outputs redoifexists autorun
```

用 R 驱动 Maxent

- 安装 Java （JDK 或者 OpenJDK）
- 一般用 dismo 包，现在 ENMeval 包也可以直接驱动 Maxent

JDK


ORACLE
Products
Resources
Support
Java / Technologies /
Java SE 16 - Downloads
Java SE Development Kit 16 Downloads
Thank you for downloading this release of the Java™ Platform, Standard Edition Development Kit (JDK™). The JDK
Java programming language.
The JDK includes tools useful for developing and testing programs written in the Java programming language and

OpenJDK

Workshop
OpenJDK FAQ
Installing
Contributing
Sponsoring
Developers' Guide
Vulnerabilities

Mailing lists
IRC · Wiki

Bylaws · Census
Legal

JEP Process

search

Source code
Mercurial
GitHub

Groups
(overview)
2D Graphics
Adoption
AWT
Build
Compatibility & Specification Review
Compiler
Conformance
Core Libraries
Governing Board
HotSpot
IDE Tooling & Support
Internationalization
JMX
Members
Networking
Porters
Quality
Security
Serviceability
Sound
Swing
Vulnerability
Web

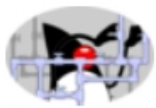
Projects
(overview)

OpenJDK

What is this? The place to collaborate on an open-source implementation of the Java Platform, Standard Edition, and related projects. (Learn more.)

Download and install the open-source JDK for most popular Linux distributions. Oracle's free, GPL-licensed, production-ready OpenJDK JDK 16 binaries are at **jdk.java.net/16**; Oracle's commercially-licensed JDK 15 binaries for Linux, macOS, and Windows, based on the same code, are here.

Learn how to use the JDK to write applications for a wide range of environments.



Hack on the JDK itself, right here in the OpenJDK Community: Browse the code on the web, clone a Mercurial repository to make a local copy, and contribute a patch to fix a bug, enhance an existing component, or define a new feature.

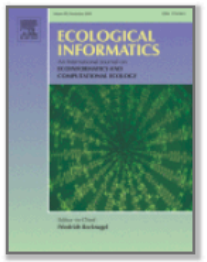
其他常见的物种分布模型

- Bioclim
- Domain
- Mahalanobis distance
- Classical regression models
- Generalized Linear Models
- Generalized Additive Models
- Boosted Regression Trees
- Random Forest
- Support Vector Machines

Assembled methods: Biomod

- assembled methods 多种方法 BIOMOD/BIOMOD2

Ecological Informatics
Volume 60, November 2020, 101150



A comparison between Ensemble and MaxEnt species distribution modelling approaches for conservation: A case study with Egyptian medicinal plants

Emad Kaky [a, b, c], Victoria Nolan [a], Abdulaziz Alatawi [a, d], Francis Gilbert [a]

Show more

Share Cite

哪种方法最优?

