生态学笔记

# 第一章 绪论

## 1. 生态学主要采用的是“假说-验证”范式，面临的困境是什么？

（1）与物理、化学等学科不同，生态学问题的答案不是简单的“是”或“不是”，通过H0显著性检验来验证生态学假说并不严谨。例如，检验群落中竞争的作用，不是通过一个简单实验就可证伪的，因为群落中同时存在竞争、捕食和寄生、干扰等多因素作用。在生态学上构建单一主导因子的H0是很难的。

（3）H0假说是小概率事件，小概率事件在一次实验中不可能发生。但生态过程不断演化，无法预见原假说是小概率事件，即。因此，生态学的所有试验验证逻辑上应倾向于“证实”假说H0成立。

（2）假说-验证范式的另一个困境是操控实验。操控实验包括正交实验设计法与析因法，无论哪种实验，都遵循重复、随机化、区组化的原则。由于研究对象所处的环境时开放的，其中影响因子及相互关系难以被发现和确定。另外，在大尺度环境下，要采取严谨和一致操作才可能排除无关干扰。可是许多专家认为，无论如何精心设计和操控，都“不可能包括所有的多样性和空间”。因此，野外实验存在着难以重复的困难，不能作为假说的判决性实验。

## 2. 何为“数据科学”范式，与“假设-验证”范式主要区别是什么？

何为“数据科学”范式：数据科学指的是利用超级计算能力，直接分析海量数据发现相关关系，获得新知识。科学发现的逻辑见图1.6（课本P9）。

与“假设-验证”范式主要区别：

第一，科学发现的逻辑起点不同。假说-验证范式主张科学发现始于科学问题，而数据科学颠覆了原来的科学发现模式，从数据出发，利用数据挖掘方法发现数据中蕴含的规律性，形成了“科学始于数据”的新模式。

第二，寻找相关性，而不是因果律。无论是经验科学范式，还是假说-验证范式，都认为科学研究的目的就是寻找现象之间的因果关系。与其他研究范式不同，数据科学范式认为分析变量之间的相关性比探寻因果律更重要，基于相关分析的预测是大数据研究的核心。

一方面，数据规律本身是从过去所积累的数据中挖掘出来的，完全可解释过去的现象或问题，另一方面，大数据包含了海量的各种现实数据，通过机器学习过去的经验来推测未来。因此，与因果律相比，基于大数据相关性的预测更准确，而且不易受偏见的影响。

第三，数据科学采用归纳方法，而不是演绎逻辑。数据科学范式沿袭了经验科学的归纳逻辑，不同的是数据科学采用的是全数据模式，即“样本=整体”的完全归纳法，客服了小样本不完全归纳法的局限性，利于发现异常值。

## 3. 什么是数据挖掘？与传统数据分析有什么区别？

数据挖掘就是从数据中发现有意义的见解和知识，并将这种发现表示为模型。因此，数据挖掘是数据科学采用的具体归纳手段和工具。

数据挖掘与数据分析不同，数据分析强调对数据的概况总结，而数据挖掘强调的是探索隐藏信息，通过机器发现的知识规则。例如，在观鸟站，有些投食点很少有鸟来取食，通过对投食点取食的鸟种类数据分析，不来取食的80%为大型鸟类，结论是不来取食的为大型鸟类，但通过编写算法自行挖掘现象背后的深层次原因，结果是投食点周围植被与筑巢栖息点有关，但很少有鸟来取食的投食点周围植被状况比较差，栖息的鸟类很少。

与传统的统计分析相比，数据挖掘在很多方法上与其是同源的。例如，数据挖掘采用的朴素贝叶斯分类就是统计理论的发展和延伸。再如，常用于数据挖掘的主成分分析和回归分析也属于统计学范畴。但数据挖掘与传统数据分析存在本质上的区别，传统数据分析需要对数据分布和变量间的关系做假设，确定用什么样的概念函数来描述变量间的关系，然后建立参数模型，并依据统计推断，确定模型的合理性，而数据挖掘并不需要对数据做任何假设，而是侧重利用机器学习方法，自动寻找变量间的关系。因此，对于海量数据，数据挖掘具有更加强大，更灵活、更高效的特点。关于数据挖掘与传统统计分析的区别间表1.2（课本P10）。

# 第二章 R语言与数据挖掘

## 1. 为何R语言被用于数据挖掘主流工具之一？它具有哪些主要特点？何为R包，即R包有哪些内容？

为何R语言被用于数据挖掘主流工具之一：

R语言是一种编程语言，它具有强大的计算和绘图能力，其语法简单，容易掌握，另外，R语言可整合其他专业软件，如QGIS，是进行生态学数据挖掘较为理想的工具。

首先R语言是开源的，无需支付任何费用。另外，R语言从一个文本编辑器发展到交互式的Rstudio和Jupyter Notebooks，这些用户友好界面极大地方便了数据操作，特别适于非计算机专业人员使用。

由于R语言丰富的内置函数和程序包，以及较高的灵活性，利用R语言可以完成数据处理任务，包括统计分析、机器学习、各种绘图以及非线性建模等，几乎涵盖了数据处理的方方面面。另外，R语言可处理多种类型数据，如数据流、Web、图形、空间数据等，被描述为一种被广泛用于数据挖掘的编程语言。

它具有哪些主要特点：

1. R语言的代码很简单，有广泛的社区支持，可帮助R语言使用者学习和掌握。
2. 如果数据量较小，可根据需要选择安装R程序包，能轻松地完成数据处理。
3. R语言具有强的整合能力，可整合地理信息系统软件，如QGIS，在QGIS中调用R也可在R中处理QGIS数据。另外，在R会话中，可交互式地使用Python，其中，reticulate包中所有函数可同时被R和Python使用，实现R和Python对象之间直接对话。

