技术与方法

随机森林是特点鲜明的模型，不是万能的模型*

李欣海 [1, 2**]

（1. 中国科学院动物研究所，北京 100101；2. 中国科学院大学，北京 100049）

摘 要 随机森林（Random forest）模型在 2001 年发表后得到广泛的关注。由于随机森林可以进行回归和判别等多种统计分析，而且不受正态性、方差齐性和自变量独立性等参数检验的前提条件的制约，其应用日益普遍，有被看作万能模型的趋势。实际上，随机森林是一种特点鲜明的模型，应用局部优化拟合观察值，在分析有偏效应关系的数据时，其结果往往不准确。本文以蝉科（Cicadidea）物种的分布数据为例，比较了随机森林在回归分析时与多元线性回归、广义可加模型和人工神经网络模型的差别，在判别分析时与线性判别分析的差别，强调了随机森林预测时的碎片化特点。结果显示随机森林在处理有多元共线性和交互作用的数据时，以及在判别分析时，其准确率最高。鉴于随机森林的局限性，建议做数据分析时选择多种模型进行比较。文中的 R 语言代码可为研究者提供参考。

关键词 随机森林；偏效应；交互作用；多元共线性；R 语言

Random forest is a specific algorithm, not omnipotent for all datasets

LI Xin-Hai[1, 2**]

(1. Institute of Zoology, Chinese Academy of Sciences, Beijing 100101, China;
2. University of Chinese Academy of Sciences, Beijing 100049, China)

Abstract Random forest has gained extensive attention since its publication in 2001. Random forest can handle both regression and classification with minimum assumptions (no need for normality, homogeneity of variance, and independence between explanatory variables), so that its applications has dramatically increased. Someone even use it as an omnipotent tool for all analysis. In fact, random forest is a specific algorithm with clear characteristics. It is an ensemble method by constructing a number of decision trees, which intends to use local optimization to fit data. When the data have strong partial effect, random forest usually does not fit well. I compared the performance of random forest with multiple regression models, generalized additive models, and artificial neural network using the occurrence data of Cicadidea species. The results showed, although the prediction of random forest looked fragmented, it outperformed the other three models. Random forest also performed better than linear discriminant analysis for classifications. Random forest has its strength and weakness. I suggestion to use multiple models for data analysis rather than one “powerful” model.

Key words random forest; partial effect; interaction; multicollinearity; R

自从 Breiman（2001a）在《机器学习》杂志上发表随机森林（Random forest）模型以来，随机森林的算法就得到日益广泛的关注，截止到 2019 年 1 月 15 日，该文被 Web of Science 数据库中的文章引用了 21 052 次，已经成为许多研究人员的常用工具。随机森林通用性很好，广泛应用在生态学（Cutler *et al.*，2007）、生物信息学（Díaz-Uriarte and de Andrés，2006）、遥感（Pal，

*资助项目 Supported projects：国家自然科学基金面上项目（31772479；31572287）

**通讯作者 Corresponding author，E-mail：lixh@ioz.ac.cn

收稿日期 Received：2019-01-10；接受日期 Accepted：2019-01-17

2005)、医学（Hallett *et al.*，2014）和社会科学（Hajjem *et al.*，2014）等领域，成为机器学习算法中的重要一员（Rodriguez-Galiano *et al.*，2015)。

人们在应用随机森林的同时，也在改进和丰富随机森林的功能。Wager 等（2014）给出了随机森林预测值的置信区间。Bader-El-Den 等（2018）提出了偏随机森林（Biased random forest）的算法来处理不均衡的分类数据，避免出现较少的类别在随即取样中被忽视。Zhao 等（2019) 提出了一个可视化的工具展示随机森林的决策路径，帮助人们理解随机森林的机理。Reis 等（2019）调整了随机森林的算法，设计了概率随机森林（Probabilistic random forest)，定义了自变量（列）和因变量类别（行）的概率分布，来应对测量值的不确定性。

随机森林在处理的数据类型上具有广谱性（Cutler *et al.*，2007)：其因变量 *Y* 可以是分类变量，也可以是连续变量；其自变量 *X* 可以是若干连续变量和若干分类变量的组合。李欣海（2013）介绍随机森林可以做判别分析、逻辑斯蒂回归和多元线性回归就是强调了这个特点。应用随机森林的前提条件非常宽松，没有参数统计所要求的正态性、方差齐性和自变量独立性的限制。因此，有些人把随机森林看作万用神器，认为可以替代其他模型。实际上，随机森林是特点鲜明的模型，在众多模型中有其明确的定位。对随机森林缺点分析的文章不多，主要集中在变量的重要性方面（Verikas *et al.*, 2011；Hallett *et al.*，2014；Nembrini，2018)。Strobl 等（2007）强调随机森林在混合有不同类型的自变量时给出的变量权重不合理，对水平较多的分类变量赋以更高的权重。

大多数研究者对随机森林的机理了解不够（Biau，2012)。本文介绍了构成随机森林的决策树的原理，详细解释了随机森林的偏效应图（Partial plot)，并与多元线性回归（Multiple linear regression)、广义可加模型（Generalized additive model) 和人工神经网络（Artificial neural network）比较了在有多元共线性和交互作用时的拟合优度（Goodness of fit）指标 R^2 值，突出了随机森林的特点。本文以蝉科（Cicadidea）物种的分布数据（GBIF.org，2018）为案例，提供 R 语言代码进行分析，为大家选择模型提供参考。文中的 R 语言代码是最核心的代码，具有经验的读者可以领会关键的函数和参数。全部的代码和数据在附属文件中，可以用来重复本文的全部分析和制图过程。

1　决策树（Decision tree）的原理

随机森林是对决策树的组合，是与袋化模型（Bagging）和推进模型（Boosting）相平行的算法（Breiman，2001a)。袋化模型又称自举平均（Bootstrap aggregating)，通过有重复地再取样建模获得参数的分布；推进模型是多个弱的分类器（不够准确的决策树）通过多次迭代重新赋予权重，发展成强的分类器（准确的决策树）的模型（Hastie *et al.*，2008)。决策树包括分类树（Classification tree) 和回归树 (Regression tree)，是通过反复二分数据解释自变量和因变量的关系（Breiman *et al.*，1984)。

