BIOMOD 操作指南

Wilfried Thuiller

Bruno Lafourcade, Miguel Araujo

张金龙 译

2009 年 6 月 8 日

译者序

BIOMOD 是 R 软件的一个程序包，是 BIOdiversity MODelling 的缩写，主要用来模型物种分布区及其随气候的变化。该软件是 Wilfried Thuiller 等三位学者 BIOMODE 是在法国蒙彼利埃 CNRS 功能与进化生态学中心开发的。W. Thuiller 是活跃在学术一线的年轻科学家，其发表的学术论文具有较大影响。

BIOMOD 程序包可以运行以下模型，对物种潜在分布区进行估算，同时也能够对相应的气候变化的情形下，物种分布区的变化给出预测。预测过程中可以设置物种迁移速率，从而更接近真实结果。

Generalised Linear Models (GLM)

Generalised Additive Models (GAM)

Classication Tree Analysis (CTA)

Articial Neural Networks (ANN)

Surface Range Envelope (SRE)

Generalised Boosting Model (GBM)

Breiman and Cutler's random forest for classication and regression (randomForest)

Mixture Discriminant Analysis (MDA)

Multiple Adaptive Regression Splines (MARS)

希望本文档对刚刚开始使用 BIOMOD 的程序包的人员有所帮助。

由于译者水平有限，在翻译过程中难免有不妥的地方，还请读者及时批评指正，若有疑问可发邮件至 jinlongzhang01@gmail.com 。

译 者

2010 年 1 月

于北京 香山

目 录

0.1 引言

BIOMOD 是 BIOdiversity MODelling的缩写。BIOMODE 是在法国蒙彼利埃 CNRS 功能与进化生态学中心开发的，得到了欧盟 FP5 AREAM 项目的部分资助。本程序包是为了模拟物种分布而开发的，但可以用来对任何分布进行模拟。惟一的限制在于变量应该编码为 0-1 二元形式。

BIOMOD 是预测物种分布的综合平台，可以详细的给出模型的不确定性，探讨物种-环境间的关系。其中包括了几种模拟物种分布的技术，利用多种方法对模型进行检验，预测物种在未来不同气候条件及迁移时的分布，估算当前物种的空间替换，绘制物种响应曲线，检验物种与预测变量的关系等。从计算的观点来讲，BIOMOD 是 R 函数的汇总，可以用独立变量将多种模型应用到变量中。

0.2 安装

安装 BIOMOD 需要安装最新版本的 R 软件。在运行 BIOMOD 之前，需要安装若干程序包，包括：rpart, MASS, gbm, gam, nnet, mda, randomForest, Design, Hmisc, reshape, plyr。

从 2009 年 3 月，BIOMOD 的函数与之前的保存方式有所不同。现在，BIOMOD 是一个 R 的程序包，可以在 http://r-forge.r-project.org/R/?group id=302 下载。

其中包括 BIOMOD 运行所需的所有函数，及相应数据。所有的函数只需键入其相应的名称即能查看其源代码。如果是新用户，则无需为此担忧，而熟悉程序的用于则可以重新编写函数，如果愿意，还可以更改内部参数，但是其风险要自己承担，因为很多函数与其他函数存在着关联。

解压缩后应该将其放入 R 的程序包文件夹，例如：C:/Program Files/R/R-2.8.0/library。程序包文件夹完全取决于 R 安装的位置及版本。

另一个名为 BIOMOD-R User Functions 的文件旨在帮助用户更好的运行 BIOMOD 软件。该脚本涵盖了数据格式准备，BIOMOD 初始化，及不同模型的运算。在任何情况下都推荐运行该脚本。你也可以根据自己的需要进行修改。

推荐的步骤，首先是创建一个称为 BIOMOD 的文件夹。之后，创建一个用来存储数据的文件夹，运行模型，保存输出结果。也就是称为工作空间“Workspace”。在下面的例子中，工作空间是“Biomod_runs”文件夹。

文件的读取和写入都将在该文件夹中进行。例如，需要将 R 函数拷贝到工作空间，以保证可以打开它们。

在较新版本的 BIOMOD 中，考虑软件到内存限制，结果将被保存在工作空间之外。 在运行 BIOMOD 时，要意识到需要创建额外的文件夹。首先，`Models()`函数将创建 `models` 和 `pred` 两个文件夹。可以想象，它们分别表示模型和预测结果。之后，`Projection()` 函数将为每一种预测的结果进行存储(参见 `Models' Projection section`)。

0.3 开始

0.3.1 数据准备

为了促进 BIOMOD 的学习，使用指南给出了模拟数据。建议用户跟随指南上的每一步运行模型对虚拟数据进行计算。通过对指南的学习，用户应该可以用自己的数据在 BIOMOD 中运行。

最开始的一步是载入 BIOMOD 包。它将自动将需要的程序包载入。

命令为 `library(BIOMOD)`

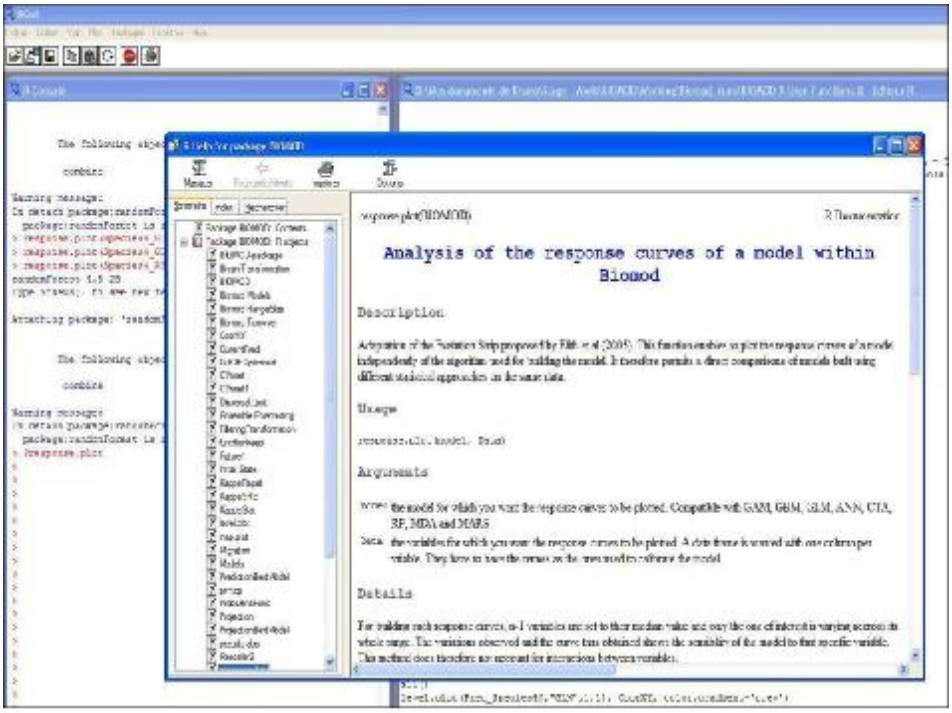
若出现 `Error in library(BIOMOD) : no package named 'BIOMOD' has been found.`

则需要安装程序包。

如果没有问题，就可以应用 BIOMOD 中的不同函数了。可以在函数名前面加一个问号"?"获取相应函数的帮助，了解该函数的用途，每个参数的解释及设定还有若干示例。一般来说，示例比本文档中的例子要更为详细。

例如，命令：

```
?response.plot
```



R 软件的内存中的信息可以保存为一个工作空间。在开始一项新的工作时，可以载入先前的工作空间，将导入所有的函数，对象，结果等，可以继续进行下一步分析。

命令为

```
save.image()
```

这个命令将在工作目录生成一个扩展名为.RData 的文件。如果按照上边的函数运行，

生成文件的实际名称为 RData，但是也可以起一个名字，如 save.image("mywork.RData")。

现在，BIOMOD 可以运行了，用户可以导入物种和环境数据。为了便于操作，我们将其存入同一个文件，载入 BIOMOD 提供的物种/环境数据。

从记事本文档中读取数据时，应用 read.table() 函数：

```
My.Data<-read.table("my_data.txt",h=T,sep="\t")
```