何为R包，它有哪些主要内容？

R主要依赖大量现成的程序包来实现它的功能，目前有10000多个函数包，涵盖了数据处理和分析方方面面，主要内容如下：

1.数据操作

(1)dplyr。可以对数据集做 subset、summarize、rearrange、join 等处理。

(2)tidyr。利用gather和spread 函数将数据集转化成格式更工整的数据集

(3)stringr。对字符串类型的数据进行正则表达式处理的工具。

(4)lubridate。处理日期和时间类型数据的工具。

(5)httr。处理 http 链接的工具集合。

2.可视化

(Wggplot2。功能强大的绘图工具包

(2)ggvis。一个可以做基于 Web的交互可视化工具包

(3)rgl。在 R中做 3D 交互可视化包。

(4)htmlwidgets。一个在R中快速建立基于JavaScript 内核的交互可视化工具包。

(5)googleVis。利用 Google Chart 工具在 R中做数据可视化。

3.数据建模

(1)统计分析与模型。包括car(方差分析)、multcomp(多重比较分析)、mgcv(广义相加模型)、lme4/nlme(线性/非线性混合效应模型)。

(2)机器学习建模。包括rendomForest(随机森林模型)、caret(模型工具包)。

4.编程

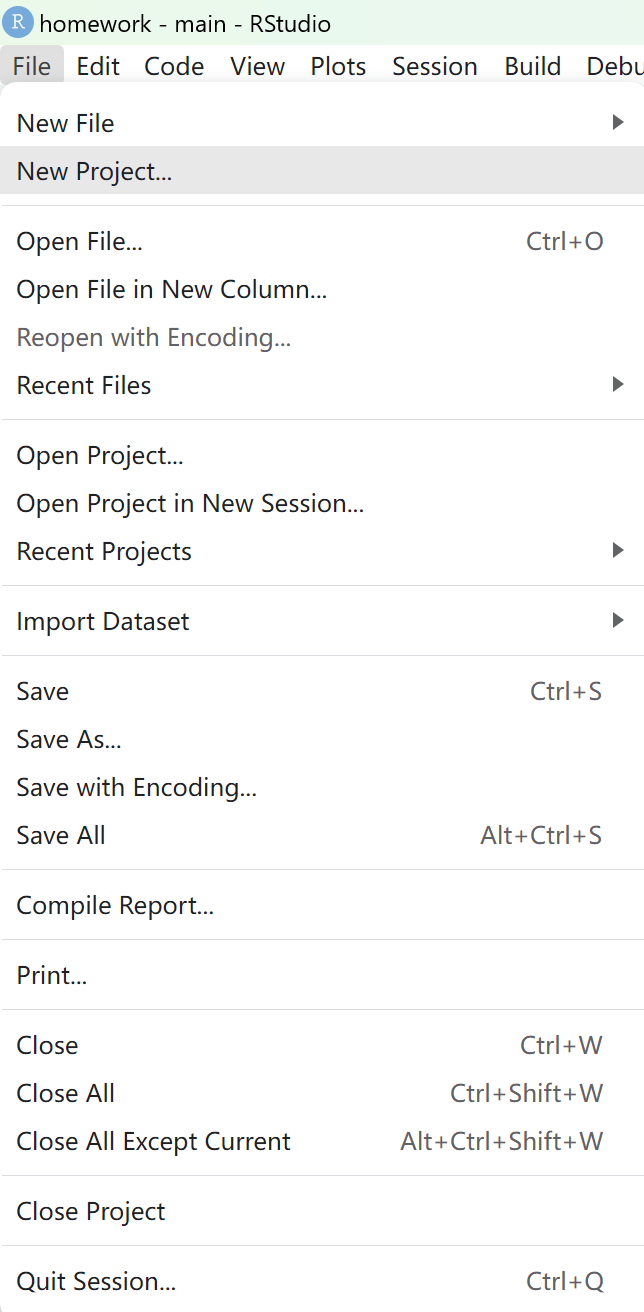
(1)Shiny。用于R交互可视化。

(2)R Makdown。处理数据分析报告的工具。

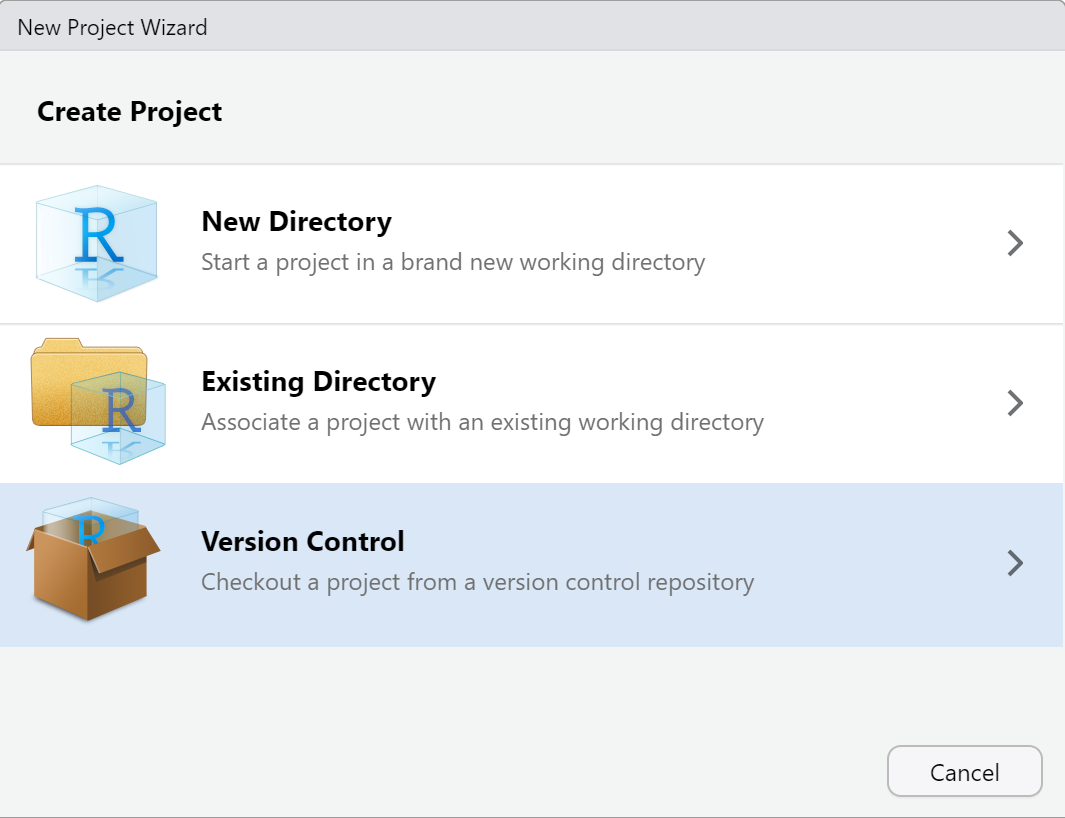