回归树的因变量是一个连续变量。本文用蝉科物种的分布数据解释回归树。蝉的生存面临的人类干扰，可能与其分布点的海拔、温度和降水等环境因素有关。我们假设分布于湿润区域或者分布于低海拔区域的个体面临的人类干扰较大。对于物种 *Hadoa texana*，其分布点的海拔与人类足迹指数有明显的负相关，即分布点在高海拔区域的个体面临的人类干扰比较低（图 1：A1)。对于这种关系，简单线性回归（人类足迹指数=截距+斜率 x 海拔）就可以很好地表现海拔和人类足迹指数的关系。在图 1（A1）中，回归树通过三次二分数据，把观测值分为四组，分别为海拔<210.5 m、210.5 m<海拔<289 m、289 m<海拔<378 m 和海拔>378 m，对应的人类干扰指数的平均值分别为 55.71、41.75、25.20 和 18.20。简单线性回归通过截距和斜率两个参数表现海拔与人类足迹指数的关系，而回归树用三个分叉点分成的四组数据，体现两者的关系。如果树的分

叉增多，则自变量对因变量的解释程度会增加。如果树的分叉细到每个观测值都有一个分支，则自变量 100%地解释因变量；但此时模型参数过多，是不可取的。决策树会用尽可能少分叉来最大可能地解释因变量的变异。对于物种 *Burbunga hillieri*，回归树拟合了复杂的自变量与因变量的关系（图 1：A2）。

分类树的因变量是一个分类变量。我们用 4 个蝉科物种的分布点的环境数据（Hijmans *et al.*，2005），判别该点的物种类别（图 1：B）。

回归树和分类树有多个自变量时，都会自动选择最合适的自变量，在最优的分叉点把观测值分成两组。分叉点的选择原则要保证组间差异最大，组内差异最小。

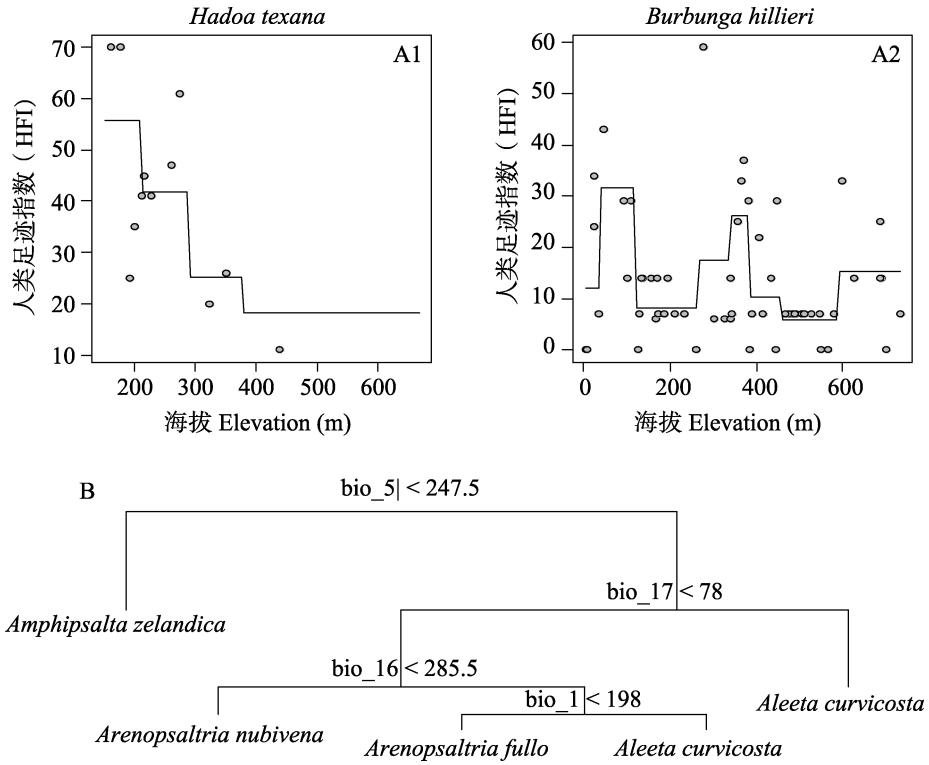

图 1　用回归树量化蝉科物种 *Hadoa texana* 和 *Burbunga hillieri* 分布点海拔和人类足迹指数的关系（A）及用分类树依据物种分布点的温湿度变量区分 4 个蝉科物种的生态位（B）

Fig. 1　Using regression tree to quantify the relationship between human footprint index and elevation of species *Hadoa texana* and *Burbunga hillieri* (A) and using classification tree to distinguish the niche of four species of Cicadidea based on the Bioclimate variables at their occurrences (B)

图 1 的分析和画图代码如下：

```
library(tree) #调用 R 语言软件包
spe <- cic[cic$species == "Burbunga hillieri", ] # 从 cic 数据库里选择物种 Burbunga hillieri 的分布点
TREE <- tree(footprint ~ alt, data=spe) #建立回归树模型
X <- seq(min(spe$alt), max(spe$alt), length. out = 100) #生成均匀分布的海拔值
Y <- predict(TREE, list(alt=X)) #计算人类足迹指数
plot(spe$alt, spe$footprint, pch=21, col="black", bg="gray", xlab = "海拔 Elevation (m)", ylab = '人类足迹指数 (HFI)', main=spe$species[1])
lines(X, Y) #画回归树折线
spe <- cic[1:500,]; spe = droplevels(spe) #选择数据库的前 500 条记录（一个随意的选择）
TREE <- tree(species ~ landuse + footprint + alt + bio_1 + bio_5 + bio_6 + bio_7 + bio_12 + bio_16 + bio_17, data = spe) #建立分类树模型
plot(TREE); text(TREE) #画分类树
```

2　随机森林的偏效应图

偏效应图是人们在应用随机森林时最常展示的结果之一，它可以给出每个自变量和因变量关系。然而很少有人真正理解这是什么关系。R

软件包 randomForest 的作者 Liaw 和 Wiener（2002）在定义偏效应图 partialPlot 时，直接说明它是画出自变量对因变量的边际效应（Marginal effect）。边际效应是指在忽视其他所有自变量的情况下一个自变量对因变量的影响。偏效应（Partial effect）是与之相对的概念，指控制其他所有自变量保持不变时一个自变量对因变量的影响。我们可以通过下列数据理解边际效应和偏效应。