键入?read.table 查询帮助文件，查看相应细节及可能的扩展功能。

间隔可以为制表符，也可以为“,”，但是以“,”作为间隔时，应该用 read.csv 函数以正确读取数据。

BIOMOD 不能识别地理坐标，也不能根据地理坐标进行排序。用户应该确保所有的数据的顺序一致。

注意： BIOMOD 中的数据不能有缺失值

下面的命令可以读取例子

```
data(Sp.Env)
data(CoorXY)
head(Sp.Env)
```

```
  Id     Var1     Var2     Var3     Var4     Var5     Var6     Var7
Species1
1  1 0.668250 4296.144 770.1200 39.32668 295.1067 16.74480 10.86612
0
2  2 0.759600 4173.614 928.1445 57.32333 348.7211 16.40933 10.51443
0
3  3 0.742437 4173.137 870.2771 50.05250 329.9908 16.40781 10.50084
0
4  4 0.554333 4264.083 620.0239 24.98943 239.0667 16.65633 10.92612
0
5  5 0.548943 4168.948 622.3300 25.16287 240.9890 16.39806 11.28096
0
6  6 0.536308 4206.206 591.7892 25.74251 222.9161 16.49461 10.12890
0
  Species2 Species3 Species4 Species5 Species6 Species7
1        0        0        0        1        0        1
2        0        0        0        0        0        1
3        0        0        0        0        0        0
4        0        0        0        1        0        0
5        0        0        0        1        0        0
6        0        0        0        0        0        0
>
- Idw: An Id to keep track of the row numbers
- X and Y: longitude and latitude of our sites (used for plotting)
- Var1 to Var7: Environmental variables (bioclimatic in that case)
- Sp281 to Sp191: Presence/absence of 8 species.
```

0.3.2 初始化

需要争取的读取数据。

函数为 Initial.State()

初始化的目的是表示哪些列是物种分布变量，哪些是环境变量等等。

其中包括

Response，即物种分布

Explanatory，即环境变量

IndependentResponse: 完全独立响应变量，用来对模型的可靠性进行检验。如果没有完全独立的数据，则应该设定为 NULL,

IndependentExplanatory： 完全独立因变量，用来对模型预测的准确性进行评估。如果没有完全独立的数据，则应该设定为 NULL

校准的步骤

理想情况下，需要用独立数据检验物种分布预测的结果。如果有这样的数据，BIOMOD 会利用独立数据校准模型。如果没有独立数据用来对模型进行检验，可以用两种方法评估模型（参见下文中 Models()函数的 NbRunEcal 选项）。

这个例子中，我们没有真正的独立数据，但是我将给出假独立数据，现在暂时忽略这一问题。

```
Initial.State(Response = Sp.Env[,c(11,13,14)], Explanatory = Sp.Env[,4:10],
IndependentResponse = Sp.Env[,c(11,13,14)], IndependentExplanatory = Sp.Env[,4:10])
ls()
```

Initial.State()函数生成两个对象，DataBIOMOD 和 DataEvalBIOMOD，后者用来对模型进行检验，特别要注意，不要更改这个对象，也不要删除它。

```
head(DataBIOMOD)
```

DataBIOMOD 开始的几列是环境变量，后边是物种出现与否的信息。DataEvalBIOMOD 的结构相同，该数据用来检验模型。

另外一个对象称为 Biomod.material，其中包含非常重要的信息，有变量数目，种数。同样要保证不要更改它。

0.4 模型

0.4.1 简要说明

Model 函数用来运行 BIOMOD 中的模型，同时包括三种检验模型的方法（kappa, True Skill Statistics, ROC 曲线）

当前可以运行 9 种模型

包括

Generalised Linear Models (GLM)

Generalised Additive Models (GAM)

Classication Tree Analysis (CTA)

Articial Neural Networks (ANN)

Surface Range Envelope (SRE)

Generalised Boosting Model (GBM)

Breiman and Cutler's random forest for classication and regression (randomForest)

Mixture Discriminant Analysis (MDA)

Multiple Adaptive Regression Splines (MARS)

模型选取的时候键入参数 T 或 F。有些模型需要设定其他的参数。参见下文。

对校正数据的每个种，已经选择的模型都将运行。下面是每个模型的简要说明和函数参数的解释。注意，解释的顺序与其在 Model 函数中出现的顺序不同。更为详尽的解释在本手册的末尾可以找到（参见模型描述一节）。

模型选择及参数

- GLM = T, TypeGLM = "poly", Test = "BIC": 分步 GLM，可以选择线性"simple"，二项式"quad"或多项式"poly"。参数选择采用 AIC 或 BIC。

- GBM = T, No.trees = 3000, CV.gbm = 5 : 运行 generalised boosting model(GBM)（boosted regression trees）。可以自己设定树的数目（默认为 3000）。用 cross-validation procedure 选择树的数目。cross-validation 的默认次数是 5.

- GAM = T, Spline = 4 : 运行广义加法模型（GAM）with a spline function with a degree of smoothing of 4 (similar to a polynomial of degree 3).

- CTA = T, CV.tree = 50 : 进行分类树分析（CTA）.树的最佳长度用 cross-calidation 进行估算（默认为 50）。

- ANN = T, CV.ann = 2 : 运行人工神经网络（ANN）.由于运行结果在不同的次数不同，weight decay 最优量，及 hidden layer 单元的数量借助 N-fold-validation（默认为 3 次）进行选择。用户也可以选择 cross-validation 的次数。

- SRE = T, Perc025=T, Perc05=F: 运行 rectilinear surace range enveloe (BIOCLIM), 用 Nix 和 Busby 推荐的百分数 0.025 或 0.05

- MDA = T: 用 MARS 函数进行模型中回归部分的混合判别分析（mixture discriminant analysis）

- MARS = T: 运行 multivariate adaptive regression spline

- RF = T: 运行 random forest model.

模型的评估

- ROC = T: 利用 AUC 曲线 Area under the ROC (receiver operating characteristic curve) Curves 评估模型

- Optimized.Threshold.ROC = T: ROC 是一种临界独立方法。但如果用户想找到最优化的 Presence、absence 正确预测结果，这一临界值可以将模型中的物种出现概率转换为 Presence、absence.

- Kappa = T: 利用 Cohen's Kappa 评价模型。保留了最优化 Kappa 的临界值。

- TSS = T: 利用 True Skill Statistics (TSS) 评价模型。保留了最优化 TSS 的临界值

- VarImport: 如果选项为真（TRUE），该参数将对不同模型的解释变量进行直接比较。模型经过训练之后，可以进行标准预测。之后，其中一个变量被随机化，并重新进行预

测。计算新预测的结果与标准预测结果之间的相关系数，从而对该参数的重要性进行评估：如果相关系数高，例如，两种方法的预测结果没有大的区别，那么被随机化的参数对模型的预测没有太大的影响。这一步将对每一个参数独立运行 n 次。

注意： VarImportance 函数运行的时候，给出的结果是 1 减去每个变量得出的相关系数的平均值。因此，分值越高表明重要性就越高。结果也可以作为相对重要性，该值已经不是相关系数，而表示的是变量的重要程度。

假缺失的使用

- NbRepPA = 0 : This will set the use of a pseudo-absences selection if higher than 0. 请仔细参考 Pseudo-absences 一节。这一步可以重复多次，这将使每个模型运行的次数大大增加。

- strategy = 'random': 用来选择 pseudo absences 的类型。可以设定为 "circles","squares", "per", "random" or "sre".

- coor = CoorXY： 具有两列数据的矩阵给出数据点的坐标。在选择"per", "circles" 和"squares"时需要用到。

- distance = 3 : 在选择"per", "circles" 和"squares" 时给出距离。

- nb.absences = 2000 : 在模型中需要运行的 pseudo absence 的次数。 根据给出的不同方式，随机在 pseudo absences 中抽取。

确定模型评价的一般步骤及运行的次数

以下的选项的值将决定模型建立的方式及检验。需要特别注意。

- DataSplit: 将原始数据划分为上文提到的校准数据和评估数据。注意：该函数确保出现数据点中用于校准或评估用的数据比例的保持恒定。

- NbRunEval : 如果没有用于评估模型的独立数据，可以用两种选择。

首先，随机将数据划分成两部分，70%和 30%(参见 Araujo et al. 2005b, Guisan and Thuiller 2005),70%的数据用于模型的校准，30%的数据用于模型的检验。

其次，也可以选择 multiple cross-validation 方法，BIOMOD 随机将数据划分 N 次，运行模型，记录预测的结果，给出 cross-calidation 的平均值。这种方法将为选择的每个模型的预测结果给出更为稳健的估计，同时给出模型对初始条件敏感性的估计，例如，物种分布资料。当然，在 PC 机上所需运行的时间也更长。

校准和评估模型的新方法： 为了使结果可信，预测必须使用独立的数据。由于这一数据常常不可用，代之以将数据随机划分为校准组和评估组的方法。校准数据用在模型的拟合或者学习中，而评估数据用于用于对预测的准确性进行评估。在 BIOMOD 中，模型的评估可以用三种方法： Kappa, ROC, TSS。 但是，这种经典的数据划分方法在多次运行模型的时候会产生不稳定性： 这是由于数据的划分是随机的，每一次运行的校准数据和评估数据都可能不相同，这就不可避免的使每次校准的模型及其相应的预测结果有所不同。为了解决这一问题，BIOMOD 允许用户利用不同的数据划分对模型的表现进行评估，之后用 100%的数据来对预测进行校准。此时，评估更为可信，而预测结果则不会受到数据随机分组的影响。但是，执行所有的评估比经典的数据进行一次有偏差的分组需要更长的运算时间。如果有独立的数据，也可以对再次对最终模型进行检验。