(3)jupyter notebook。提供编程环境

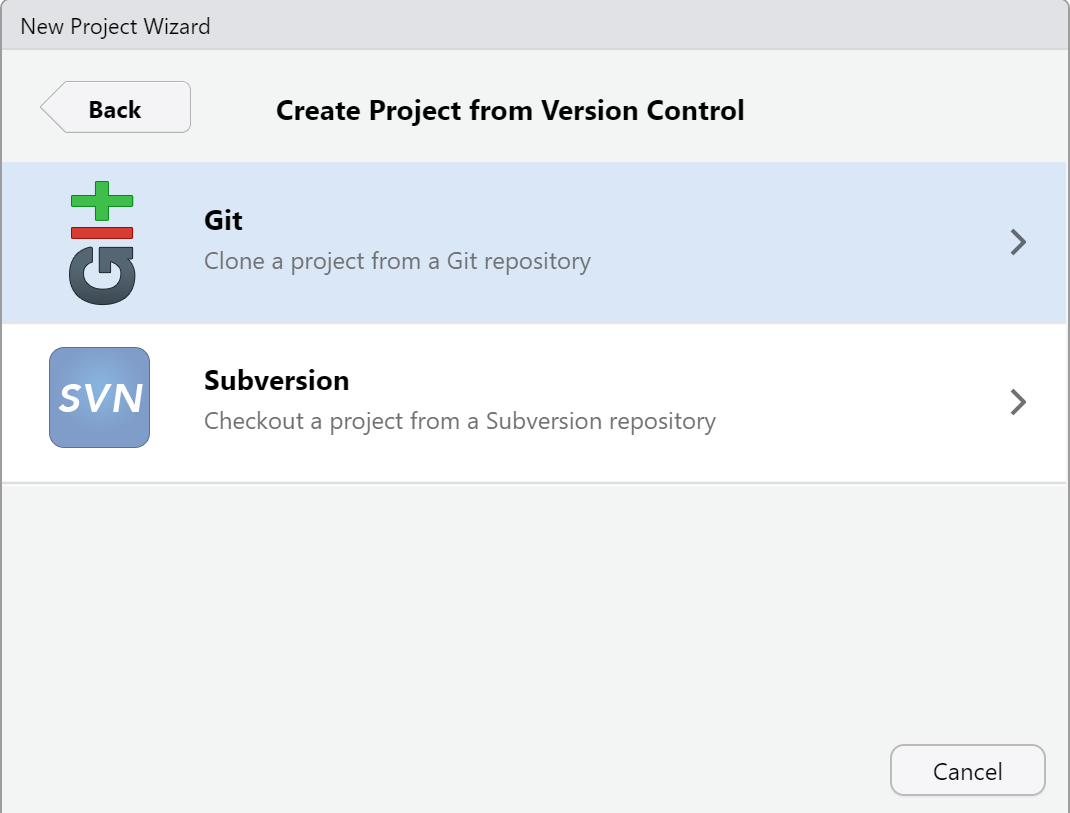
## 2. 通过git/VPN版本控制，用github托管project和代码的步骤有哪些？

File→New project



Version Control→Git





## 2. 机器学习建模与传统统计模型有哪些不同？（从算法和评估模型方面）

(1)统计建模是通过假设一个合适的数据模型,然后根据数据估计模型参数。相反,机器学习方法不是从模型开始,而是使用一种算法来学习响应与预测变量之间的关系。机器学习方法是假设生成过程是复杂和未知的,试图通过观察输入和响应发现主导模式。

(2)传统统计方法选择了一个模型,没有检验任何其他假设。机器学习方法首先给出的是模型集合,包括多项式模型、指数模型等,目标是从不同模型选择最佳的一个。当然,最终的结果与回归算法结果可能是一致的。

(3)统计模型重点是描述数据与结果变量之间的关系,评估模型的合理性是通过置信区间、显著性检验和其他检验对回归参数进行分析,即统计推断。机器学习涉及训练集和测试集,模型优劣通过测试集评估。

(4)统计建模更多关注变量之间的关系和意义,是可解释的。然而,机器学习强调预测性能,而不在于模型是否具有可解释性。

## 3. 什么是贝叶斯理论？为何可用于机器学习建模的？

什么是贝叶斯理论：

在没有X与Y联合概率下,可利用X的条件概率计算Y的条件概率,这种由已知的先验概率P(Y)和条件概率P(X|Y),推算出后验概率P(Y|X)的方法,就是Bayes规则或定理。