*X*1 <- c(1, 2, 3, 4, 5, 6, 7, 8, 9) # 简写为 c(1:9)

*X*2 <- c(2, 2, 2, 4, 4, 4, 6, 6, 6)

Y <- c(3, 2, 1, 6, 5, 4, 9, 8, 7)

对于上述三个变量 X_1、X_2 和 Y，如果忽视 X_2，做回归 lm（Y~X_1）得到斜率为 0.8，即 X_1 增加 1 会让 Y 增加 0.8。这个斜率 0.8 是 X_1 对 Y 的边际效应。如果考虑 X_2，做回归 lm($Y \sim X_1 + X_2$) 得到斜率为-1，即 X_1 增加 1 会导致 Y 降低 1。这个斜率-1 是 X_1 对 Y 的偏效应，是控制 X_2 时 X_1 对 Y 的真正的效应。X_1 从 1 累进到 9，Y 的取值也是大致由小到大，所以 X_1 与 Y 有明显的正相关。然而在 X_2 为 2 时，Y 的取值与 X_1 相反。在 X_2 为 4 和 6 时，情况依然。应用多元线性回归，X_1 和 X_2 可以 100%地解释 Y 的变化（图 2：A）。应用随机森林，X_1 和 X_2 只能解释 57.1%的 Y 的变化（方差和，Sum of square）（图 2：B）。而且，随机森林给出的 X_1 对 Y 的偏效应图（图 2：B）并不是真正的偏效应，而是类似边际效应。通过回归模型和随机森林可以分别预测 Y 在 X_1-X_2 的整个取值面上的值（图 2），可以看出随机森林的预测是碎片化的（图 2：B），这是决策树的本质（二分数据）所决定的。

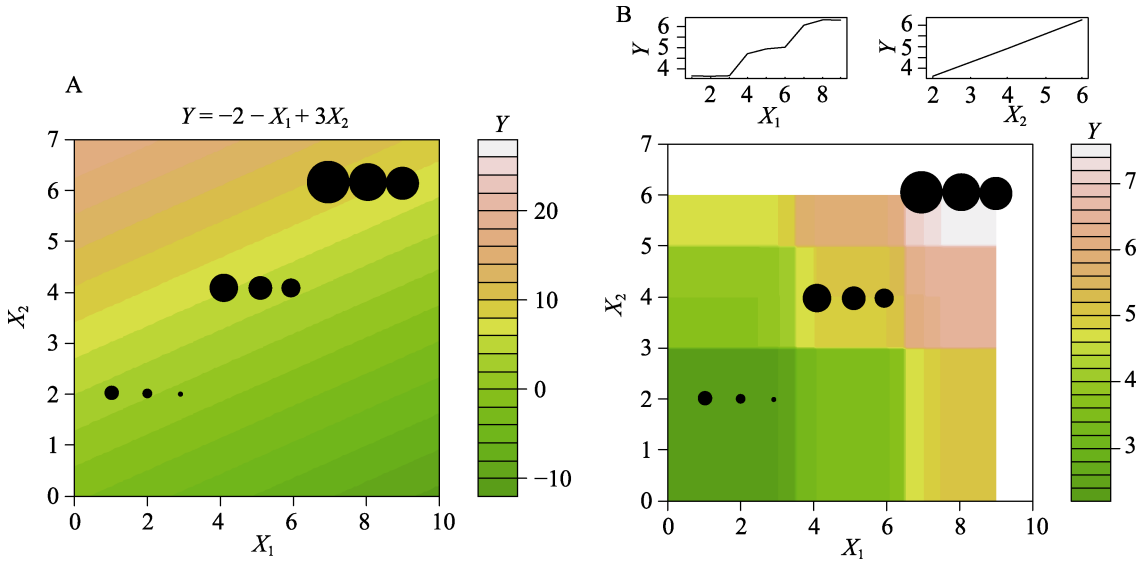

图 2　用多元线性回归（A）和随机森林（B）对具有偏效应的变量 X_1、X_2 和 Y 进行预测

Fig. 2　Predicting the values of *Y* using multiple regression (A) and random forest (B) on the basis of two variables X_1 and X_2, which have strong partial effect on *Y*

圆圈大小表示 Y 的取值大小，背景颜色表示模型的预测值。图 B 上部为随机森林给出的偏效应图。

The sizes of the circles represent the actual values of *Y*, the colors represent the predicted values of *Y*. The upper panel of Fig. 2: B shows the partial plots of X_1 and X_2 based on random forest.

图 2 的分析和画图代码如下：

```
# 多元线性回归
fit <- summary(lm(Y ~ X1 + X2))
interception <- fit[[4]][1, 1]
coef_X1      <- fit[[4]][2, 1]
coef_X2      <- fit[[4]][3, 1]

X.1 <- seq(min(X1), max(X1), length = 100); X.2 <- seq(min(X2), max(X2), length = 100)
f <- function(X.1, X.2) { r <- interception + coef_X1*X.1 + coef_X2*X.2 };
Y <- outer(X.1, X.2, f)
filled.contour(X.1, X.2, Y, main='', color = terrain.colors, xlab=eXpression(paste(X[1])), ylab=
```

```
expression(paste(X[2])))
    points(X1, X2, pch=16, cex=Y) # filled.contour 有bug，需要手工调整
    # 随机森林
    RF <- randomForest(Y ~ X1 + X2, importance= TRUE, ntree=1000)
    X.1 <- seq(min(X1), max(X1), length= 100);
    X.2 <- seq(min(X2), maX(X2), length= 100)
    data <- expand.grid(X.1, X.2);
    names(data) = c('X1', 'X2')
    Y <- predict(RF, newdata=data);
    Y <- matrix (Y, nrow=100, ncol=100)
    filled.contour(X.1, X.2, Y, main=paste(''), color = terrain.colors, xlab=expression(paste(X[1])), ylab=expression(paste(X[2])))
    # 偏效应图
D <- cbind(X1, X2, Y)
    partialPlot(RF, D, X1, xlab=expression(paste (X[1])), Ylab='Y', main='')
partialPlot(RF, D, X2, xlab=expression(paste (X[2])), ylab='Y', main='')
```