- Yweights: 设定响应变量的权重（一个具有 N 列数据，即 N 种分布数据的矩阵）。这类似于每个点上物种的可见程度，对于更为可信的出现不出现数据给出更大的权重。可以划分响应的等级，在模型运算中作为权重。要了解更多信息，参见 weights 在 R 中怎样工作。

- KeepindependentPred: 如果设定为真，（同时给出真正的独立数据），则独立数据的预测也将被保存。如果为假，则仅保存独立数据预测的准确性（ROC，TSS，Kappa）

0.4.2 模型的运行

现在可以针对我们的种运行不同的模型。每个模型的运行将要花费一段时间。这里，我们将要对每个种运行 9(个模型)*4(3 次重复+最终模型)*2（PA 重复），总共是 72 个模型，这将花费几分钟的时间。

如果你的计算机仅仅是简单的个人计算机例如（如 laptops），特别是数据有几万行的时候，也许需要执行一部分模型，以减少运算时间。通过将参数设定为真，可以一次运行所有模型。

注意 与以前版本的 BIOMOD 相比，最好不要一次运行一个模型，因为每个种的预测结果现在分别存储。运行几次模型将会给输出数据的分析带来麻烦。

应该注意的还有，NbRunEval 和 BbRepPA 两个选项，在增加每个种运行次数的同时将延长运算时间。如果没有足够的耐心或性能良好的计算机，不要额外调高这两个值。

输入如下命令：

```
Models(GLM = T, TypeGLM = "poly", Test = "AIC", GBM = T, No.trees
= 2000, GAM = T,
Spline = 3, CTA = T, CV.tree = 50, ANN = T, CV.ann = 2, SRE = T,
Perc025=T, Perc05=F, MDA = T,
MARS = T, RF = T, NbRunEval = 3, DataSplit = 80, Yweights=NULL, Roc
= T, Optimized.Threshold.Roc = T,
Kappa = T, TSS=T, KeepPredIndependent = T, VarImport=5, NbRepPA=2,
strategy="circles",
coor=CoorXY, distance=2, nb.absences=1000)
```

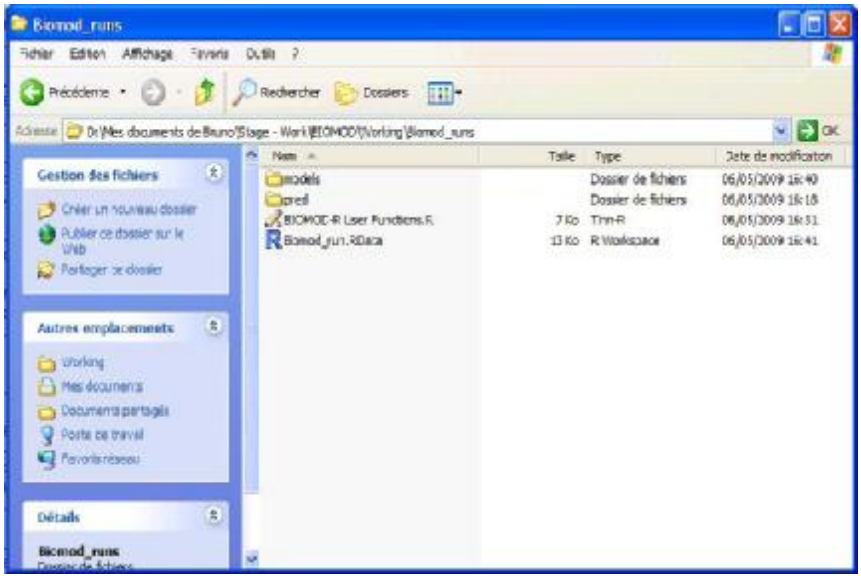
为了举例说明，我们在例子中这里用到了 2 个 pseudo absence。注意，对于 Sp277，与需要的数量相比，absence 的数据太少，所以只运行了一次 PA。nb.absences 参数设定为 1000

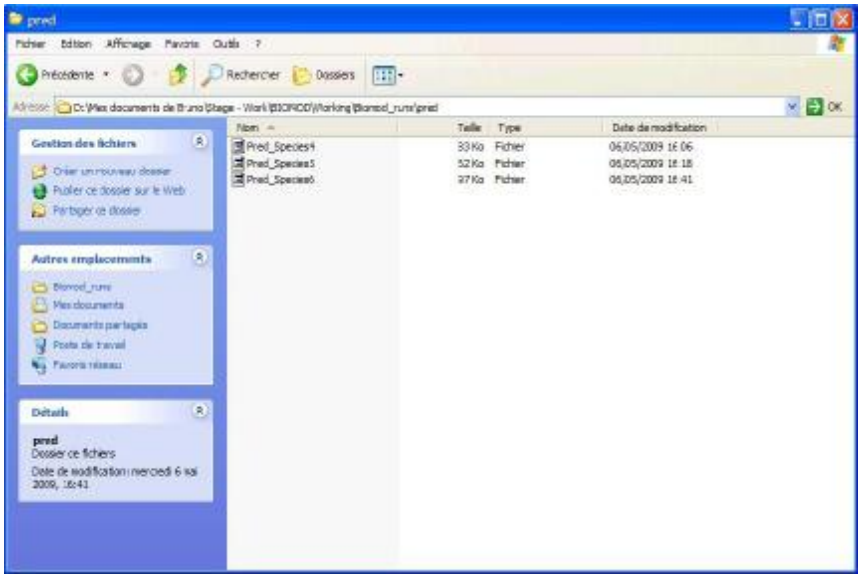
但是

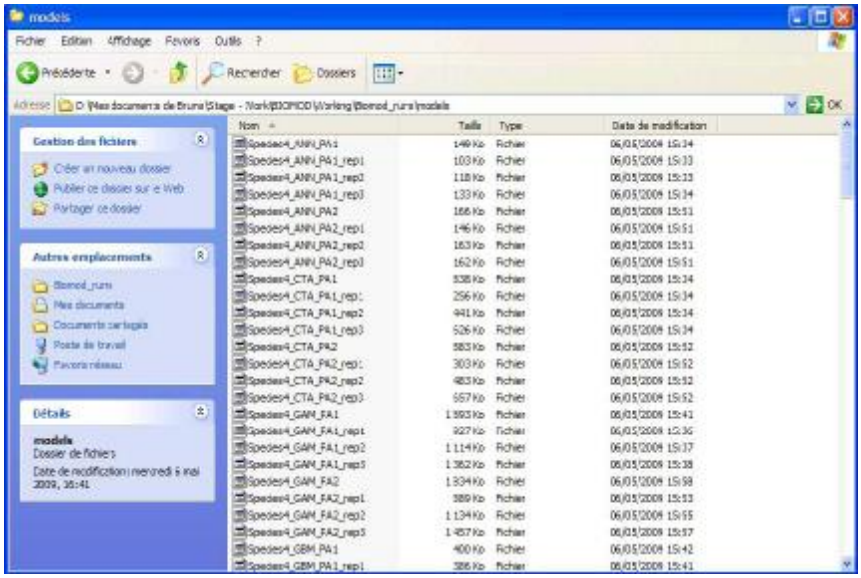
```
> #the number of data selected by the pseudo-absences procedure
> length(Biomod.PA.data$Sp277)
[1] 1791
> #the number of presences for Sp277
> sum(Sp.Env[,"Sp277"])
[1] 1080
> #Hence, the number of absences available for calibration
> length(Biomod.PA.data$Sp281) - sum(Sp.Env[,"Sp277"])
[1] 679
```

absence 数据量太少。这种情况下，pseudo-absences 运行一次。

此时，你的工作文件夹应该如下所示：


Biomod_runs
Fichier Edition Affichage Favoris Outils ?
Précédente
Rechercher
Dossiers
OK
Gestion des fichiers
Créer un nouveau dossier
Publier ce dossier sur le Web
Partager ce dossier
Autres emplacements
Working
Mes documents
Documents partagés
Poste de travail
Favoris réseau
Détails
Biomod_runs
Nom
Taille
Type
Date de modification
models
Dossier de fichiers
06/05/2009 16:40
pred
Dossier de fichiers
06/05/2009 16:18
BIOMOD-R User Functions.R
7 Ko
Tinn-R
06/05/2009 16:51
Biomod_run1.RData
13 Ko
R Workspace
06/05/2009 16:41


pred
Fichier Edition Affichage Favoris Outils ?
Précédente
Rechercher
Dossiers
OK
Gestion des fichiers
Autres emplacements
Biomod_runs
Mes documents
Documents partagés
Poste de travail
Favoris réseau
Détails
pred
Dossier de fichiers
Date de modification : mercredi 6 mai 2009, 16:41
Nom
Taille
Type
Date de modification
Pred_Species4
33 Ko
Fichier
06/05/2009 16:06
Pred_Species5
52 Ko
Fichier
06/05/2009 16:18
Pred_Species6
97 Ko
Fichier
06/05/2009 16:41