为何可用于机器学习建模：

贝叶斯定理为思考和模拟机器学习提供了一个有用的框架,即在数据集上测试假设空间H中每个假设(h1,h2,h3,…)为真的概率,即每个类为真的概率,并寻找最大化。在此框架下,数据(D)的概率是恒定的,可从计算中删除。另外,若没有任何关于类的先验信息，可以分配一个同样概率,将其视为常数,也从计算中删除,最终得到如下公式：

P(class|data)= P(data|class)

依据公式对每个类执行此计算,并选择最大概率的类作为最终结果。

## 4. 生态学上常用树模型，包括随机森林、提升回归树，可用于分类和回归。

## 1）有哪几个重要参数？建模过程中，采用何种方法优化这些参数？

随机森林模型中有3个重要参数:

(1)森林中决策树的个数(number of trees)。增加决策树的个数会降低预测结果的方差,这样在预测时会有更高的精确度,但训练时间延长,训练时间大致与树的个数之间存在一个线性增加关系。可根据均方差(Mean Squared Error, MSE)或平均绝对值误差(Mean Absolute Error，MAE)来确定。

(2)最大树的深度(max of depth)。在CART中,每棵树深度大,意味着预测更有力，但训练时间长,也倾向过拟合。对于随机森林,不容易出现过拟合,可选择比CART更大的树深度,一般为整数(integer)或none,如果为none,所有叶子节点都是纯的,或叶子节点包含的样例数小于 min\_sample\_split。

(3)最大特征数(max of features)。对于普通决策树,从所有n个样本特征中选择一个最优特征进行左右子树划分,而随机森林是从几个特征中选择m个特征,从m中选择一个最优的特征做分割点,以增强模型泛化能力。如m=n,则RF与普通的CART决策树没有区别。m 越小,模型越健壮,模型方差会减小,但偏倚会增大。在实际中,一般选择m= (对于分类问题)或m=n/3(对于回归问题),可通过交叉验证调参获取一个合适的m 值。

除此而外,还有几个次要参数,用于终止树的生长:

(1)内部节点最少样例数(min sample of split)。具体数字(int)或数目的百分比(float)

(2)叶子节点上应有的最小样例数(min sample of leaf)。叶子节点中的样例数为具体数字或数目的百分比。叶子节点样例数过少,易遭受噪音影响,一般取值大于50。

提升回归树：

BRT的两个重要参数分别为学习速率r(也称为收缩参数,决定每棵树的贡献),和复杂度c(控制是否拟合交互，tc=1拟合一个加性模型,tc=2拟合一个具有two-way 交互作用模型),这两个参数决定了最优预测所需树的数量(nt)。

## 2）分别给出评价分类、回归模型的具体指标是什么？

随机森林：

当数据集样本少时,用交叉验证(Cross-validation)方法选择模型参数,评估模型的预测性能。常用10-折交叉验证(10-fold Cross-validation),其基本原理及分析过程如下:

(1)把原始数据集分解成10个大小近似相等的子数据集,把第1个子数据集作为验证数据集,把其余9个子数据集合并后,用于构建模型。基于该模型对验证数据集的因变量进行预测,并计算预测误差的平方和。

(2)类似地,把第2个、第3个……直至第10个子数据集分别作为验证数据集,并把其余的9个子数据集合并后构建模型,基于验证数据集计算因变量预测误差的平方和。

(3)计算前述10个预测误差平方和的平均值,平均值最小的模型为最优模型:

交叉验证的一个特例是将每个观察值作为一个子数据集,然后使用该观察值之外的其他所有观察值估计模型参数,并计算预测误差的平方。

提升回归树：

R中有许多包实现 BTR,最流行的有 gbm(generalized boosted regression models),它是广义提升回归模型,可以解决回归问题,也可以解决分类问题。gbm 包有两个主要的训练函数:gbm 和 gbm.fit,gbm 使用 formula 接口来指定模型,而 gbm.fit需要分离的x和y矩阵,用于处理多变量,函数提供了很多参数以进行模型调优。

(1)distribution。计算损失函数时,需要对输出变量的数据分布做出假设。一般来说,对于分类问题,选择bernoulli或者 adaboost;对于连续因变量,选择 gaussian或者laplace。

(2)shrinkage。学习速率,即每一步迭代中向梯度下降方向前进的步长,默认shrink-age为0.001.

(3)n.trees。即迭代次数,一般为 3000~10000,以最小化 MSE(Mean Squared Error)损失函数,gbm中的gbm.perf可以估计出最佳迭代次数。

(4)interaction.depth 和 n.minobsinnode。子决策树即基础学习器的深度和决策树叶节点包含的最小观测数,默认interaction.depth为1

在BTR模型中引人一些随机性,以提高精度和速度,并减少过拟合。gbm包通过一个“袋分数(bag fraction)”来控制随机性,该“袋分数”规定了在每一步中要选择的数据比例。默认的袋分数为0.5,这意味着在每次迭代时,50%的数据是从完整训练集中随机抽取的,没有替换。通过比较不同袋比例的预测性能和模型的可变性,可以确定最佳袋分数。

## 5. caret包提供了统一建模框架，包括分割数据，选择特征，评估模型等。

## 1）一般将数据集分割训练集与测试集，用到哪个函数？

createDataPartition()

## 2）优化模型参数用到验证集，该数据集包含在训练集或是测试集？

训练集

## 3）什么是选择特征，如何选择的，用到caret包中哪个函数？

什么是选择特征：

通过可视化方法,直接选出对目标变量影响较大的特征。

如何选择的，用到caret包中哪个函数：

选择特征的一个较好方法是递归特征消除(Recursive Feature Elimination，RFE)。递归消除特征法使用一个基模型来进行多轮训练,首先用所有特征训练模型得到每个特征权重,并剔除拥有最小权重特征,之后再基于其余特征训练,重复前面的过程,如此往复递归,直至剩余的特征数量达到所需的特征数量。

caret 使用rfe()函数实现RFE,并且可以通过定义 refControl()函数来控制 rfe 使用什么算法以及何种交叉验证方法。另外，在特征比较多的情况下，可通过size指定模型大小，即拟需要的特征数量。