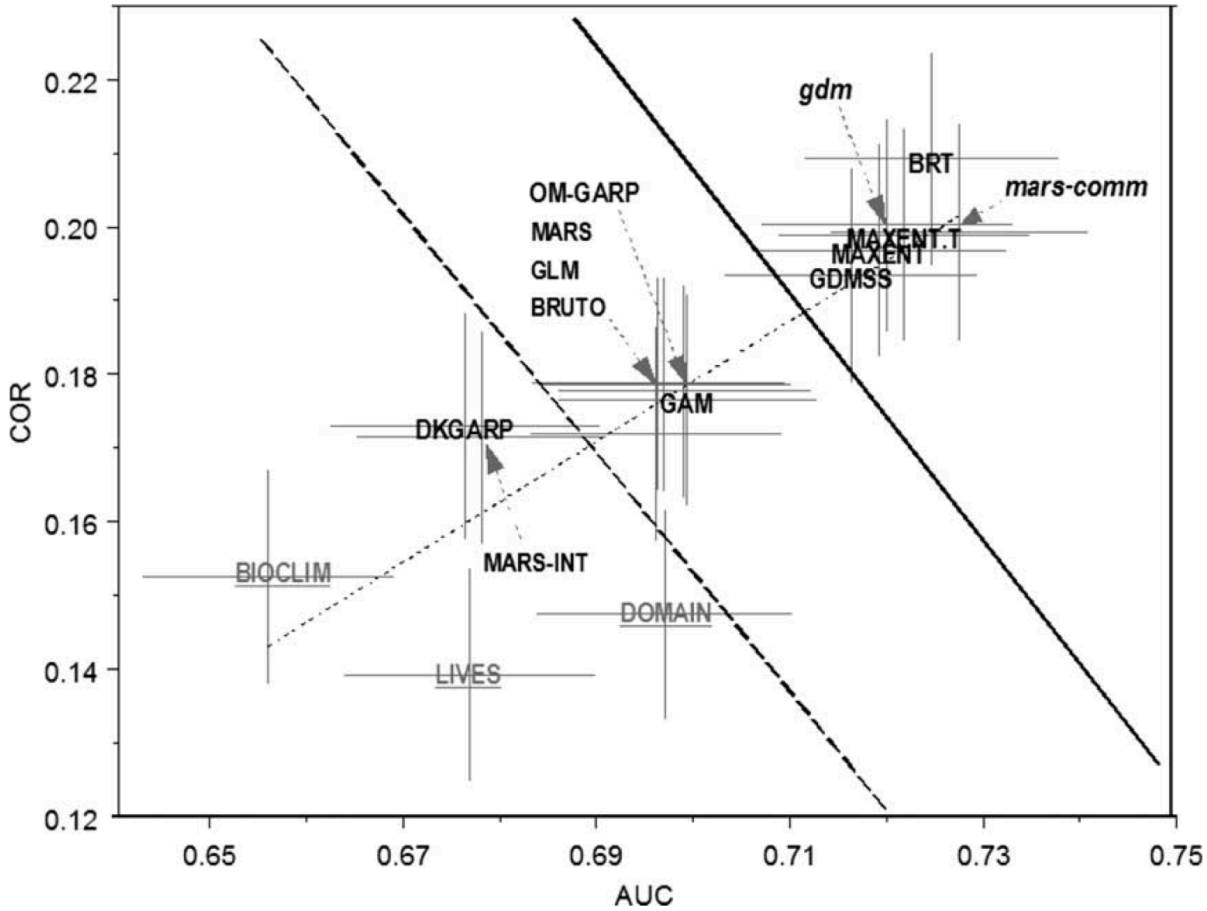


Fig. 3. Mean AUC vs mean correlation (COR) for modelling methods, summarised across all species. The grey bars are standard errors estimated in the GLMM (see Appendix), reflecting variation for an average species in an average region. The labels are broad classifications of the methods: grey underlined = only use presence data, black capitals = use presence and background samples, black lower case italics = community methods.

Jane Elith et al. (2006)

现在一般认为 BRT（boosted regression tree）和 Maxent 的效果最为“稳健”

什么意思?

- 对小数据不太敏感
- 对极端值不太敏感
- 对预测用图层的数量和分辨率不太敏感
- 对物种数据的不完整，不太敏感

在上述情况下，均能够得到较为“稳定”的结果。但这个稳定是用什么判断的呢?

模型的评价

将数据随机分成两份，

- 一份用来训练模型 (training data)
- 一份用来检验训练出来的模型 (testing data)

常用的指标

https://rspatial.org/raster/sdm/5_sdm_models.html#model-evaluation

评价模型

- ROC (receiver operating characteristic curve)
- AUC (Area under the ROC curve)

dismo 程序包 evaluate 函数的例子：我试着理解每个参数的含义，不过后来放弃了 ……

```
e <- evaluate(p=p, a=a) #

## class           : ModelEvaluation
## n presences     : 50
## n absences      : 50
## AUC             : 0.6624
## cor             : 0.3010577
## max TPR+TNR at : 0.4274477

threshold(e) # 物种出现的临界值
kappa,  spec_sens,  no_omission, prevalence, equal_sens_spec,
```

用 AUC 判断模型的准确性靠谱吗?