0.4.3 结果分析

在工作空间中储存了若干对象

```
ls()
```

其中一些为运行的结果，如 Evaluation.results 或 the VarImportance objects，另外一些在内存中有一定的用处，还有一些数据框 (Sp.Env, DataBIOMOD)

除了 SRE 模型之外，每个模型都将生成一个储存多种参数的对象，包括每个变量的重要值（GBM,GAM,和 randomForest）,对变量显著性的方差分析（GLM，GAM）等等。这些输出结果对于进行预测是十分重要的，可让人们知晓模型选择了哪些变量。

模型本身现在从 R 工作空间提取出来，直接存储在用户的硬盘上。以物种模型的方式命名。如 Sp164_MDA。在重复或者 pseudo-absence 的运行结果，添加相应的后缀，所以会有 Sp164 MDA PA1 rep2 的结果。

重新载入模型非常方便，使用 load()函数，可以输入相应的路径直接载入 R 工作空间中存储的模型。同样，也可以用该函数调用其他 R 以外的输出结果（predictions and projections）. 这里举例查看一下 GLM 的模拟结果（也许命令行不太方便，需要熟悉一段时间）。

```
> load("models/Sp277_GLM_PA1")
> Sp277_GLM_PA1
```

结果显示最终模型中保留的变量。输出结果页给出了不同的系数，自由度，残差，及最终模型的 AIC 值。当然，每个模型给出的信息都是不同的，这与模型本身有关。每个模型的描述如下（cf. OUPUTS 和 INTERPRETATION)）

此时，我们拥有的输出结果包括，Initial.State()函数的输出：

- Sp.Env
- CoorXY
- DataBIOMOD
- Biomod.material

也有 Model()函数的输出结果，Model 的输出结果也保存在硬盘上，包括：

- Evaluation.results.Roc
- Evaluation.results.Kappa
- Evaluation.results.TSS
- VarImportance
- Models.information.

前三个对象包含模型评估过称的得分，及每个模型和每个种的区别。VarImportance 主要是解释性的，包括变量贡献量分析。用户对 Models.information 可能缺乏兴趣，它包括用来更改模型中情景的信息。

如果 NbRepPA 大于 0，我们还将得到

- Biomod.PA.data
- Biomod.PA.sample
- SpNoName.circles.2 (or something close)

Biomod.PA.data 包含了 pseudo-absence 函数运行后，内部数据的可应用性。Biomod.PA.sample 包含从 DataBIOMOD 中获取的每个种的校准数据和每个 PA 数据。最后一个对象是 pseudo-absence 的运行结果，在此没有重要的作用（参见 Pseudo-absence 一节）。

对原始数据的预测每个种单独存储在一个对象中，以 Pred.Speciesname 的格式命名，其中包每次模型运行的物种出现概率（生境适宜度指数）

注意： 为了运算和节省内存的需要，该值设定为 0 到 1000.为了获得真实的出现概率，需要将数据标准化到 0，1 之间，即除以 1000.

对于独立数据，也将产生同样的对象，结果也将以同样的方式展示。

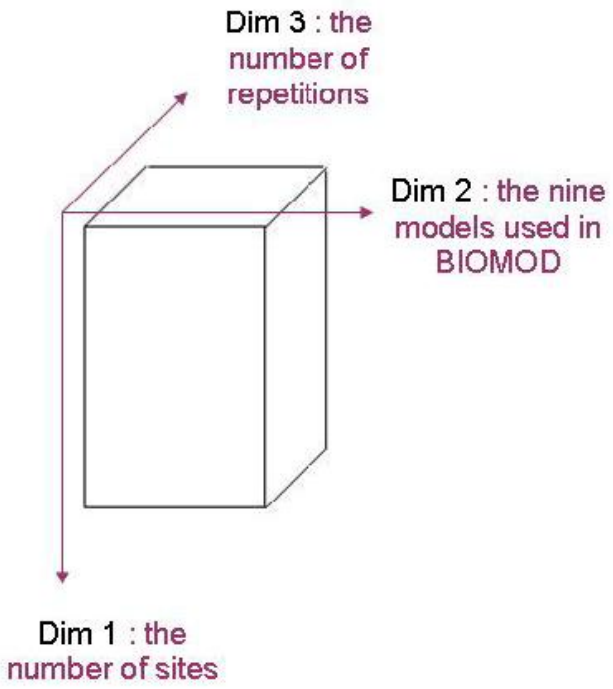
```
load("pred/Pred_Sp277")
```

让人迷惑的是，这些对象已经不是矩阵，而是多维数组（4 维）。这四个维度可以按照下面的方法进行可视化。

前两个维度每一列是一个模型的预测结果，组成一个矩阵。行数与建立模型的数据相对应。

```
Pred_Sp277[1:20,,1,1]
```

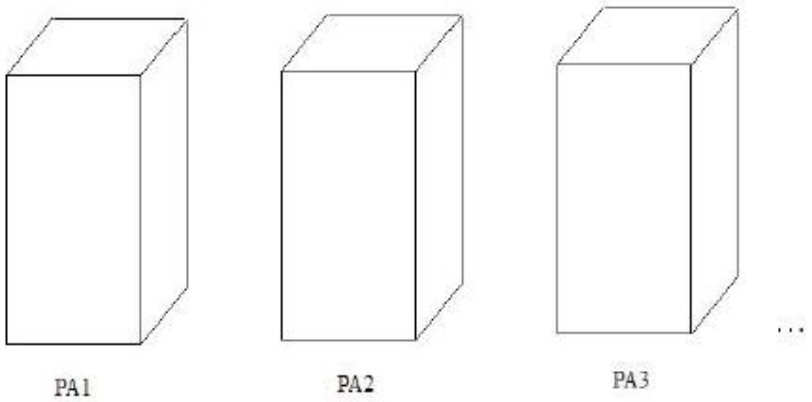
第三个维度由 2D 矩阵组成，一个挨一个，与每次重复次数相连。该维度最小为 1。BIOMOD 总是生成 100%数据校正后的最终模型，则第三个维度的长度是 NbRunEval 参数+1。例如，NbRunEval=1o，将生成 11 个层。



需要记住的是，第一层总是最终模型。

```
> #the final model
> Pred_Sp277[1:20,,1,1]
> #the first repetition model
> Pred_Sp277[1:20,,2,1]
> #the second repetition model
> Pred_Sp277[1:20,,3,1]
```

第四维显示的是 pseudo absence 重复的次数。在 NbRepPA=0 时，维度为 1（不是 0）。



虽然你永远也不可能用 R 看到它。这里只是它怎样排列的一个抽象观察。有用的函数包括 dim(), dimnames(). 前者给出每个维度的层数，后者给出他们的名称。

例如

```
> load("pred/Pred_Sp281")
> dim(Pred_Sp281)
> dimnames(Pred_Sp281)
> #you can avoid having the rownames to be printed in the console
as they
> #are generally not very usefull
> dimnames(Pred_Sp281)[-1]
```

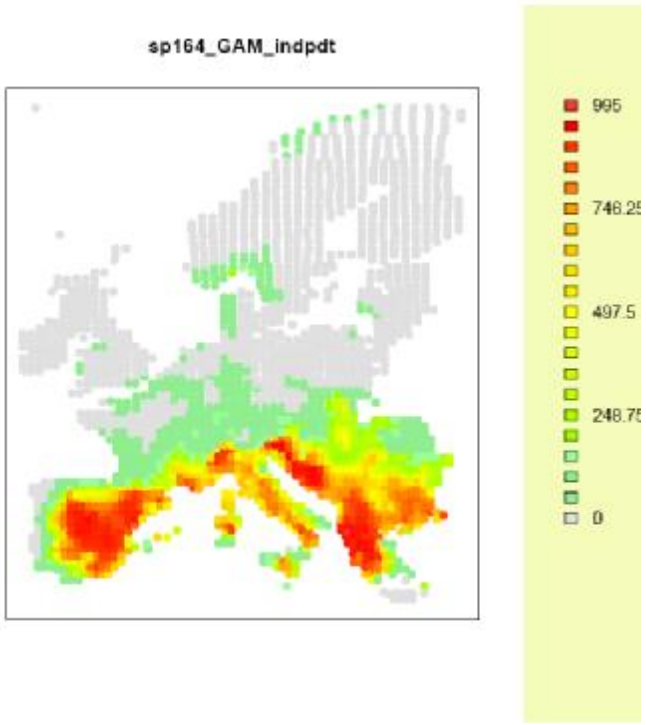
举例来说，我们要查看第一种用 GBM 模型预测的出现的概率。这里，我们展示了 20 行。

```
> #if you don't inform the 3rd and 4th dimension (you still need
commas), you will have all of them
> #at once in a matrix.
> load("pred/Pred_Sp281")
> Pred_Sp281[481:500,"GBM",,]
```