R 软件包 randomForest 的偏效应函数 partialPlot 给出的是随机森林的偏效应，即在随机森林算法中全面考虑了所有其他变量的影响后，计算出的一个自变量对因变量的影响。下面用蝉科物种 *Atrapsalta encaustica* 分布点的数据进一步解释随机森林的偏效应。对于该物种的 135 个分布点，我们可以用其海拔和年降水量解释人类足迹指数，检查分布于较高海拔较干燥区域的个体是否受到的人类干扰较低。图 3 显示了 4 种情况下海拔对人类足迹指数的作用：1. 构建随机森林模型 RF1，用海拔解释人类足迹指数，画出海拔对人类足迹指数的偏效应（图 3 中的红线）；2. 构建随机森林模型 RF2，用海拔和年总降水解释人类足迹指数，画出海拔对人类足迹指数的偏效应（图 3 中的蓝线）；3. 用 RF1 预测海拔的取值梯度内人类足迹指数的大小（图 3 中的绿线）；4. 人类足迹指数的局部平滑曲线（图 3 中的灰线）。可以看到随机森林的偏效应图在是否考虑其他自变量时有很大不同（图 3 中红线和蓝线）。用 RF1 预测的人类足迹指数与 RF1 的偏效应图非常类似，然而所用的数据不同：前者用了从最低到最高海拔范围内 100 个均匀分布的海拔值进行预测，后者用实际的 135 个海拔值进行计算。

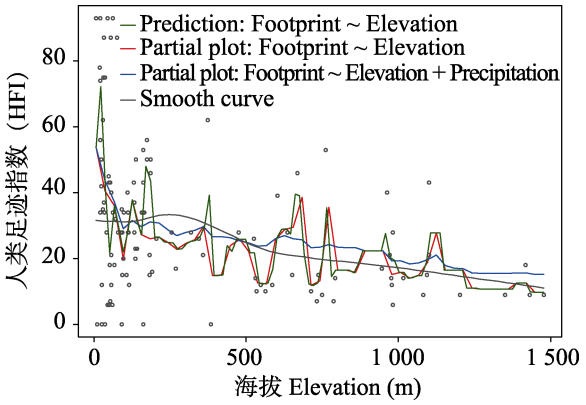

图 3 随机森林自变量对因变量的偏效应

Fig. 3 The partial plot in random forest

蝉科物种*Atrapsalta encaustica*分布点的人类干扰指数可以被海拔这一个变量解释，形成随机森林模型 RF1；用海拔和年降水两个变量解释，形成随机森林模型 RF2。图中显示应用 RF1 预测人类干扰指数的取值（绿线）、RF1 的偏效应图（红线）、RF2 的偏效应图（蓝线）和人类干扰指数的局部平滑曲线（灰线）。

The human footprint index (HFI) at the occurrences of species *Atrapsalta encaustica* was explained by elevation only (random forest model 1 (RF1)), and by both elevation and annual total precipitation (random forest model 2 (RF2)). The green line is the predicted HFI using RF1 based on 100 evenly distributed elevations. The red line is the partial plot (elevation vs. HFI) of RF1. The green line is the partial plot (elevation vs. HFI) of RF2. The gray line is the smooth line for elevation-HFI relationship.

图 3 的分析和画图代码如下：

```
spe <- gal[gal$species == 'Atrapsalta encaustica',]
    RF1 <- randomForest(footprint ~ alt, data = spe, prox = TRUE, importance = TRUE, ntree = 1000)
    part1 <- partialPlot(RF1, spe, alt) #红线
    RF2 <- randomForest(footprint ~ alt + bio_12, data=spe, prox=TRUE, importance = TRUE, ntree = 1000)
    part2 <- partialPlot(RF2, spe, alt) #蓝线
    plot(part1, type='l', xlab = "海拔 Elevation (m)", ylab = '人类足迹指数 (HFI)', main='',
    col=2, ylim = c(0, max(spe$footprint)))
    lines(part2, col = "blue")
    legend("topright", legend=c("Prediction: Footprint ~ Elevation", "Partial plot: Footprint ~ Elevation",
           "Partial plot: Footprint ~ Elevation + Precipitation", "Smooth curve"),
           lty = 1, col = c('darkgreen', 'red', 'blue', 'gray'), lwd = c(1,1,1,2), cex = 0.8)
    points(spe$alt, spe$footprint, col = adjustcolor ("black", alpha=0.5), cex =.6, pch = 16)
    x.1 <- seq(min(spe$alt), max(spe$alt), length = 100)
    data <- as.data.frame(x.1); names(data) = c('alt')
    Y1 = predict(RF1, newdata = data) #绿线
```

```
lines(data$alt, Y1, col = "darkgreen") #绿线
smooth <- loess.smooth(spe$alt, spe$footprint) #灰线
lines(smooth, col = "gray", lwd = 2) #灰线
```

随机森林量化的自变量和因变量的关系比局部平滑曲线要详细得多。这样对因变量的解释程度会大幅度提升，但是损失了通用性，本质上是过度拟合（Overfit）。Breiman（2001a，2001b）反复强调随机森林不会过度拟合，是指相对于一个决策树（Breiman *et al.*，1984），随机森林应用了成百上千棵树，并没有进一步地过度拟合。决策树用大量节点划分数据，本身就有过度拟合的倾向（Hastie *et al.*，2008）。

随机森林的偏效应图中，纵坐标的意义有时令人费解。当因变量是连续变量时，其纵坐标是其预测值的平均值，定义如下：$\bar{f}(\mathrm{x})=\frac{1}{n}\sum_{1}^{n}f(x,x_{\mathrm{iC}})$，其中 x 是产生偏效应的自变量，x_{iC} 是其他自变量。如果因变量是分类变量，纵坐标是对预测类别投票比例的 logit 转换值（log（投此类别的票数/投其他类别的票数））。

3　与其他模型的比较

本文比较了随机森林在两种数据情形下与其他几种模型的差别。两种数据情形是有多元共线性（Multicollinearity）的数据和有交互作用（Interaction）的数据。与随机森林相比的模型是多元线性回归、广义可加模型和人工神经网络。广义可加模型是进行了局部拟合的回归模型（Hastie and Tibshirani，1986）；人工神经网络是经典的机器学习模型（Hopfield，1982）。两者是常用的拟合优度较好的模型之一（Li and Wang，2013）。