Received: 3 February 2016 | Revised: 2 October 2017 | Accepted: 9 October 2017

DOI: 10.1111/geb.12684

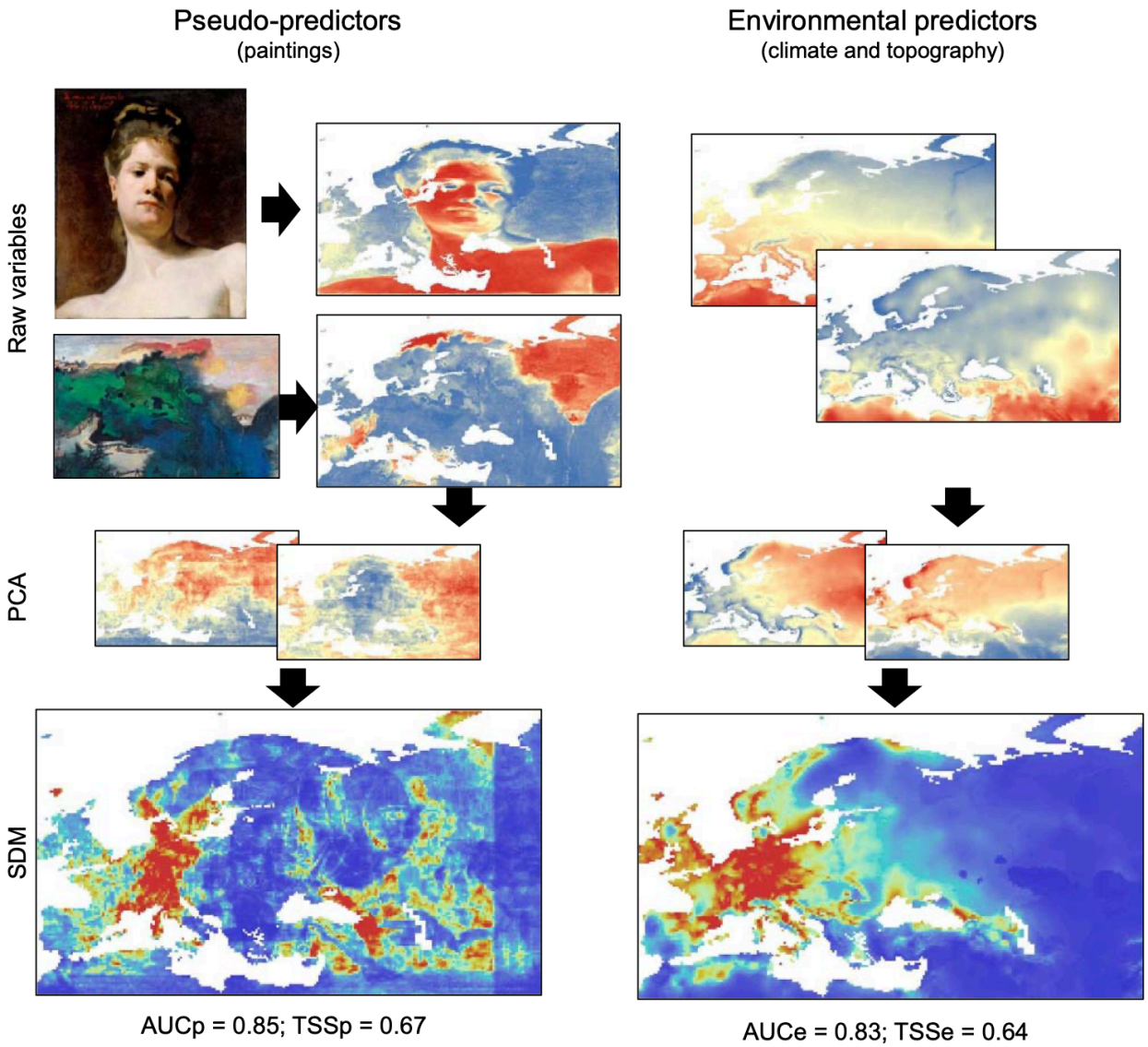
MACROECOLOGICAL METHODS

WILEY Global Ecology and Biogeography A Journal of Macroecology

Paintings predict the distribution of species, or the challenge of selecting environmental predictors and evaluation statistics

Yoan Fourcade[1,2] | Aurélien G. Besnard[1,3] | Jean Secondi[1,4,5]

由绘画转换成的图层预测结果所得 AUC 更高……



将 percentage 转换为 present-absence

0.5 作为物种是否出现的临界值是否合理?