## 4）通过什么函数指定优化参数方法？

caret 包中的 train()函数功能强大,配合 trControl、tuneGrid或 tuneLength参数,可进一步优化模型。

# 第三章 共享数据与数据库

## 1. 生态学数据可分为生物属性数据与环境数据，主要数据来源是什么？

现在有数千个在线数据库,如re3data.org提供了2000多个存储库的信息,OpenDAOR 包含3500以上的数据库。NCEAS、NESCent、SESYNC等机构开发的一些共享数据系统,管理和保存与生态学关系密切的数据。例如,CESTES是一个群落生态学全球数据库,汇编了80个数据集,涉及群落物种丰度或存在/缺失、物种性状、各地点的环境变量数据,涵盖不同类群、生态系统类型、人类干扰水平和空间尺度,为从事生物地理学、宏观生态学、群落生态学以及生物多样性-生态系统功能的研究人员提供重要数据源。生态学常用数据库名称、网站等信息见表3.3（课本P62）。

环境数据来源也相当广泛,概括起来包括两个方面:

(1)遥感影像及产品。如 NASA的 MODIS 和 Landsat,可解译地形地貌,包括海拔、坡度、坡向、高程等,以及土地利用、植被指数(NDVI)、火灾等数据。另外,土壤湿度、地表温变等也可从遥感影像获取。

(2)其他环境要素数据。环境要素涉及气候、大气成分、降水、土壤理化特性等,以及人口数据、生态足迹、生物资源交易等,主要是各种共享数据库,如 EPI。

根据文献及 Ecological Data Wiki、Free Gis Datasets,各国都在收集有关全球尺度的环境数据,有关的数据库见表3.4（课本P63）。

## 2. 遥感影像是重要共享数据源，举例说明生态学上可利用遥感影像获取哪些数据？

可解译地形地貌,包括海拔、坡度、坡向、高程等,以及土地利用、植被指数(NDVI)、火灾等数据。另外,土壤湿度、地表温变等也可从遥感影像获取。

## 3. 生态学常用的数据库是postgresql或postgis，QGIS自带的是哪种？可存储非空间数据？

QGIS自带的数据库是SQLite。

虽然SQLite主要用于存储非空间数据，但QGIS也支持在SQLite数据库中存储空间数据，例如矢量数据的几何信息。

## 4. 什么是数据共享？主要共享平台或途径是什么？

在大数据时代,没有一个人或机构能够容纳、管理和有效分析所有形式的生态数据。妥善管理数据,与他人分享数据是数据科学的一个重要部分。数据共享具有如下好处:

(1)通过共享数据,科学家能够通过获取各种不同的环境观测值研究不同生态系统类型的现象,可以避免重新收集数据,节省时间和科研成本。

(2)通过数据共享、开放,支持可重复性的生态学研究,促进科学团体协同解决共同关注的复杂生态问题。

(3)此外,研究数据公开,促进科学研究的过程和成果更加透明,加快知识传播,以确保科学和社会最大限度地受益。

共享平台或途径：

目前已开发一些工具,包括 DMPtools、DataOne、R、Morpho、re3data.org、KeplerVisTrails,可作为生态学家进行研究和管理数据的工具。有关数据、新软件和算法以及工作流也可以通过 DMPtools、KNB、Dryad 和myExperiment工具与他人共享。研究者可将数据发布至公共数据存储库,如 Dryad(http://datadryad.org/)、figshare(https://figsharecom/)、zenodo(https://www.zenodo.org/)、DataOne等平台,供他人免费下载使用。另外,Ecology Archives(http://esapubs.org/archive/)是ESA期刊上论文附带的数据材料。这些基本是在项目完成后的数据共享方案。

在众多选择中,比较方便的是通过 GitHub生成和发布开放、共享数据集,关键是确保Git、SVN 和 SSH 在 Rstudio 和 GitHub一致。

# 第四章 探索性数据分析

## 1. 探索性数据分析主要包括哪些主要内容，其目的是什么？

探索性数据分析（Exploratory Data Analysis，简称EDA）是统计学中一种重要的数据分析方法，主要包括以下主要内容：

1.数据可视化： EDA的核心是通过图表、图形和统计图来可视化数据，包括直方图、箱线图、散点图、密度图等。数据可视化可以帮助发现数据的分布特征、异常值、趋势和模式。

2.集中趋势度量： 这些度量包括均值、中位数、众数等，用于描述数据的中心位置。

3.离散程度度量： 这些度量包括方差、标准差、四分位距等，用于描述数据的离散程度或变异性。

4.相关性分析： 通过相关系数等方法来探索不同变量之间的相关关系，包括正相关、负相关或无相关。

5.异常值检测： 识别和处理可能存在的异常值或离群点，这些异常值可能会影响数据分析的结果。

6.缺失值处理： 分析并处理数据中可能存在的缺失值，以确保数据完整性和准确性。

7.数据分布和形态检查： 通过观察数据的分布形态，例如是否服从正态分布，来判断数据的特性。

8.数据变换： 对数据进行变换以满足统计分析的要求，例如对数变换、幂变换等。

9.模式识别： 探索数据中可能存在的模式、趋势和周期性变化，以及不同子集之间的差异。

10.探索性模型建立： 尝试建立简单的模型来描述数据，例如线性模型、非线性模型等，以便进一步分析和推断。

探索性数据分析的主要目的是帮助研究者更好地理解数据，发现数据中的规律、趋势和模式，为后续的统计推断、建模和决策提供基础。通过EDA，研究者可以对数据进行初步的分析和解释，发现数据中的潜在信息，为后续深入分析提供指导。