3.1　多元共线性

随机森林擅长处理有多元共线性的数据（Breiman，2001b）。对于蝉科物种 *Cicadetta petryi*，其分布点的海拔和年降水有强烈的正相关（图 4）。如果用这两个变量预测分布点的人

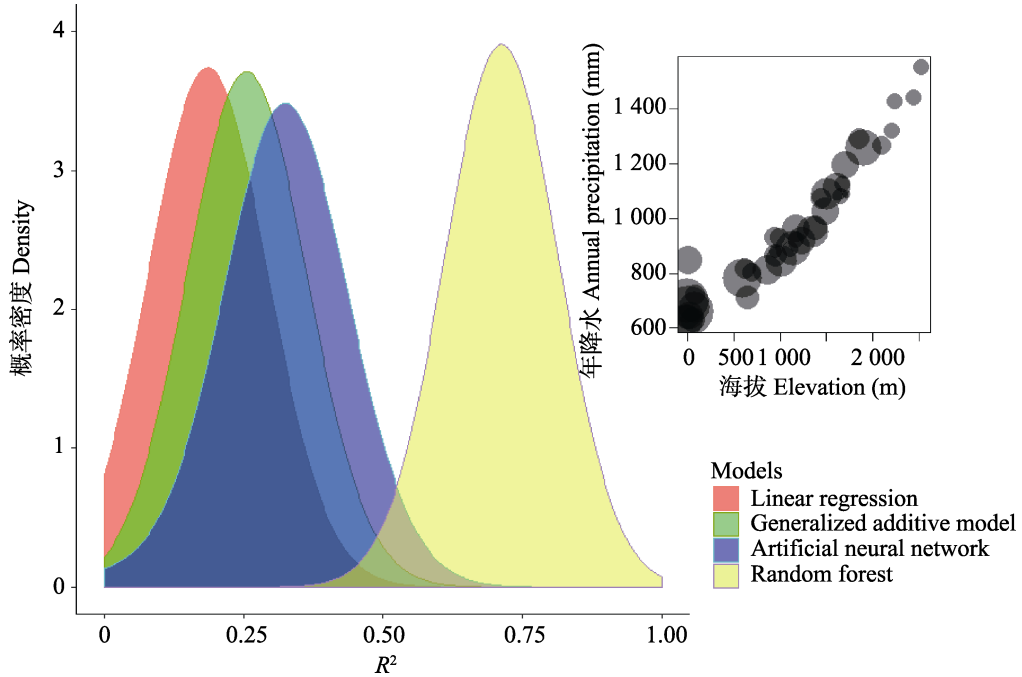

图 4　用 R^2 的分布比较多元线性回归、广义可加模型、人工神经网络和随机森林对具有强烈多元共线性的数据的拟合效果

Fig. 4　The distribution of R^2 values using linear model, generalized additive model, artificial neural network and random forest for the model (human footprint index - elevation + precipitation)

模型是用蝉科物种 *Cicadetta petryi* 分布点的海拔和年降水预测其人类干扰指数，数据是 100 次随机无重复 90%取样的分布点。

Based on 100 samples of the occurrences data of *Cicadetta petryi*.

类足迹指数，多元共线性是不可避免的。对于多元线性回归，多元共线性会导致回归系数不稳健（Robust），严重影响模型的准确性。物种 *Cicadetta petryi* 有 41 分布点。为了计算模型的稳健性，随机抽取其中 90%的分布点，重复 100 次，每次分别用多元线性回归、广义可加模型、人工神经网络和随机森林进行拟合，得到 R^2 值。对于广义可加模型，应用了 mgcv 包的 gam 函数，没有限定自变量的自由度。对于人工神经网络，用了 neuralnet 包的 neuralnet 函数，给定了 3 个隐含层，节点数分别为 20、10 和 5，远超过数据要求的复杂度。R^2 值都是通过预测值与观测值手工计算的（见下面的代码）。随机森林给出的方差解释率（% Var explained）是最后一棵树的 R^2 值，与其他树的 R^2 值有差别，因此本文弃之不用，而是用预测值与观测值手工计算。汇总的结果可以看到随机森林的 R^2 值远远超出了其他 3 个模型（图 4）。

图 4 的分析和画图代码如下：

```
library(mgcv); library(neuralnet)
spe <- gal[gal$species == 'Cicadetta petryi', ] #Cicadettana calliope 备选
plot(spe$alt, spe$bio_12, cex = spe$footprint/10, col=adjustcolor("black", alpha=0.5), pch=16,
xlab = "海拔 Elevation (m)", ylab = '年降水 Annual precipitation (mm))')
R2_L <- numeric(); R2_R <- numeric();
R2_ G <- numeric(); R2_A <- numeric()
for (i in 1:100){
  sam <- sample(1:nrow(spe), floor(nrow(spe)/ 1.1), rep = F); ite <-spe[sam,]
  # 多元回归
  fit <- summary(lm(footprint ~ alt + bio_12, data = ite)); R2_L[i] <- fit$adj.r.squared
  # 随机森林
  RF <- randomForest(footprint ~ alt + bio_ 12, data = ite, importance=F, ntree=1000)
  pred <- numeric(); pred <- predict(RF, newdata = ite)
  R2_R[i] <- 1 - var(ite$footprint-pred) / var(ite$footprint)
  # 广义可加模型
  fit <- gam(footprint ~ s(alt) + s(bio_12), data = ite); pred <- predict(fit, newdata = ite)
  R2_G[i] <- 1 - var(ite$footprint-pred)/var (ite$footprint)
  # 人工神经网络
  Data = ite[,c(4, 9, 3)] #选取海拔、降水和人类足迹指数
  maxs <- apply(Data, 2, max); mins <- apply (Data, 2, min)
  Data <- as.data.frame(scale(Data, center = mins, scale = maxs - mins)) #very important!!!
  nn <- neuralnet(footprint - alt + bio_12, data = Data, hidden = c(20, 10, 5), threshold = 0.01, linear.output = T)
  pred <- compute(nn, Data[, c(1,2)])$net. result
  pred <- pred *(maxs[3]-mins[3])+min (mins[3])
  R2_A[i] <- 1 - var(ite$footprint-pred)/var (ite$footprint)
}
#画图
library(ggplot2)
D1 <- data.frame(ID = 1:100, Models = "Linear regression", index = R2_L)
D2 <- data.frame(ID = 1:100, Models = "Generalized additive model", index = R2_G)
D3 <- data.frame(ID=1:100, Models = "Artificial neural network", index = R2_A)
D4 <- data.frame(ID=1:100, Models = "Random forest", index = R2_R)
DD <- rbind(D1, D2, D3, D4)
ggplot(DD, aes(index, fill = Models, colour = Models)) +
  geom_density(alpha = 0.5, bw=0.1) + xlim(0, 1) + ylim(0, 4) +
  theme(panel.grid.major = element_blank(), panel.grid.minor = element_blank(),
        panel.background = element_blank(), axis.line = element_line(colour = "black")) + xlab(expression(paste(R^2))) + ylab("概率密度 Density") +
  scale_fill_manual( values = c("red", "green", "blue", "yellow"))
```