研究告诉我们，应该选择 `equal_sens_spec`

- Liu, C., White, M., & Newell, G. (2013). Selecting thresholds for the prediction of species occurrence with presence-only data.*Journal of biogeography*,*40*(4), 778-789.

绘制地图

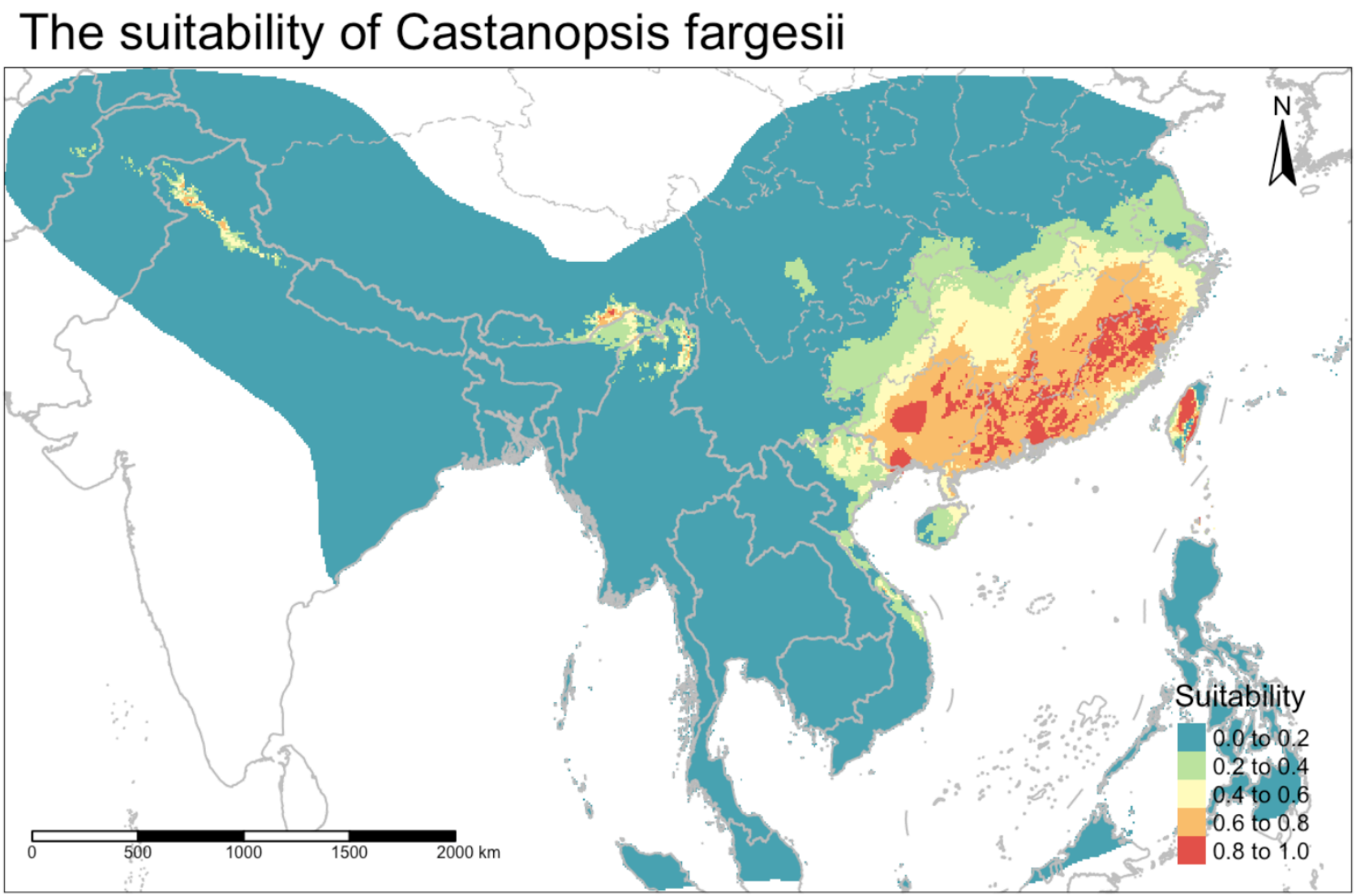
以下软件均可

- ArcGIS
- QGIS
- R
- Python

地图的几个基本要素

- 点、线、面
- 颜色、透明度、粗细、字体大小
- 指北针
- 比例尺

下面这个地图合格吗?


The suitability of Castanopsis fargesii
N
Suitability
0.0 to 0.2
0.2 to 0.4
0.4 to 0.6
0.6 to 0.8
0.8 to 1.0
0
500
1000
1500
2000 km

绘制地图的程序包 tmap

类似 ggplot2，也是基于 grid 程序包，图层可以叠加

两种常用的软件包组合：

- sf + tmap
- sf + ggplot2

研究的可重复性

1. 建立工作流程的概念，并尽量将各种操作整合到流程中。
2. 所有操作，尽量用代码完成，保证每个更改都有详细的记录
3. 代码保存好，做好注释，按照通用的规则命名和缩进
4. ....

物种分布区模型的可重复性

ECOGRAPHY

A JOURNAL OF SPACE AND TIME IN ECOLOGY

Editor's Choice and Review and synthesis | Open Access

A standard protocol for reporting species distribution models

Damaris Zurell, Janet Franklin, Christian König, Phil J. Bouchet, Carsten F. Dormann, Jane Elith, Guillermo Fandos, Xiao Feng, Gurutzeta Guillera-Arroita, Antoine