## 2. 群落研究中，针对样方的分析（Q-mode）是基于距离、相关系数或协方差？根据此结果的聚类，有哪几种主要聚类方法？

基于距离。

根据Q-mode分析的结果，常见的主要聚类方法包括：

1.层次聚类法（Hierarchical Clustering）： 将样方或变量逐步进行聚类，形成一个层次化的聚类结构。层次聚类方法包括聚合聚类（Agglomerative Clustering）和分裂聚类（Divisive Clustering）两种主要类型。

2.K均值聚类法（K-means Clustering）： 将样方或变量划分为预先指定的K个簇，通过迭代优化来最小化簇内的差异性，从而实现聚类分组。

3.密度聚类法（Density-based Clustering）： 基于样方或变量的密度来识别簇，例如DBSCAN（Density-Based Spatial Clustering of Applications with Noise）算法。

4.模型聚类法（Model-based Clustering）： 基于统计模型来识别样方或变量之间的聚类结构，例如高斯混合模型（Gaussian Mixture Model，GMM）。

5.谱聚类法（Spectral Clustering）： 将样方或变量之间的相似性转化为图的拉普拉斯矩阵，并通过特征向量分解来实现聚类。

这些聚类方法在Q-mode分析中可以根据不同的研究目的和数据特点选择合适的方法来进行样方或变量的聚类分析，以揭示样方之间的相似性或差异性，并识别出潜在的生态群落结构。

## 3. 针对物种或栖息地环境变量分析（R-mode）是基于距离还是相关系数、协方差？根据此结果进行排序，排序的方法有哪些？

基于相关系数、协方差。

常用的排序方法包括间接排序（非约束排序）和直接排序（约束排序）。

## 4. 排序可依据线性模型或是单峰模型，主要判断方法是R中哪个包或函数？限制性排序与非限制性排序主要区别是什么？

使用vagan包中的decorana函数

限制性排序和非限制性排序是排序分析中两种常见的方法，它们之间的主要区别在于排序过程中是否对变量之间的关系进行限制。

限制性排序（Constrained Ordination）： 在限制性排序中，排序过程受到一定的约束或限制，这些约束通常来自于研究者的假设或预先确定的模型。常见的限制性排序方法包括主成分分析（PCA）、冗余分析（RDA）、典型相关分析（CCA）等。在这些方法中，排序的结果受到了对应变量之间的约束，使得排序结果更符合预先设定的模型或假设。

非限制性排序（Unconstrained Ordination）： 在非限制性排序中，排序过程不受到任何约束，排序的结果主要基于变量之间的相似性或差异性。非限制性排序的方法包括主坐标分析（PCoA）、多维标度分析（MDS）等。这些方法通常用于探索数据的整体结构，发现变量之间的模式或趋势，而不受特定模型或假设的限制。

因此，限制性排序更适合于验证特定的假设或模型，而非限制性排序则更适用于探索数据的结构和模式。选择哪种排序方法取决于研究的目的以及对数据的理解程度。

## 5. 在双序图中，点表示样方，矢量表示物种或环境，图解排序结果要看标度，当scaling=1，矢量长度和夹角分别表示的是什么？当scaling=2，矢量长度和夹角分别表示的是什么？

当scaling=1时，矢量长度表示的是物种（或环境）在样方中的相对丰度或相对重要性。夹角则表示物种（或环境）之间的相似性或差异性。夹角越小，表示相似性越高，夹角越大，表示差异性越大。

当scaling=2时，矢量长度表示的是物种（或环境）在样方中的绝对丰度或绝对重要性。夹角仍然表示物种（或环境）之间的相似性或差异性，其含义与scaling=1时相同。

# 第五章 空间数据探索与建模

## 1. 空间数据有哪些类型？样方、研究区域、地形地貌分别以何种类型数据存储的？

地理空间数据是具有地理坐标及与坐标相关的数据,如地图、遥感数据和物种分布数据等，分为矢量数据和栅格数据。

研究区域：矢量数据

地形地貌：栅格数据

## 2. 操作空间数据可以用QGIS和R，两者可以互用，在R中有哪些包在两者之间建立沟通的桥梁，举例说明。

RQGIS、qgisprogresss

## 3. 原始卫星影像资料要经过哪些处理步骤，才能用于生态学有关分析？哪些层级的遥感资料可直接使用？其主要用途是什么？

处理步骤：

1.辐射定标:用radation=DN\*gain+offset(定标公式)将每个波段的DN转化为大气顶层辐射,然后通过近似太阳辐射的归一化,得到大气表观反射率。

2. 大气校正(atmosphericcorrection)：采用暗物体减法(DOS)等方法,消除太阳反射、气溶胶散射,获得地标信息。

3. 正射校正(topographicillumination correction)：在几何校正基础上，通过高程数据(DEM)消除地形起伏引起的图像变形。

4. 几何配准(geolocation correction)：将不同时间、不同波段、不同遥感器系统所获得的同一地区的图像(数据)，经几何变换使同名像点在位置和方位上完全叠合。

可直接使用的：

Level2(经过辐射校正和系统级几何校正)、Level3(采样地面控制点对几何校正模型进行修正)、Level4(采用地面控制点和数字高程模型对几何校正模型进行修正)。

主要用途：

遥感影像是宏观生态学研究的重要数据源,通过解译影像可获取一些重要环境参数,如地表温度(Land Surface Temperature,LST)、地表覆盖类型(河流、森林、草原或居民区)以及植被状况(植被指数 VI)等,进一步可获得土地利用类型,这些参数常作为生态建模的预测变量。

