**课堂笔记**

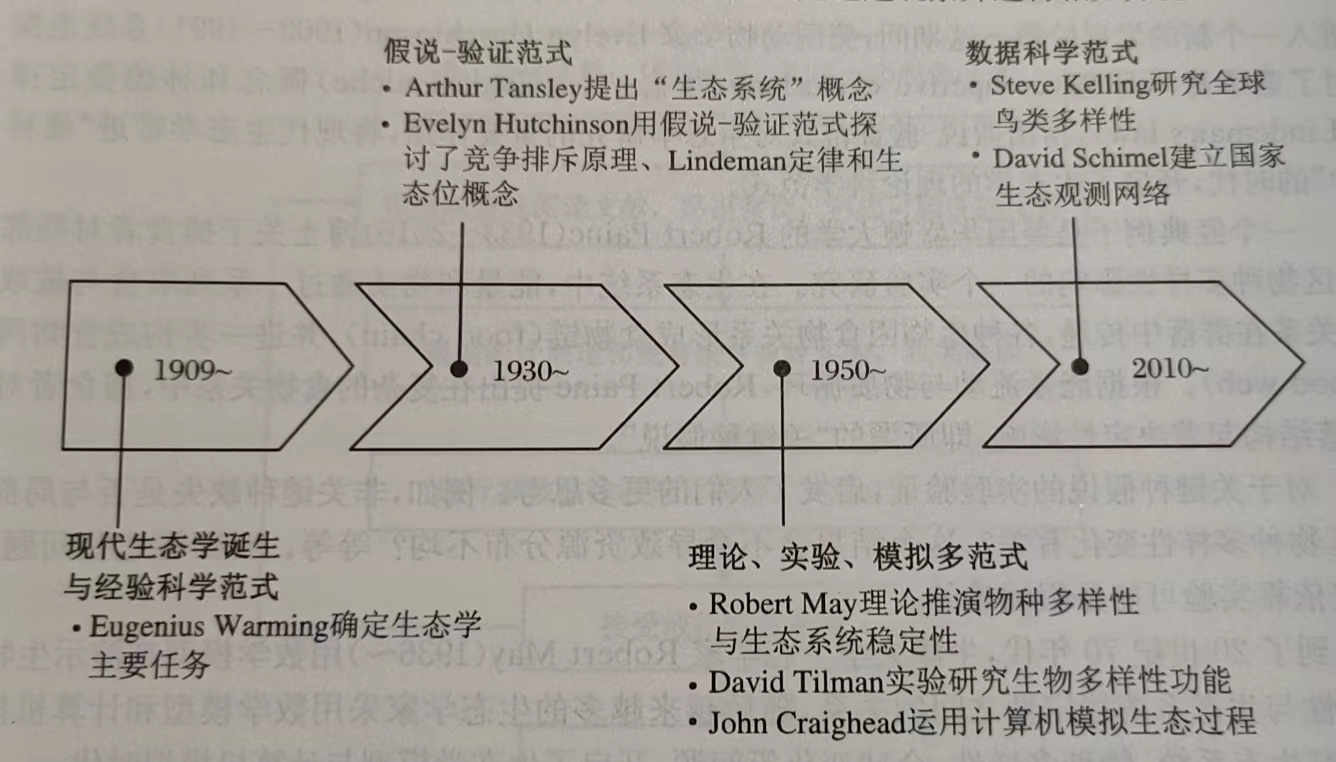
梁梦琪

**第1章 生态学研究及其范式**

**1、研究范式及演变**

范式（paradigm）是一个哲学术语，通俗地讲，它是一个学科的理论基础和实践规范的全部，包括定律、理论体系以及研究的仪器设备和具体方法，是从事某个学科研究成员公认的模式。

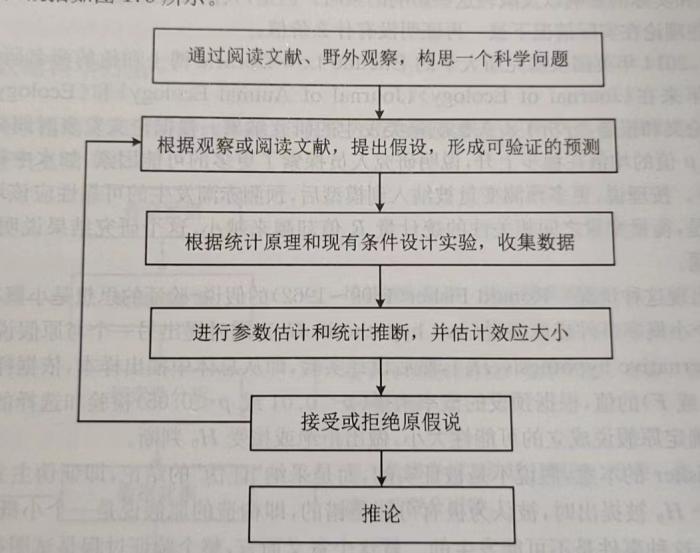
研究范式是学科范式的方法论部分。按照Jim Gray的观点，科学研究范式可分为四类：以观察实验描述自然规律为主的经验科学，用模型推演得到结论的理论科学，利用电子计算机对科学实验进行模拟仿真的计算科学，以及直接分析海量数据发现相关关系而获得新知识的数据密集型科学。