3.2 交互作用

随机森林擅长处理有交互作用的数据（Breiman，2001b）。交互作用是指自变量 X_1 的变化会导致因变量 Y 与另一个自变量 X_2 的关系发生改变。对于蝉科物种 *Zammara smaragdula*，其分布点的海拔和年降水在解释人类足迹指数时有明显的交互作用（图 5）。本文分别用多元线性回归、广义可加模型、人工神经网络和随机森林预测了在海拔和年降水的取值平面上人类足迹指数的大小（图 5），四种模型的 R^2 分别为 7%、57.3%、7.5%和 87.1%。随机森林的拟合优度遥遥领先。随机森林能充分量化交互作用，但是它并不提示交互作用是否很强。研究者一般是通过其他模型，如线性模型，发现交互作用是否显著的。

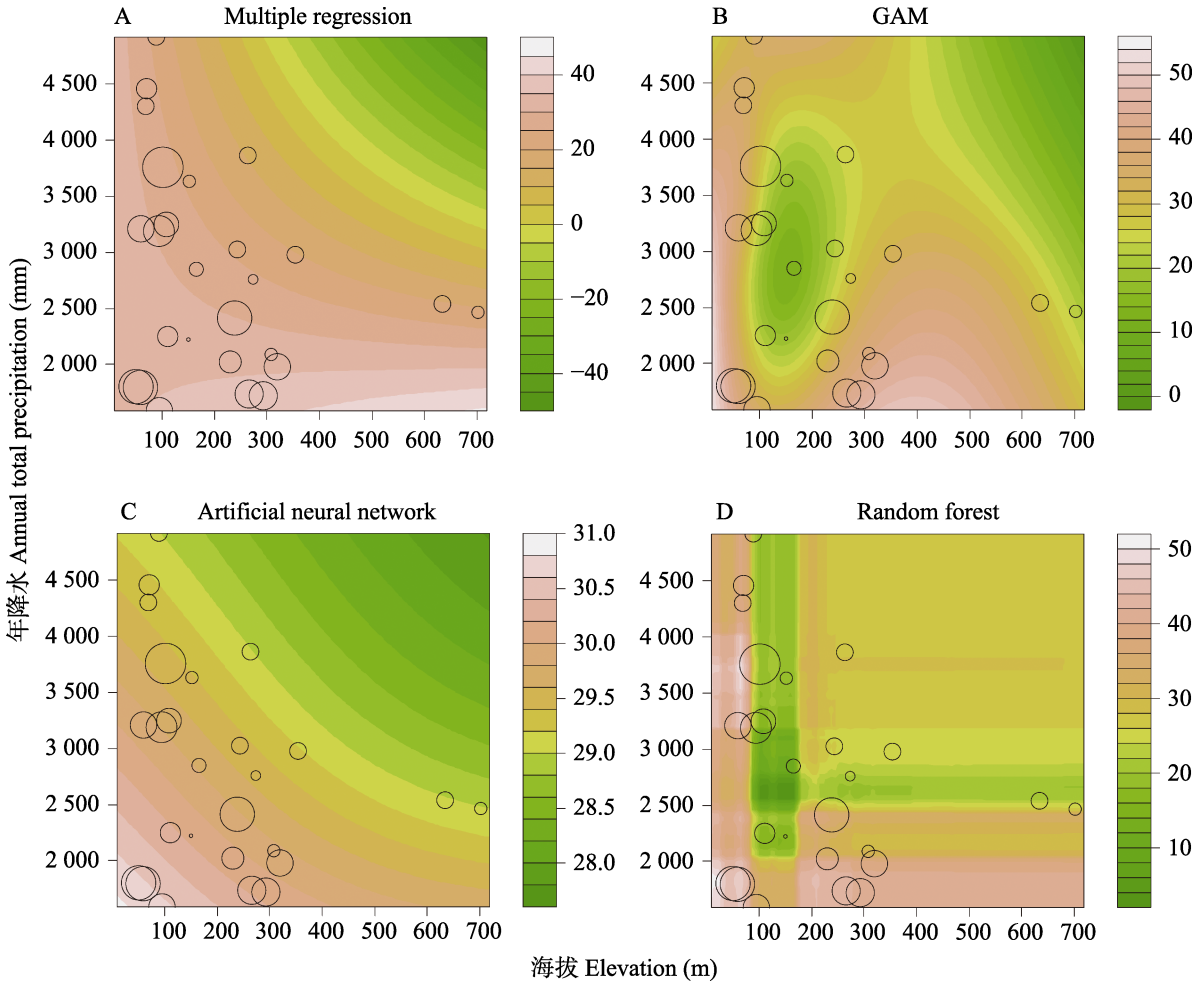

图 5　分别用多元回归（A）、广义可加模型（B）、人工神经网络（C）和随机森林（D）预测了蝉科物种 ***Zammara smaragdula*** 分布点海拔和年降水的取值平面上的人类足迹指数

Fig. 5　Using multiple regression (A), generalized additive model (B), artificial neural network (C), and random forest (D) to predict the human footprint index (the gradient color) at the occurrences of the species *Zammara smaragdula*

横坐标是海拔（m），纵坐标是年降水量（mm），圆圈大小表示分布点人类足迹指数的大小。

Based on elevation (the X axis) and annual total precipitation (the Y axis). The sizes of the circles represent the values of human footprint index of the occurrences.