## 4. 利用空间数据，可绘制静态研究区域地图，也可建立预测模型，能否就全球变化导致物种分布改变，说明具体的建模步骤？

利用空间数据进行物种分布模型建立是一种常见的方法，以下是具体的建模步骤：

（1）数据收集：收集相关的空间数据，包括物种分布数据、环境因素数据和地理信息数据。物种分布数据可以是已知物种分布的观测数据或者物种出现的点数据；环境因素数据包括温度、降水量、海拔等影响物种分布的环境因素；地理信息数据可以是地形、土壤类型等与物种分布相关的地理信息。

（2）数据预处理：对收集到的数据进行预处理，包括数据清洗、缺失值处理和数据转换。清洗数据可以去除异常值和错误数据，确保数据的准确性；缺失值处理可以使用插补方法填充缺失值；数据转换可以对数据进行标准化或者归一化，以便后续建模分析。

（3）特征选择：根据物种分布和环境因素之间的关系，选择合适的特征变量。可以借助统计方法、地理信息系统（GIS）和专业知识来确定哪些特征变量对物种分布具有显著影响。

（4）模型选择：选择合适的建模方法来建立物种分布模型。常用的模型包括物种分布潜在分布模型（Species Distribution Models, SDMs）和最大熵模型（MaxEnt）。SDMs基于统计学原理，通过训练数据建立物种与环境因素之间的关系模型；MaxEnt模型则基于最大熵原理，将物种分布视为最不确定的分布。

（5）模型训练：使用已知的物种分布数据和环境因素数据，对选定的模型进行训练。训练过程中，模型会学习物种与环境因素之间的关系，并生成一个预测模型。

（6）模型评估：使用评估数据集对建立的模型进行评估。评估指标可以包括准确率、召回率、F1-score等，用来评估模型的预测效果。

（7）模型应用：使用建立好的模型进行预测。输入待预测的环境因素数据，模型会输出相应的物种分布预测结果。这些预测结果可以用来分析全球变化对物种分布的影响，进而制定保护措施和环境管理策略。

# 第六章 时间序列数据与建模

## 1. 时间序列数据纬度高，处理难度大，通常要降维，目前降维的主要方法有哪些？R中有哪些包可帮助实现重表达？

降维的主要方法：

分段表达时间序列是最简单的降维方法,即通过重采样,将时序p的长度m减少到n。该方法大致分为:

(1)等间隔重取样。通过等间隔重采样,减少采样点数,达到降维的目的。但如果采样率太低,采样/压缩时间序列形状将失真。

(2)感知重要点(Perceptually Important Points,PIP)表达。对于时序中所有数据点重要性进行排序,其中,第一个数据点p和最后一个数据点P。分别是两个 PIP,第三个PIP是与前两个PIP距离最大的点,第四个PIP点是与两个相邻的PIP(第一个 PIP 和第二个PIP之间)距离最大的点……直到序列中所有点作为PIP列出,这样得到关键点序列,基本上可以保持原序列变化趋势。用关键点来表示时间序列,利于异常检测。

(3)分段线性表示。一种改进的重采样方法,即不用采样点的值而是采样间隔的平均值表示,也称为分段聚合逼近(Piece-Wise Aggregate Approximation,PAA)。除了分段均值表示外,还可用分段变化的和(SegmentedSum of Variation,SSV)来表示时间序列每一段,或用直线近似分段(Piecewise inear Representation,PLR),即用数据点连线给出分段逼近。

(4)符号聚合逼近(Symbolic Aggregate Approximation,SAX)。它是用字符表示分段的均值,先将序列的时间轴X轴离散为多个相等的段,并计算段内平均值,即PAA,然后在序列值的纵轴Y按照高斯分布,分割等大小区域,依次将每个间隔映射为一个符号,然后将PAA每个分段均值转换为所在等概率区间对应的符号,即所谓符号聚集近似,这样序列简约为SAX单词。

对于每个时间序列,需要用normz函数(z-得分)或norm\_min\_max(min-max)方法进行均值为0和标准差为1的标准化,然后才能转换为其他表示。

此外,也可将时间序列从时域转换为频域,然后在频域降维。在频域中,一些数据降维方法包括奇异值分解(Singular Value Decomposition,SVD)、离散傅里叶变换(DiscreteFourier Transform,DFT)、离散小波变换(Discrete Wavelet Transform,DWT)

R中有哪些包可帮助实现重表达：

对于季节性较明显的时间序列,建议采用基于模型的重表达,通过简单平均(或中位数)或提取季节性回归系数来实现,包括 Mean Seasonal Profile(repr\_seas\_profile)、SeasonalLinear Models(repr \_lm),Seasonal Additive Model(repr\_gam),或Seasonal ExponentialSmoothing Coefficients(repr exp)。

R中的 TSrepr 包提供了多个函数平滑高噪声时间序列,包括repr\_paa()、repr\_dwt()、repr dft()等,分别用于 PAA、DWT、DFT。有关 TSrepr 包的用法,可参阅相关文献。

## 2. 不同于表格数据，时间序列数据通常只有测定时间和测定值，没有建模用的特征，需要自己构建特征，一般从时间戳或数值本身创建，基于数值创建特征的方法有哪些？

(1)日期时间特征(date time features)。即观察值与具体日期或季节有关,如日低温与雨季等有重要关系,可以选择把日期或季节作为特征。

(2)滞后特征(lagfeatures)。如今年的11月份数据与以往年份的11月份数据更相关即要关注历史上的今天。

(3)窗口特征(window features)。建模精度不仅与采用或选择的滑动窗口大小密切相关,而且与窗口内的均值、最大值等具体数据也有关。

## 3. 基于数值特征可构建预测模型，其主要依据是什么？如何确定或判读这个依据是否存在？

基于数值特征构建时序预测模型的主要依据是数据的时间顺序。时序预测模型的目标是根据过去的数值特征值来预测未来的数值特征值。

确定或判断这个依据是否存在通常可以通过以下几种方法进行：

（1）时间序列图：绘制数据随时间变化的曲线图，观察数据是否呈现出一定的趋势、周期性或季节性。如果数据呈现明显的趋势或周期性，那么有可能适合应用时序预测模型。

（2）自相关性分析：通过计算数据的自相关系数（ACF）来检测时间序列数据是否存在相关性。自相关系数反映了数据在不同滞后时间点上的相关性，如果自相关系数超过阈值或呈现显著的波动，说明数据具有时间相关性。