Guisan, José J. Lahoz-Monfort, Pedro J. Leitão, Daniel S. Park, A. Townsend Peterson, Giovanni Rapacciuolo, Dirk R. Schmatz, Boris Schröder, Josep M. Serra-Diaz, Wilfried Thuiller, Katherine L. Yates, Niklaus E. Zimmermann, Cory Merow

... See fewer authors

First published: 01 June 2020 | https://doi.org/10.1111/ecog.04960 | Citations: 35

物种分布区模型的可重复性

Global Ecology and Biogeography
A Journal of Macroecology

MACROECOLOGICAL METHODS

Species' range model metadata standards: RMMS

Cory Merow ✉, Brian S. Maitner, Hannah L. Owens, Jamie M. Kass, Brian J. Enquist, Walter Jetz, Rob Guralnick

First published: 28 August 2019 | https://doi.org/10.1111/geb.12993 | Citations: 6

物种分布区模型的可重复性

Perspective | Open Access | Published: 23 September 2019

A checklist for maximizing reproducibility of ecological niche models

Xiao Feng ✉, Daniel S. Park, Cassondra Walker, A. Townsend Peterson, Cory Merow & Monica Papeş

Nature Ecology & Evolution **3**, 1382–1395 (2019) | Cite this article

以下第二部分（代码）

以下第二部分（代码）