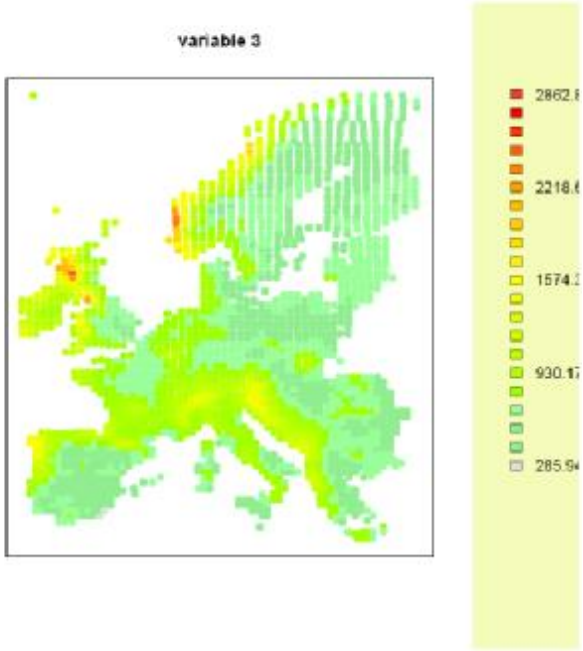
注意，由于 GBM 包含一个随机成分，你的结果与例子中的可能稍有不同。

绘制预测的地图，需要用到 level.plot()函数。 它需要连个变量：你想要绘制的向量的值及数据点的坐标。这一函数对于任何类型的数据都可以处理。由于我们在运行模型中用到了 pseudo-absence 数据，绘制部分预测结果不太方便。而我们将绘制假 GAM 的独立数据（整套数据），以及用来校准模型的一个变量。

```
> load("pred/Pred_Sp164_indpdt")
>       level.plot(Pred_Sp164_indpdt[,"GAM",1,1],       CoorXY,
title='sp164_GAM_indpdt')
```




```
> #and the level plot for the third variable used
> level.plot(Sp.Env[,6], CoorXY, title='variable 3')
```

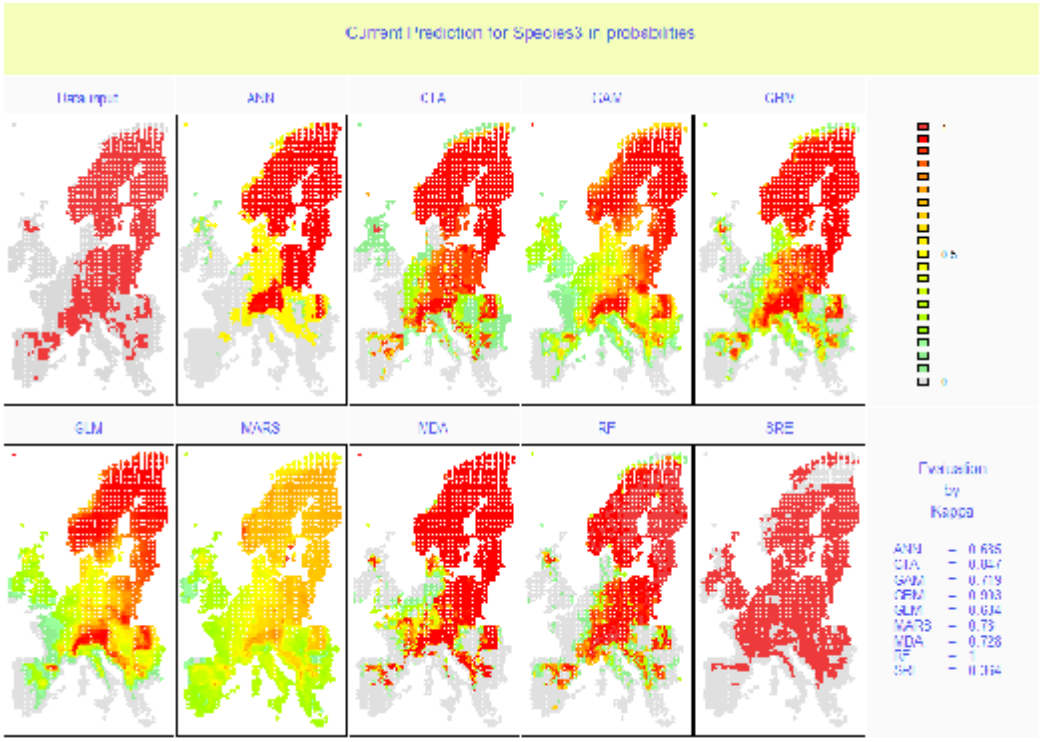


注意，独立预测只针对最后100%模型，而不是针对重复。可以用以下命令查看：

```
Pred_Sp164_indpdt[1:10,,,]
```

level.plot 函数的升级版本是 map.plot 函数。它是专门为 BIOMOD 设计的，可以让不同模型的输出结果同时输出，便于比较。需要输入的是选择哪种模型及想要输出的物种，同时也包括输入数据的格式，及用哪种方法进行输出（例如，是否为 01 数据，是否需要设定临界值等）。例如，此处我们对第一种预测的所有模型给出的概率的原始数据绘制了地图。

```
> #this example was made on prior versions, the function is under maintenance
> map.plot(Sp=1, model='all', method='Kappa', format.type='probs', wanted='prediction')
```



通过设定 color.gradient 选项，可以更改颜色梯度为红（默认），蓝或灰色。

0.5 输出与解释

0.5.1 GLM 的解释和用法

GLM 模型包含的各种成分取决于模型建立过程中的选项，即 GLM 对象。这里是对校准过程的备忘。为通过逐步选项选取的非独立变量及残差和零模型给出解释。

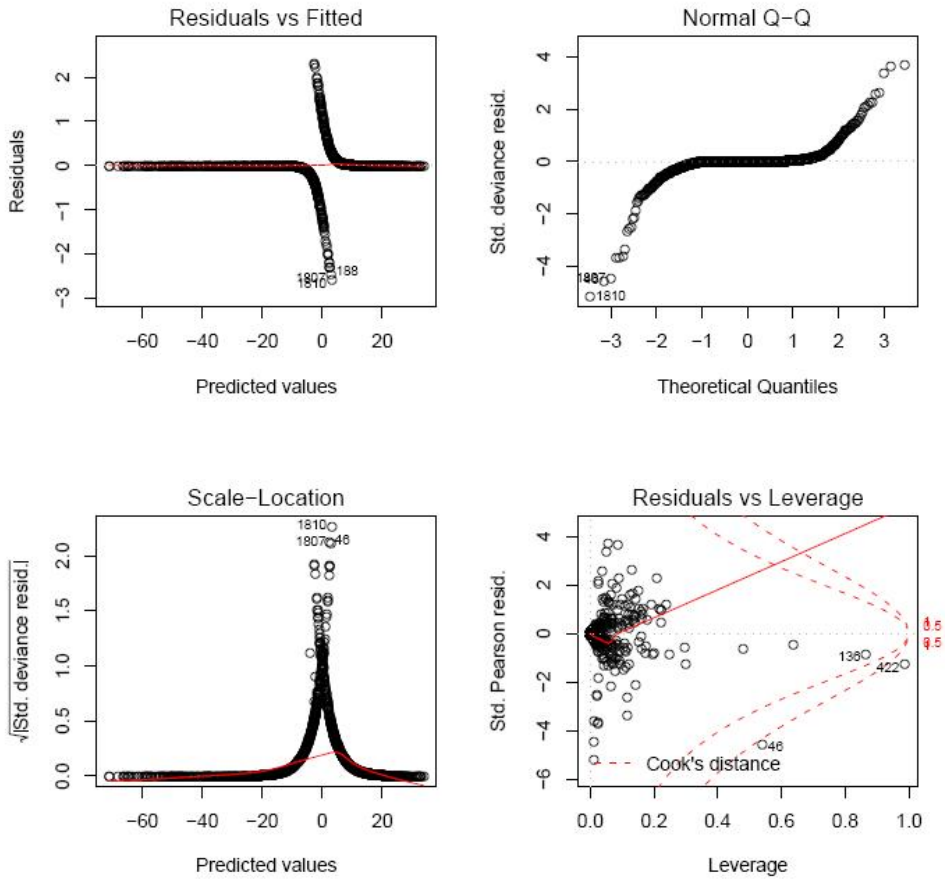
```
load("models/Sp277_GLM_PA1")
Sp277_GLM_PA1
summary(Sp277_GLM_PA1)
```