**图1.1 生态学四种研究范式**

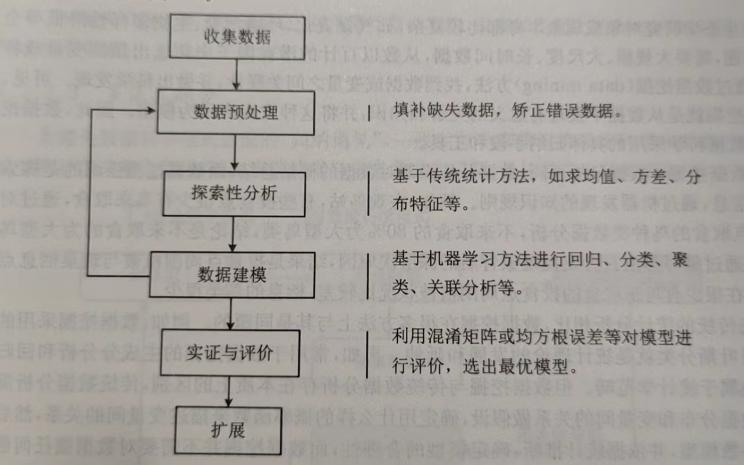
2、假说-验证范式

假说-演绎科学研究模型如下图所示。



**图1.2 假说-演绎科学研究模型**

3、数据科学方法



**图1.3 数据科学方法逻辑图**

（1）数据科学范式与其他范式的区别：

数据科学范式在科学发现的逻辑上不同于其他范式，主要表现在如下三个方面：

第一，科学发现的逻辑起点不同。经验科学范式认为“科学始于观察”，即在自然和实验观察的基础上，通过归纳提炼出科学理论，假说-验证范式主张科学发现始于科学问题，而数据科学顺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。

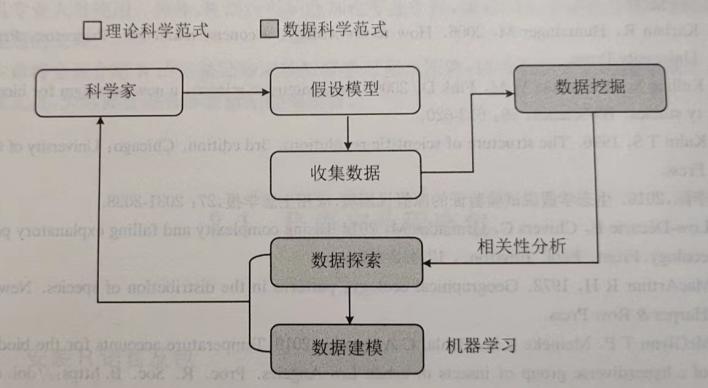
第二，寻找相关性，而不是因果律。无论是经验科学范式，还是假说-验证范式，都认为科学研究的目的就是寻找现象之间的因果关系。与其他研究范式不同，数据科学范式认为分析变量之间的相关性比探寻因果律更重要，基于相关分析的预测是大数据研究的核心。

相关不能揭示因果关系，那么数据挖掘发现的规律是否是科学规律？按照科学哲学观点，如果一个命题能够解释以往出现的现象，又能预测未来可能出现的新现象和新问题，其科学性就得到了检验。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。经验科学采用的是不完全归纳。例如，19世纪德国生物学家Carl Bergman(1814~1865)通过归纳法提出了“同一物种在越冷的地方，其个体的体积越大，外形越接近球形”这一规律。另外，从达尔文的生物进化论到种群指长的Logistic方程都采用了归纳方法。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，克服了小样本不完全归纳法的局限性，利于发现异常值。