图 5 的分析和画图代码如下：（图 5 的大部分绘图代码前面已经包括，下面列出广义可加模型的交互作用的代码）

```
spe <- gal[gal$species == "Zammara smaragdula", ]
x1 <- spe$alt; x2 <- spe$bio_12; y <- spe$footprint
fit <- gam(y ~ s(x1, k = 3) + s(x2, k = 3) + te(x1, x2))
# ti((RP, WB)) is better when main effect is large
```

3.3　判别分析

随机森林的算法特别适合进行判别分析，因为决策树的基本方法就是通过反复二分数据来分类的。对于蝉科的数据，选择前 500 个分布点，包含 5 个物种分布数据（这是一个随意的选择，只为演示判别分析）。本文分别构建线性判别分析和随机森林，利用每个分布点的 9 个变量，区分 5 个物种的生态位。结果显示线性判别分析的正确率是 89.6%，随机森林的准确率是 97.8%。这个准确率高得令人惊讶，反映了这 5 个物种的生态位（温湿度、海拔和人类干扰）分离得非常充分。

图 6 的分析和画图代码如下：

```
spe <- gal[1:500,]; spe <- droplevels(spe)
RF   <- randomForest(species ~ landuse + footprint + alt + bio_1 + bio_5 + bio_6 + bio_7 + bio_12+
bio_16+ bio_17, data = spe, importance = T, ntree = 1000)
```

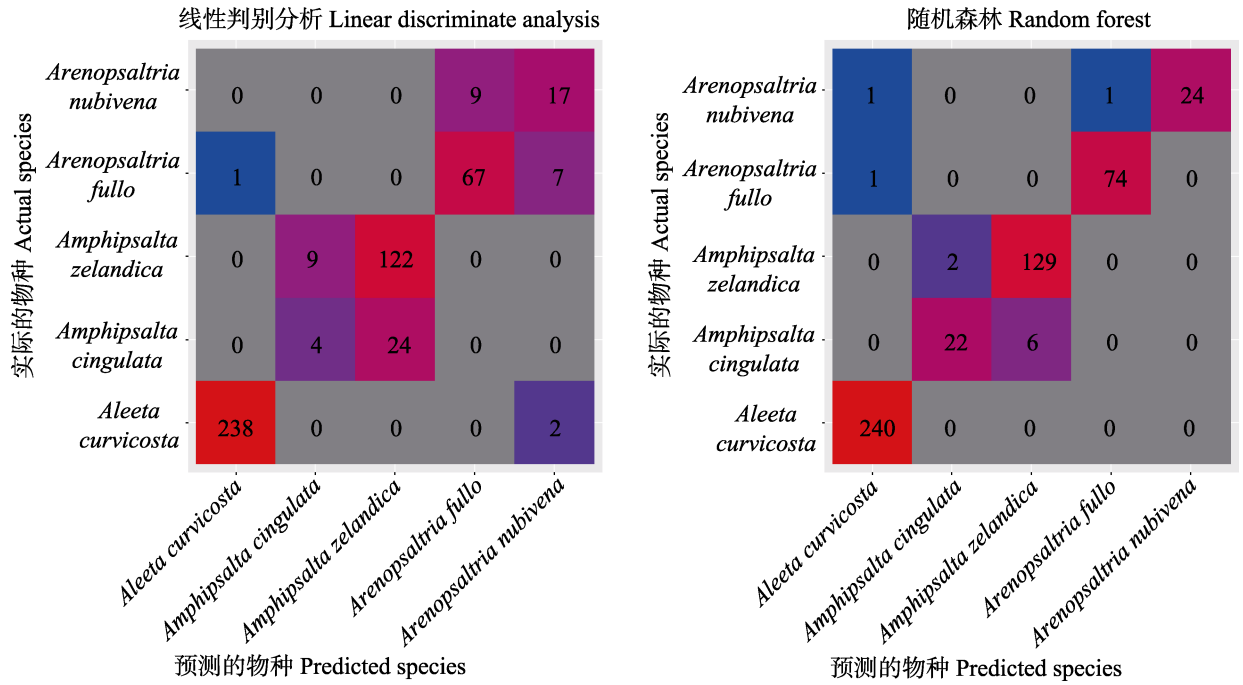

图 6　应用线性判别分析和随机森林利用分布点的 9 个环境变量进行蝉科物种判别的混淆矩阵

Fig. 6　The confusion matrix for classifying species of Cicadidea at the occurrences using linear discriminant analysis (left panel) and random forest (right panel) based on 9 environmental variables

```
RF.predicted <- predict(RF, newdata=spe) #随机森林的预测值
library(MASS)
lda.result <- lda(species ~ landuse + footprint + alt + bio_1 + bio_5 + bio_6 + bio_7 + bio_12 +
bio_16 + bio_17, data=spe)
lda.predict <- predict(lda.result, spe) #线性判别分析得而预测值
compare <- data.frame(predicted_class = lda.predict$class, actual_class = spe$species)
compare <- data.frame(predicted_class = RF.predicted, actual_class = spe$species)
confusion_matrix <- as.data.frame(table (compare))
ggplot(data = confusion_matrix, mapping = aes(x = predicted_class, y = actual_class)) +
geom_tile(aes(fill = Freq)) + geom_text(aes(label = sprintf("%1.0f", Freq)), vjust = 1) +
scale_fill_gradient(low = "blue", high = "red", trans = "log")
```

4　讨论

随机森林作为决策树的组合模型，具有鲜明的决策树的特点，在量化自变量和因变量的关系时，把数据分成若干组，分别对应因变量和若干自变量的组合。在预测上，其表现为区域化、碎片化的模式（图 2：B 和图 5：D），与回归模型的梯度变化有很大不同。这种典型的局部优化的、非参数特点的模型行为，在量化多变量（即高维数据）、复杂关系（交互作用、高次项等非线性关系）的情形下往往表现优异（Kim *et al.*，2009；Winham *et al.*，2012）。

随机森林的偏效应图信息量很大，给出了详细的自变量与因变量的关系（Gröemping，2009）。这个偏效应图在数值上与线性模型的边际效应类似，貌似是忽视了其他自变量后的 *X*-*Y* 关系，实质上是随机森林算法下的偏效应（图 3）。随机森林算法下的偏效应有时并不准确，在 *X*-*Y* 关系明确时其碎片化的拟合反而扭曲了真正的关系（图 2）。

决策树在预测上往往是过度拟合的（Hastie *et al.*，2008）。随机森林通过投票的机制降低了过度拟合的程度（Breiman，2001a），但是与线性模型相比，其预测依旧会过度拟合（Elith and Graham，2009），表现为对现有数据匹配得很好，而对没有数据的区域预测得非常保守，假阴性（Omission）的概率较高（李欣海等，2019）。

不可否认，随机森林是非常优秀的模型，在大量的同其他模型的比较中表现优异

（Kampichler *et al.*，2010）。本文强调了随机森林的特点，目的是避免对这个方法的神化，在分析数据时针对具体数据情况选择多种模型，更好地解决科学问题。

参考文献 (References)

Bader-El-Den M, Teitei E, Perry T, 2018. Biased random forest for dealing with the class imbalance problem. *IEEE Transactions on Neural Networks and Learning Systems*, doi: 10.1109/TNNLS.2018.2878400.