接下来的程序输出包括方差分析结果及逐步过程的细节。独立变量按照 AIC 重要值排序。

```
Sp277_GLM_PA1$anova
```

R 的 plot()函数将给出 GLM 的基础和常规输出。这些结果也是重要的，但是在逻辑斯蒂回归时关系不大。

```
par(mfrow=c(2,2))
plot(Sp277_GLM_PA1)
```



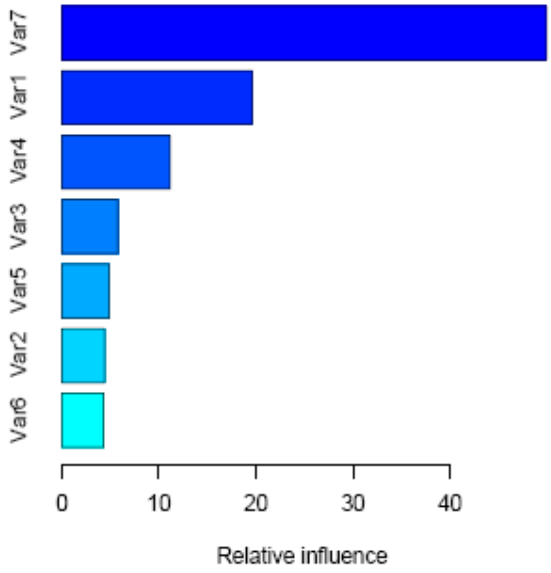
0.5.2 GBM 的解释和用法

GBM 和其他模型的输出结果与 GLM 类似。这里我们只讲一下 GBM 的输出结果。

summary 函数将算出 gbm 对象中每个变量的相对影响力。返回值每个变量简化的属性值，该值表示每次重复中预测过程中梯度的误差平方和。它描述的是 loss 函数简化过程中每个变量的影响力。返回值为数据框，第一项是变量名称，第二项是计算出的相对影响力，是标准化到 100 的值。

需要确保 GBM 程序包已经被载入了。

```
> load("models/Sp281_GBM_PA1")
> summary(Sp281_GBM_PA1)
```

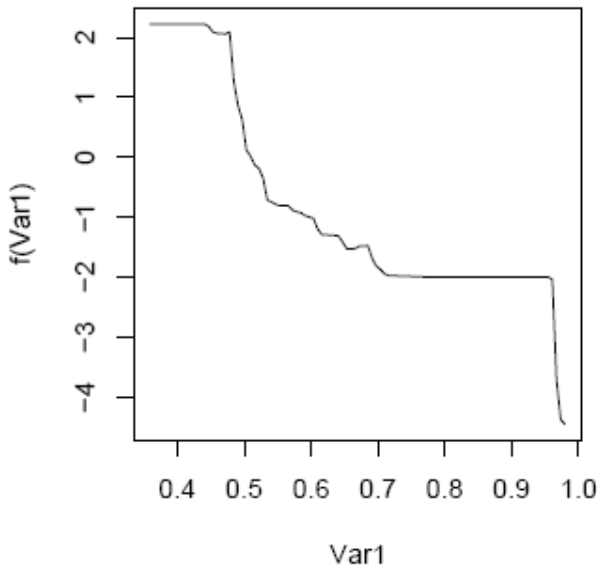


响应曲线

GBM 程序包允许用户绘制模型中物种相对于环境的响应曲线。

i.var: 包含变量系数和名称的向量，用来绘图。如果用系数，变量系数出现的次序与开始 gbm 公式中出现的顺序是一致的。例如，这里 BIOMOD 将绘制模型中的第一个变量。

```
> plot(Sp281_GBM_PA1, i.var=1)
```

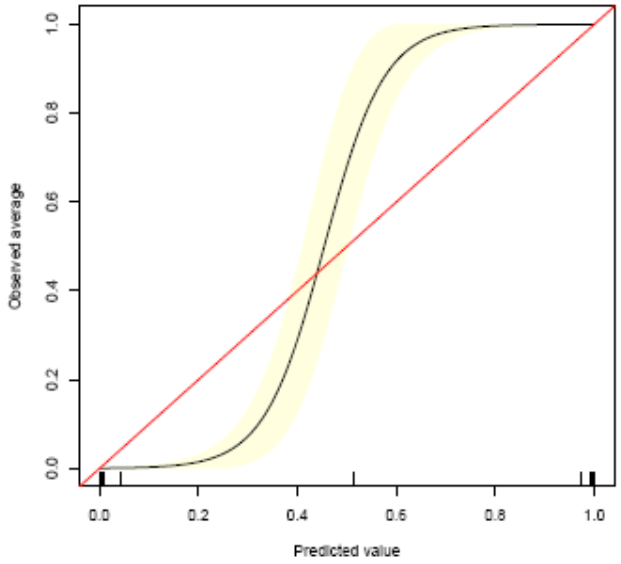


用户也可以用 response.plot 函数绘制 GLM 中的响应曲线。

gbm 包也提供了实验性质的诊断工具，将拟合的数值相对实际平均值作图。利用 gam 来估算 E(y|p)。校正精准的预测意味着 E(y|p)=p。该图也包含 95 点区间估计。

```
> library(gbm)
```

```
> #let's store the data that was used for calibration of the first
PA run
> #for Sp277 to simplify the code
> data.used <- DataBIOMOD[Biomod.PA.sample$Sp277$PA1,"Sp277"]
> calibrate.plot(data.used, Pred_Sp277[,"GBM",1,1]/1000)
```



该函数需要所选种的实际 presence-absence 数据，以及预测结果。该函数能应用于 R-BIOMOD 的任意模型中。

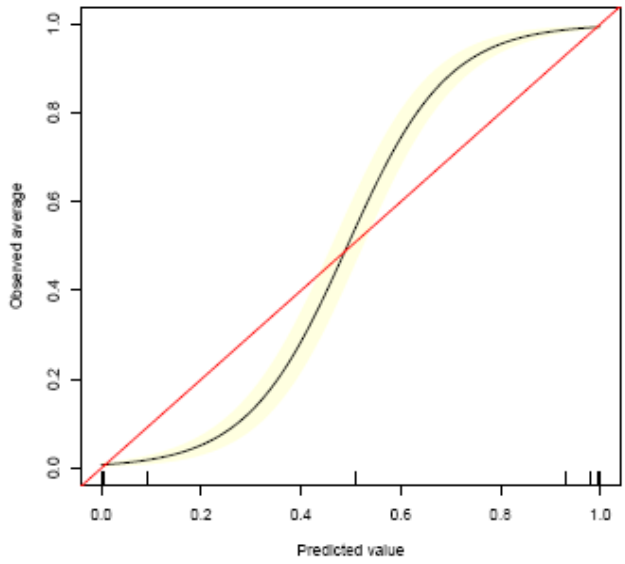
0.5.3 GAM 的解释和用法

输出结果与 GLM 十分相似。

响应曲线可以用内部函数十分方便的绘制出来。

R 中所有的与 GAM 有关的结果都能查看（前提是 GAM 包已经载入）。如 GBM 结果所示，用户可以用 gbm 包中的 calibrate.plot 函数查看模型的精确程度。

```
calibrate.plot(data.used, Pred_Sp277[,"GAM",1,1]/1000)
```



0.5.4 CTA 的解释和用法

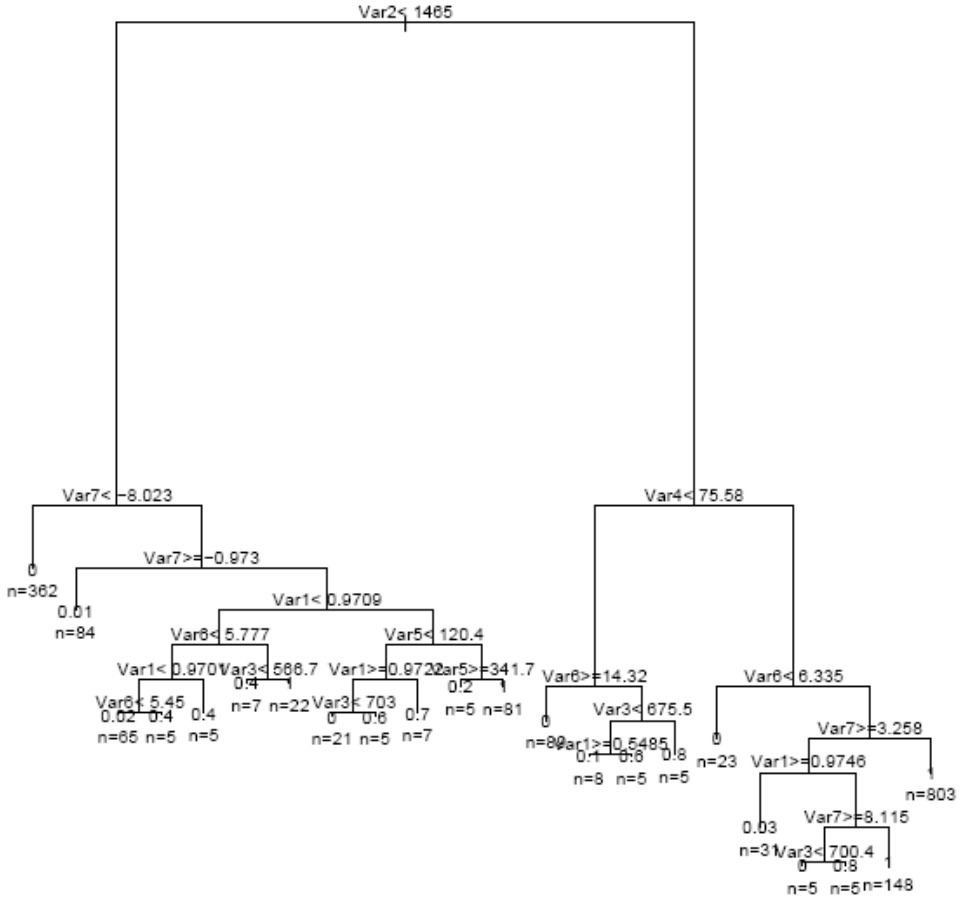
CTA 模型输出结果很有用。最重要的一个名为 frame，给出了节点的细节，每个节点解释的方差及出现的概率。

```
load("models/Sp277_CTA_PA1")
names(Sp277_CTA_PA1)
Sp277_CTA_PA1$frame
```