（2）多范式融合

为避免数据科学范式造成的“归纳偏见”，一些生态学家建议将理论科学和数据密集型科学两种范式整合起来，通过多范式融合来避免这些问题。



**图1.4 多范式融合途径**

**第2章 R语言与数据挖掘**

数据挖掘主要是基于机器学习（machine learning）方法，从大量的，不完全的、有噪声的数据中提取隐含的信息。在纷繁复杂的生态学研究中，通过数据挖掘可捕捉生态过程和格局细节信息，在更大的混沌周期中发现变量之间的关系模式。

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，适合非计算机专业人员使用。另外，R语言可整合其他专业软件，如QGS，是进行生态学数据挖掘较为理想的工具。

1、安装R语言及包

R/RTools：https://cran.r-project.org/bin/windows/base/

RStudio：https://posit.co/products/open-source/rstudio/download

二、R功能与函数

R中内置了许多函数，可直接在R环境中运行，以完成一些任务，包括生成数据，提供基本运算、可视化分析以及绘图等。

**1、生成数据**

（1）read.table()。通常接受一个文件输人，并从文件读取数据。

（2）c()。接受一个范围或逗号分隔的数值列表，并用它创建一个向量。

（3）matrix()。接受一组向量或一个范围值，创建一个矩阵。常见参数nrow（指定矩阵中的行数）、ncol（指定列数）和“”。

（4）seq()。接受一个开始值和结束值，产生一个数值序列。

（5）数据转换。as.data.frame()、as.vector()、as.list()、as.matrix()接受一个输入，并强制将该值分别转换为数据框、向量、列表和矩阵。

（6）apply家族。用于在特定数据类型（列表，矩阵等）的项上重复执行，解决数据循环处理。

**2、各种运算符**

（1）“+”“-”“\*”“/”。用于数字、向量或矩阵的加、减、乘、除运算。

（2）“\* \*”或“^”。作为指数函数用，运算符前面的值是基，后面的值是幂。

（3）“%%”。模运算符，运算符前面的值除以运算符后面的值，并返回两个值相除的余数。

（4）“%\*%”。矩阵乘法符号，它的前、后各有一个矩阵，此操作将返回一个矩阵，该矩阵是前、后两个矩阵的乘积。

（5）“~”。用于R定义一个关系模型。一个简单的线性回归的格式是y~x，y~x+z 表示多元回归模型。为了得到这样一个模型的系数，使用lm作为模型参数。

（6）常见逻辑运算符：

①！=表示“不等”；②=表示相等”；③>表示“大于”；④>=表示“大于等于”；⑤|表示“或”；⑥&表示“与”。

**3、绘图函数**

（1）plot()。接受x和y坐标的向量，并在图形窗口中绘制它们。

（2）pairs()。接受数字矩阵或数据框输入，并生成一系列由输入中任意两列构成的配对图。

（3）points()。接受x和y坐标的向量，并在图形窗口中绘制指定的点

（4）abline()。接受要绘制的直线（a和b）的斜率和截距，并在图形窗口上绘制直线。

（5）hist()。接受一个向量，并绘制一个直方图，其中，x轴上的值的频率在y轴上。

（6）boxplot()。生成多个输入的比较图，对于比较多个数据样本特别有用。

**4、函数包**

目前有10000多个函数包，涵盖了数据处理和分析方方面面，任何人都可以访问、下载和使用，下面是一些常用的R包。

（1）数据操作

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

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

③stringr。对字符串类型的数据进行正则表达式处理的工具。

④lubridate。处理日期和时间类型数据的工具。

⑤httr。处理http链接的工具集合。

（2）可视化

①gplot2。功能强大的绘图工具包。

②ggvis。一个可以做基于Web的交互可视化工具包。

③rgl。在R中做3D交互可视化包。

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

⑤googleVis。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randomForest(随机森林模型)、caret(模型工具包)。

（4）编程

①Shiny。用于R交互可视化。

②R Makdown。处理数据分析报告的工具。

③jupyter notebook。提供编程环境。

**安装tidyverse函数包代码示例：**

install.packages("tidyverse")

#download and install the tidyverse package

library(tidyverse)

#load the tidyverse package

packageDescription("tidyverse")

#access information about the tidyverse package

help(package="tidyverse")

#get help document

**5、R基本操作**

（1）准备工作

①设置工作目录

开启RStudio，首先需要设置一个工作目录，即在本地PC机上新建一个文件夹，将R限制在该目录下，以便加载数据或输出结果，避免R找不到程序和工作对象。

②命令行和脚本

在RStudio中操作数据，主要通过输入R命令或脚本

③读取和保存数据

最常用的是read.csv()和write.csv()函数。read.csv()可读取矩阵或表类型的数据，并将数据（包括表头）作为数据框读到R中，经过处理可用write.csv()把数据保存起来。