Biau G, 2012. Analysis of a random forests model. *Journal of Machine Learning Research*, 13(4): 1063–1095.

Breiman L, 2001a. Random forests. *Machine Learning*, 45(1): 5–32.

Breiman L, 2001b. Statistical modeling: the two cultures. *Statistical Science*, 16(3): 199–215.

Breiman L, JFriedman JH, Olshen RA, Stone CJ, 1984. Classification and Regression Trees. New York: Chapman and Hall. 358.

Cutler DR, Edwards Jr TC, Beard KH, Cutler A, Hess KT, 2007. Random forests for classification in ecology. *Ecology*, 88(11): 2783–2792.

Díaz-Uriarte R, de Andrés SA, 2006. Gene selection and classification of microarray data using random forest. *BMC Bioinformatics*, 7: 3.

Elith J, Graham CH, 2009. Do they? How do they? Why do they differ? on finding reasons for differing performances of species distribution models. *Ecography*, 32: 66–77.

GBIF.org, 2018. GBIF occurrence download. https://doi.org/10.15468/dl.mqaniq (29 December 2018).

Gröemping U, 2009. Variable importance assessment in regression: linear regression versus random forest. *American Statistician*, 63(4): 308–319.

Hajjem A, Bellavance F, Larocque D, 2014. Mixed-effects random forest for clustered data. *Journal of Statistical Computation and Simulation*, 84(6): 1313–1328.

Hallett MJ, Fan JJ, Su XG, Levine RA, Nunn ME, 2014. Random forest and variable importance rankings for correlated survival data, with applications to tooth loss. *Statistical Modelling*, 14(9): 523–547.

Hastie T, Tibshirani R, Friedman J, 2008. The Elements of Statistical Learning (2nd ed.). Stanford: Springer. 745.

Hastie TJ, Tibshirani R, 1986. Generalized additive models. *Statistical Science*, 1(3): 297–310.

Hijmans RJ, Cameron SE, Parra JL, Jones PG, Jarvis A, 2005. Very high resolution interpolated climate surfaces for global land areas. *International Journal of Climatology*, 25(12): 1965–1978.

Hopfield JJ, 1982. Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences of the United States of America-Biological Sciences*, 79(8): 2554–2558.

Kampichler C, WielandR, Calmé S, Weissenberger H, Arriaga-Weiss S, 2010. Classification in conservation biology: a comparison of five machine-learning methods. *Ecological Informatics*, 5(6): 441–450.

Kim Y, Wojciechowski R, Sung H, Mathias RA, Wang L, Klein AP, Lenroot RK, Malley J, Bailey-Wilson JE, 2009. Evaluation of random forests performance for genome-wide association studies in the presence of interaction effects. *BMC Proceedings*, 3(Suppl.7): S64.

Li XH, Gao EH, Li BD, Zhan XJ, 2019. Estimating abundance of Tibetan wild ass, Tibetan gazelle and Tibetan antelope using species distribution models and distance sampling. *Scientia Sinica Vitae*, 49, doi: 10.1360/N052018-000171. [李欣海, 郜二虎, 李百度, 詹祥江, 2019. 用物种分布模型和距离抽样估计三江源藏野驴、藏原羚和藏羚羊的数量. 中国科学: 生命科学, 49, doi: 10.1360/N052018-000171.].

Li XH, 2013. Using random forest for classification and regression. *Chinese Journal of Applied Entomology*, 50(4): 1190–1197. [李欣海, 2013. 随机森林模型在分类与回归分析中的应用. 应用昆虫学报, 50(4): 1190–1197.]

Li XH, Wang Y, 2013. Applying various algorithms for species distribution modeling. *Integrative Zoology*, 8 (2): 124–135.

Liaw A, Wiener M, 2002. Classification and regression by randomForest. *R News*, 2(3): 18–22.

Nembrini S, 2018. Bias in the intervention in prediction measure in random forests: illustrations and recommendations. *Bioinformatics*, doi: 10.1093/bioinformatics/bty959.

Pal M, 2005. Random forest classifier for remote sensing classification. *International Journal of Remote Sensing*, 26(1): 217–222.

Reis I, Baron D, Shahaf S, 2019. Probabilistic random forest: a machine learning algorithm for noisy data sets. *The Astronomical Journal*, 10.3847/1538-3881/aaf101.

Rodriguez-Galiano V, Sanchez-Castillo M, Chica-Olmo M, Chica-Rivas M, 2015. Machine learning predictive models for mineral prospectivity: an evaluation of neural networks, random forest, regression trees and support vector machines. *Ore Geology Reviews*, 71(12): 804–818.

Strobl C, Boulesteix AL, Zeileis A, Hothorn T, 2007. Bias in random forest variable importance measures: illustrations, sources and a solution. *BMC Bioinformatics*, 8: 25.

Verikas A, Gelzinis A, Bacauskiene M, 2011. Mining data with random forests: a survey and results of new tests. *Pattern Recognition*, 44(2): 330–349.

Wager S, Hastie T, Efron B, 2014. Confidence Intervals for random forests: the jackknife and the infinitesimal jackknife. *Journal of Machine Learning Research*, 15: 1625–1651.

Winham S, Wang X, de Andrade M, Freimuth R, Colby C, Huebner C, Biernacka J, 2012. Interaction detection with random forests in high-dimensional data. *Genetic Epidemiology*, 36: 142.

Zhao X, Wu Y, Lee DY, Cui W, 2019. iForest: interpreting random forests via visual analytics. *IEEE Transactions on Visualization and Computer Graphics*, 25: 407–416.