为了更方便的查看数据，可以同时绘制树图。要保证已经载入了 rpart 包。

注意，plot 函数默认情况下不显示标签和文字，用户必须用 text 函数添加文字。

```
plot(Sp277_CTA_PA1, margin=0.05)
text(Sp277_CTA_PA1, use.n=T)
```



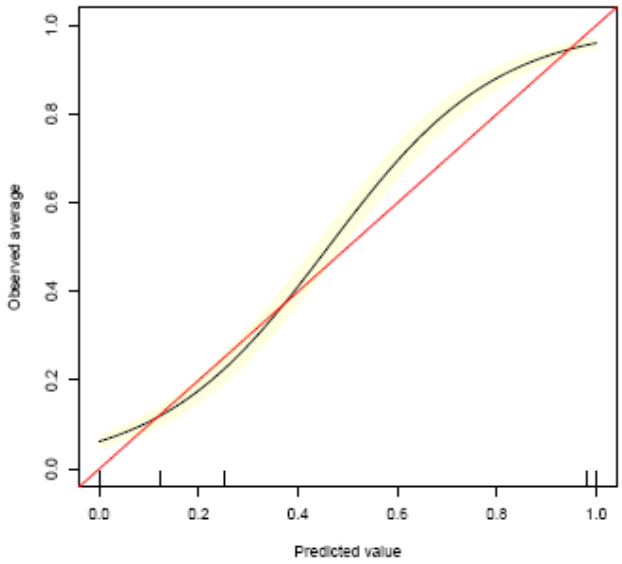
即使对于 CTA，仍然可以绘制响应曲线。这是基于等级的方法，使得关系非常精准。

0.5.5 ANN 的解释和用法

与 GLM，GAM，CTA 相仿，可以用 plot.response 函数绘制所选物种和环境变量的响应曲线。

用户同时可以用 gbm 包的 calibrate.plot 函数绘图，表示模型拟合程度的好坏。

```
> load("models/Sp277_ANN_PA1")
> names(Sp277_ANN_PA1)
calibrate.plot(data.used, Pred_Sp277[,"ANN",1,1]/1000)
```



0.5.6 SRE 的解释和用法

在这里，没有那个模型比这种线性分室模型更简单，与其他模型的结果一样，预测结果也被保存。

SRE 不能给出物种出现的概率，只能给出物种的出现或不出现，因此 ROC 评价曲线也不可用。calibrat.plot 函数也就不能工作了。只有 TSS 和 Kappa 可用。

0.5.7 MDA 的解释和用法

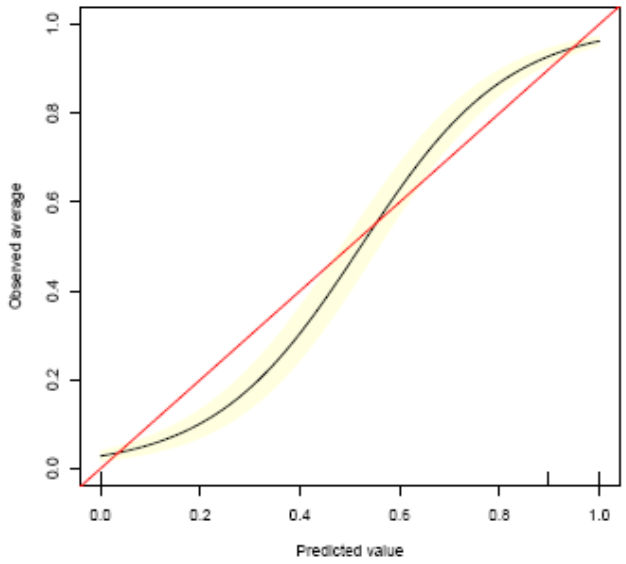
取决于在建立模型的时候的选项，模型可包括以下组分，例如对于第一种来说：

```
> load("models/Sp277_MDA_PA1")
> summary(Sp277_MDA_PA1)
Sp277_MDA_PA1
```

与前面的模型类似，我们也可以用 response.plot 绘制所选物种对于环境变量的响应曲线。

用 gbm 包的 calibrate.plot 函数显示模型拟合程度的好坏。

```
calibrate.plot(data.used, Pred_Sp277[,"MDA",1,1]/1000)
```



0.5.8 MARS 的解释和用法

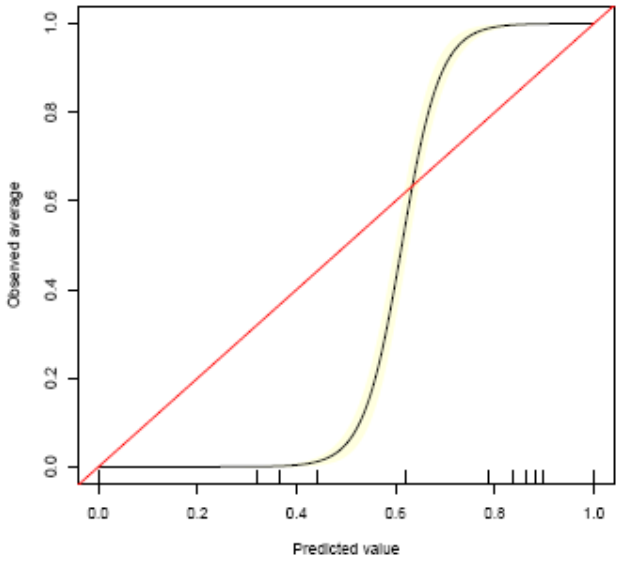
取决于在建立模型的时候的选项，模型可包括以下组分，例如对于第一种来说：

```
> load("models/Sp277_MARS_PA1")
> summary(Sp277_MARS_PA1)
```

与前面的模型类似，我们也可以用 `response.plot` 绘制所选物种对于环境变量的响应曲线。

用 `gbm` 包的 `calibrate.plot` 函数显示模型拟合程度的好坏。

```
calibrate.plot(data.used, (Pred_Sp277[,"MARS",1,1]/1000))
```



0.5.9 RF 的解释和用法

```
> load("models/Sp277_RF_PA1")
> summary(Sp277_RF_PA1)
```

由 random Forest 产生的每个变量，可以提取出来

```
Sp277_RF_PA1$importance
```

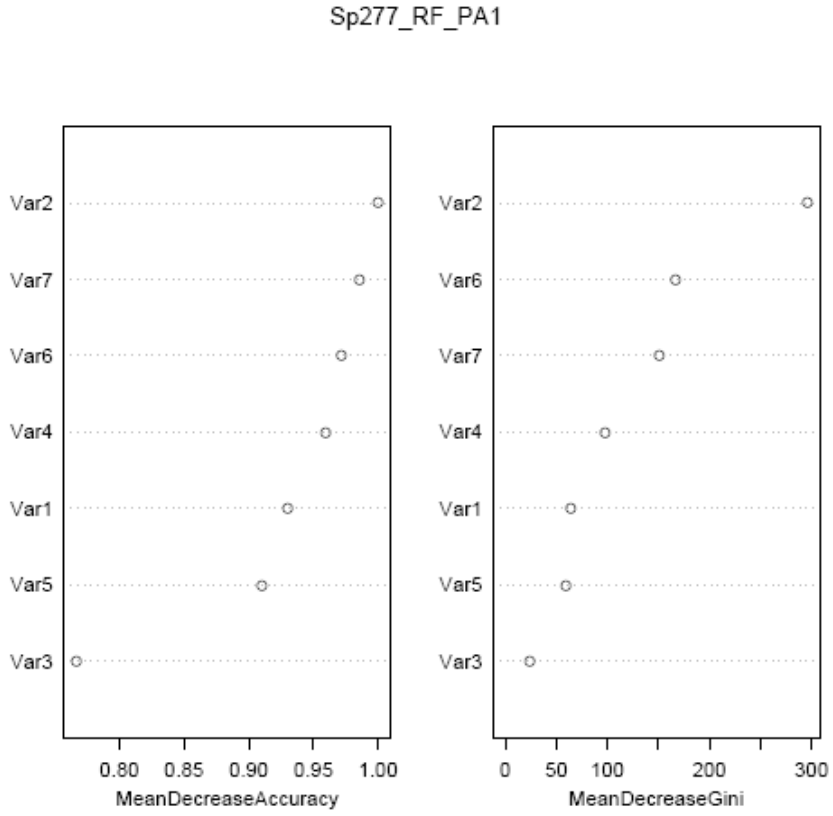
下面是变量重要值的定义

- Mean Decrease Accuracy: 对于每个树，数据离开袋子（out-of-bag）的比例——预测的准确性被记录。对每个预测变量进行同样的变换。计算所有树间两个预测精度的差别，用标准差进行标准化。