（2）数据类型及操作

在R中，常用数据类型为向量、矩阵、列表和数据框。

①向量

常用来存储一维数据，包括数字向量（存储整数或小数）和字符向量（存储字符串）。

②矩阵

经常被用来保存多维数据，A%\*%B表示矩阵A与B相乘。

dim()可查看矩阵维度，dim()也可将一个向量转换为一个矩阵。使用cbind()和rbind()函数连接两个向量。但要确保两个向量都有相同的元素数，否则，它将返回NA值。

③列表

它是其他对象的集合，单个列表可以同时包含一个字符向量（字符串）、一个数字向量和一个矩阵。

④数据框

带有data.frame的列表，组件可以是向量、矩阵、列表或其他，允许多个类型数据放在一起，添加到数据框中的每个组件必须具有相同的行数。在R中，主要数据格式为数据框，理解数据框是非常重要的。

⑤索引并取值

R允许方便的访问大数据集，从而提取样本。索引运算符“[]”不仅可以用于从向量、列表、矩阵或数据框中选择单个元素，而且可以用于执行其他任务。

⑥数据转换

R提供一整套函数，用于将数据从一种类型转换为其他类型。如数据之间的转换，可执行as.data.frame()、as.vector()、as.list()、as.matrix()等代码。

（3）简单运算

①加、减、乘、除运算

R console可以用作交互式计算器，从事一些基本的计算或各种计算组合。相关代码如下：

10-(2+3)#加、减

(3\*8)/(2\*3)#乘、除

1og(12)#对数

sqrt(121) #开方

②创建变量

使用<-或=符号创建变量。

**示例：以数据框为例，编写一个简短的代码来说明一些用于数据处理的函数或包**

#import and save data

#data file import/output

emp.data<- data.frame( #Creating data frame

name = c("Raman","Rafia","Himanshu","jasmine","Yash"),

salary = c(623.3,915.2,611.0,729.0,843.25),

start\_date = as.Date(c("2012-01-01", "2013-09-23", "2014-11-15", "2014-05-11","2015-03-27")),

dept = c("Operations","IT","HR","IT","Finance"),

stringsAsFactors = FALSE

)

# save the dataframe as a csv file

install.packages("csv")

library("csv")

write.csv(emp.data, file = "D:/data-driven-ecology/liangmengqi\_exercises\_2024/employee.csv", # save in a file

)

csv\_data <- read.csv("D:/data-driven-ecology/liangmengqi\_exercises\_2024/employee.csv")

print(csv\_data)

#check data structure

print(is.data.frame(csv\_data))

#check whether a column or row has missing data

any(is.na(csv\_data[1, ]))

any(is.na(csv\_data[1, ]))

#extract values from a column or select/add a column

max\_sal<- max(csv\_data$salary) # Getting the maximum

print(max\_sal)

#transform a wider table to a long format

download.file("tinyurl.com/dcmac2017dec/data/surveys\_wide.csv",

dest="D:/data-driven-ecology/liangmengqi\_exercises\_2024/surveys\_wide.csv")

library(tidyverse)

surveys\_wide <- read.csv("D:/data-driven-ecology/liangmengqi\_exercises\_2024/surveys\_wide.csv")

surveys\_long <- surveys\_wide %>%

gather(key = species\_abbrev, value = count, -(month:plot\_id))

# inspect the data structure

str(surveys\_long)

#reverse the operation

spread(surveys\_long,key=species\_abbrev,value=count)

#visualize the data

surveys\_complete <- surveys\_wide %>%

filter(

!is.na(DM), # remove missing DM

!is.na(RM), # remove missing RM

)

write\_csv(surveys\_complete, "D:/data-driven-ecology/liangmengqi\_exercises\_2024/surveys\_complete.csv")

# Plotting scatter plot

surveys\_complete <- read\_csv("D:/data-driven-ecology/liangmengqi\_exercises\_2024/surveys\_complete.csv")

ggplot(data = surveys\_complete)

ggplot(data = surveys\_complete,

aes(x = DM, y = RM)) # define aes

picture <- ggplot(data = surveys\_complete,

aes(x = DM, y = RM)) +

geom\_point() # dot plots

picture + theme\_bw() +

theme(panel.grid = element\_blank())

ggsave("D:/data-driven-ecology/liangmengqi\_exercises\_2024/survey\_complete\_plot.png", picture, width = 15, height = 10)