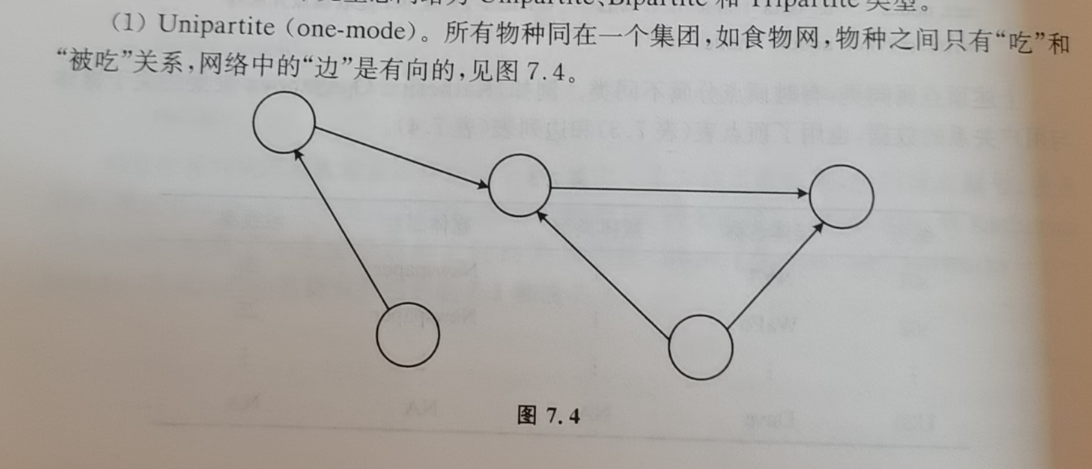
（3）平稳性检验：平稳性是指时间序列数据的统计特性在时间上保持不变。通过平稳性检验（如ADF检验、单位根检验等）判断数据是否平稳。如果数据是平稳的，那么可以应用许多经典的时序预测模型，如ARIMA、VAR等。

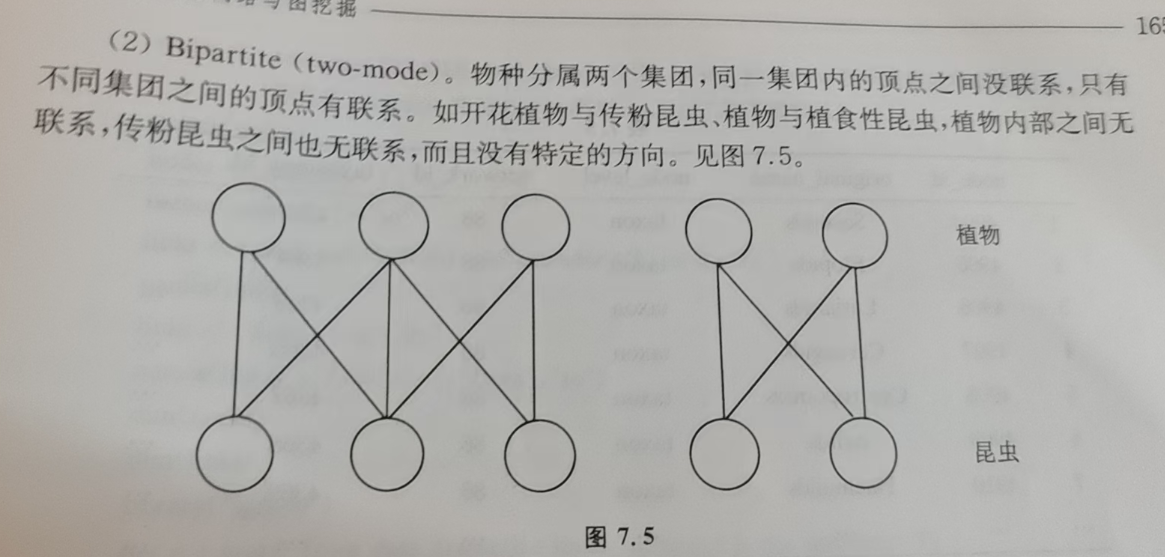
（4）季节性分解：对时间序列数据进行季节性分解，将其分解为趋势、季节和残差三个部分。如果季节性分量明显，并且趋势和残差相对平稳，那么可以考虑应用季节性模型，如季节性ARIMA、季节性指数平滑等。

（5）模型评估：使用合适的时序预测模型（如ARIMA、LSTM等）构建预测模型，并通过交叉验证、均方根误差（RMSE）、平均绝对误差（MAE）等指标来评估模型的预测能力。

# 第七章 生态网络数据与建模

## 1. 生态网络主要有one-mode和two-mode类型，主要区别是什么？可用矩阵形式表达，请举例说明。





## 2. 生态网络特性（properties）与建模特征（features）是否相同，请举例说明生态网络有哪些主要特征？

如前所述,生态网络具有一些基本特征,包括物种水平上的特征和网络水平上的特征在物种水平上,主要特征包括度、中心性和特异性(specializaton或 specility)等。在网络水平上,网络特征参数主要包括:连通性(connectance)、模块性(modularity)、嵌套性(nestedness)和特异性(specialization)。

## 3. 建立网络链接模型，基于哪些顶点及其特征，基本原理是什么？建模步骤有哪些？