4、大数据与机器学习

在生态学研究中，解释变量通常比较多。例如，鸟类的体重与年龄，性别和取食集团(Guil)都有关，植被类型与坡向、降雨和土壤养分有关等等。在大数据时代，数据量大，数据类型更加多样，变量之间存在复杂交互作用，依靠传统统计分析很难完成。机器学习中的树模型比较直观，且模型不受预测变量类型、标度、缺省值影响，特别适用于生态学建模。

树模型包括决策树(Decision Tree，DT)、随机森林(Random Forest，RF)、提升回归树(Boosting Regression Tree，BRT)等。DT是基础模型，RF基于许多决策树，最终预测是从许多树平均得出的。BRT的起点也是决策树，与随机森林不同，每棵树不是相互独立的，而设置为前一棵树的残差（预测误差），通过对建模不良特征给予更多权重，逐步减少模型与观察到的目标数据之间的差异。

**随机森林**

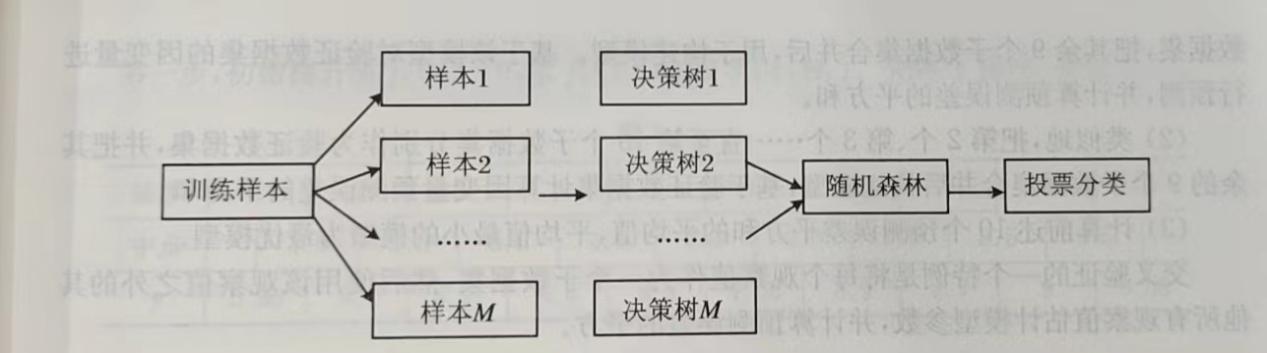
集成学习(Ensemble Learning)是使用多个学习器进行学习，然后整合各个学习器的结果得到更好学习效果的机器学习算法。主要有两类算法：

①bagging集成学习。即从总体数据集中采用有放回抽样得到N个数据集，也称为bagging方法，每个数据集训练出一个模型，然后采用N个模型预测投票（分类问题）或计算均值（回归问题）输出最后结果。随机森林采用bagging集成学习算法。

②boosting集成学习。首先根据训练集用一个初始权重训练出第一个弱学习器，再根据第一个弱学习器误差率调整样本权重，加大误差大的训练样本的权重，然后利用调整权重后样本训练第二个学习器，如此重复，直到弱学习器达到预先指定数，最后通过均值输出结果。提升回归树采用boosting集成学习算法。

随机森林是通过自助法(boot-strap)从样本N中有放回地抽样，选择N个样本（有样本可能相同)，基于Gini指数法选择特征变量，构建决策树，重复M次，构建出M棵决策树，最终形成一个含有M棵决策树的森林。

随机森林算法的原理如下图所示。



**图2.1 随机森林算法的原理图**

随机森林的建立过程如下：

①有N个样本，有放回地随机选择N个样本（每次取1个样本，放回抽样），用选择好的N个样本训练一个决策树。

②当每个样本有n个属性时，随机从这n个属性中选取出m个属性，然后从这m个属性中采用Gini指数来选择一个属性，作为该节点的分裂属性

③在决策树形成过程中，每个节点都按照步骤②来分裂，如果下一次该节点选出来的属性是刚刚父节点分裂时用过的属性，则该节点已经达到了叶子节点，无需继续分裂，注意整个决策树形成过程中没有剪枝。

④重复上述三个步骤，建立M决策树，如此形成RF 对于一个包含M棵决策树的森林，当有一测试样本，让森林中每棵决策树各自独立判断，看这个样本应该属于哪一类，然后根据M棵树中，被归为哪一类的树最多，就将测试样本归为该类别。因此，由随机森林输出的类别是由各个树输出的类别的众数决定的。

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个样本特征中选择个最优特征进行左右子树划分，而随机森林是从n个特征中选择m个特征，从m中选择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 =√n(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。