- Mean Decrease Gini: 第二个量度是所有树的变量分割的节点不纯性减少的平均值（The second measure is the total decrease in node impurities from splitting on the variable, averaged over all trees. ）。澄清一下，节点的不纯性用 Gini 系数表示。

类似的，随机森林中变量的重要性可以用 dotchart 图表示。

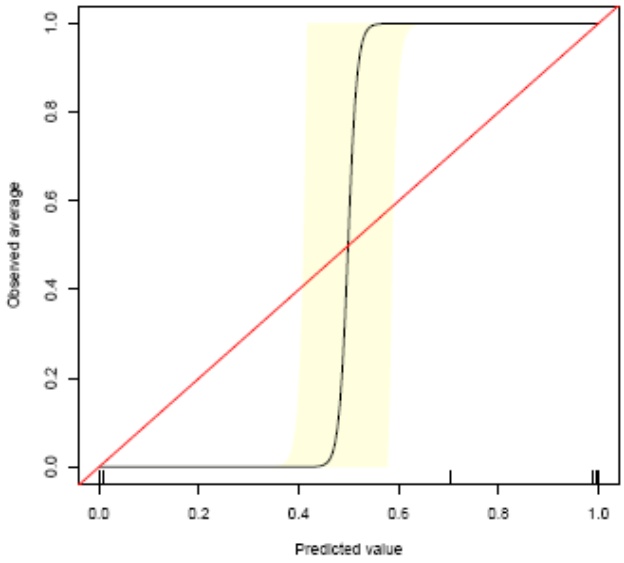
```
varImpPlot(Sp277_RF_PA1)
```



用变换过程（对于每个模型都类似）评价每个变量的重要值，在此通过选项也可以实现。类似的，用 plot.response 可以绘制响应曲线。

可以利用 gbm 包的 calibrate.plot 函数绘制拟合度好坏的图形。

```
calibrate.plot(data.used, (Pred_Sp277[,"RF",1,1]/1000))
```



0.5.10 预测结果的评估

有三种方法可以评估模型的表现（c.f. 0.14 Predictive Performance description）。如果选择了 ROC,Kappa 和/或 TSS,将计算相应的 cross-validation 分步模型,最后用 100%的数据校准模型。每次运行的每个种的预测结果将单独保存。

Model 函数将生成一个汇总表，名为“Evaluation.results.method”,其中包含了方便对于每个方法和分类单元的比较的结果。

```
> #Here we only display the info for the first species modelled
> Evaluation.results.Kappa[1:8]
```

以 PA 为例，这里包含 4 个不同的矩阵，每个是对应于不同的运行次数（3 次重复，80%~20%的数据分割，最终 100%的数据）。第一次重复（Sp277_PA1_rep1），第一列数据是模型校准后其余 20%数据的得分。最后一列是 80%数据用于校准的数据和 20%生于数据的组合。

对于最终模型（Sp277_PA1），第一列是所有重复的平均 cross-calidation。第二列是利用独立数据（如果有的话）对模型评估时的得分。接下来的四列是从最终模型得到的。

可以看到 Sp277 的 PA2 是空矩阵。这是因为该种的 PA 只运行了一次（参见 13 页）。

如果要获得第二种经过 GLM 模型的 ROC 预测精度，则可以用

```
> Evaluation.results.Roc$Sp277_PA1["GLM",]
```

可以看到，GLM 对于本种的预测精度很高。评估相对校准的微小减小表明该模型没有过度拟合。

如果选取了 ROC 曲线的临界值，利用全部数据的校准的估计可以去掉一些数据（第四列）。这将显示评估数据的基于临界概率的物种存在或不存在的最优化结果。临界值相关的敏感性和特异性在最后两列给出。这一临界值在后面将概率转换为存在或不存在（二元数据）数据或特定范围值的时候将会用到。

概率值已经被转换为 0-1000。

ROC，Kappa, TSS 方法的结构相同。

0.5.11 每个变量的重要性

很难对不同模型的预测进行比较，由于不同模型关于物种分布和环境的关系时，其算法、计算内容和假设是不同的。在模拟的步骤中，BIOMOD 可以计算出一个关于每个变量独立于模型的重要性的值。考虑到预测准确性，每个模型的每个物种该值单独存储。用汇总表提取结果会更方便。

运行 Models 函数将产生一个 VarImportance 的对象（当 VarImp 设定为大于 0）。查看方法如下：

```
VarImportance
```

要注意的是，只针对最终模型计算了变量的重要值。

记住，每个变量的重要值是 1 减去原始预测与模拟变量预测的相关系数。分值越高意味着该变量越重要，0 意味着根本不重要。

注意：

获得的相关系数可能为负。我们认为这种情况下变量的影响比相关系数为 0 的影响要大。此时，变量的重要值仍然是用 1 减去该相关系数，因此，重要值将大于 1.这种情况并不罕见。

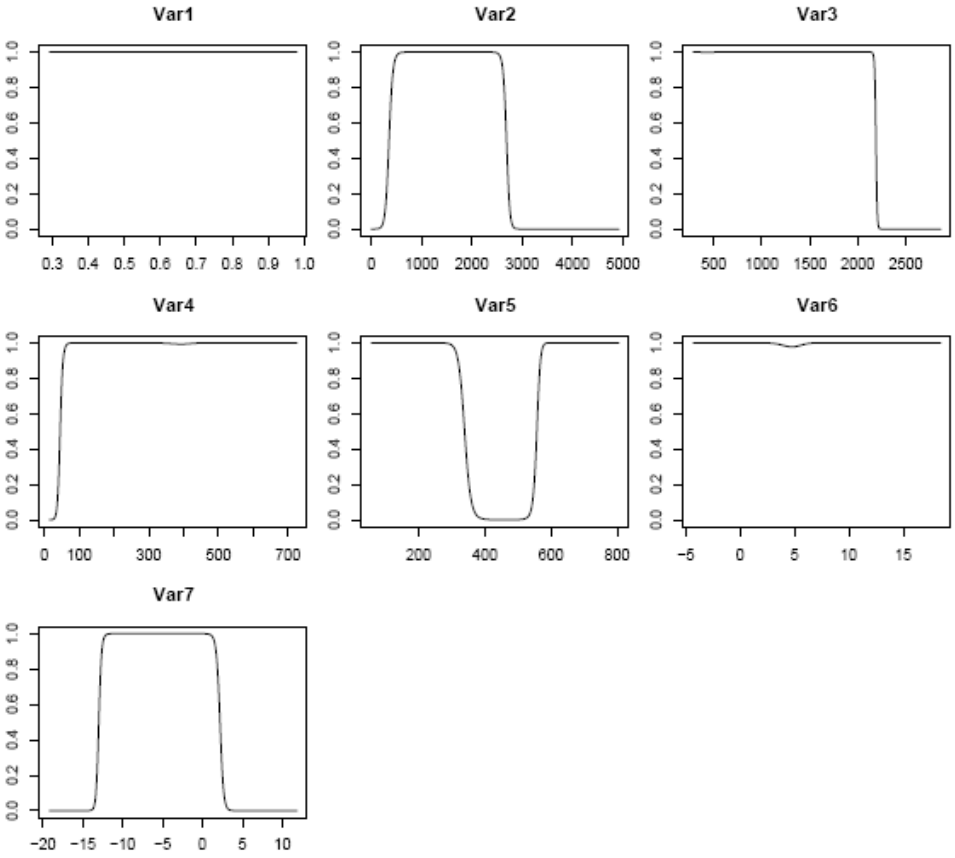
0.5.12 响应曲线

BIOMOD 在尺度合适的情况下，可以绘制每个模型的响应曲线。此时必须用到 response.plot. 该函数要用到所选的模型，所选的物种以绘制响应曲线。

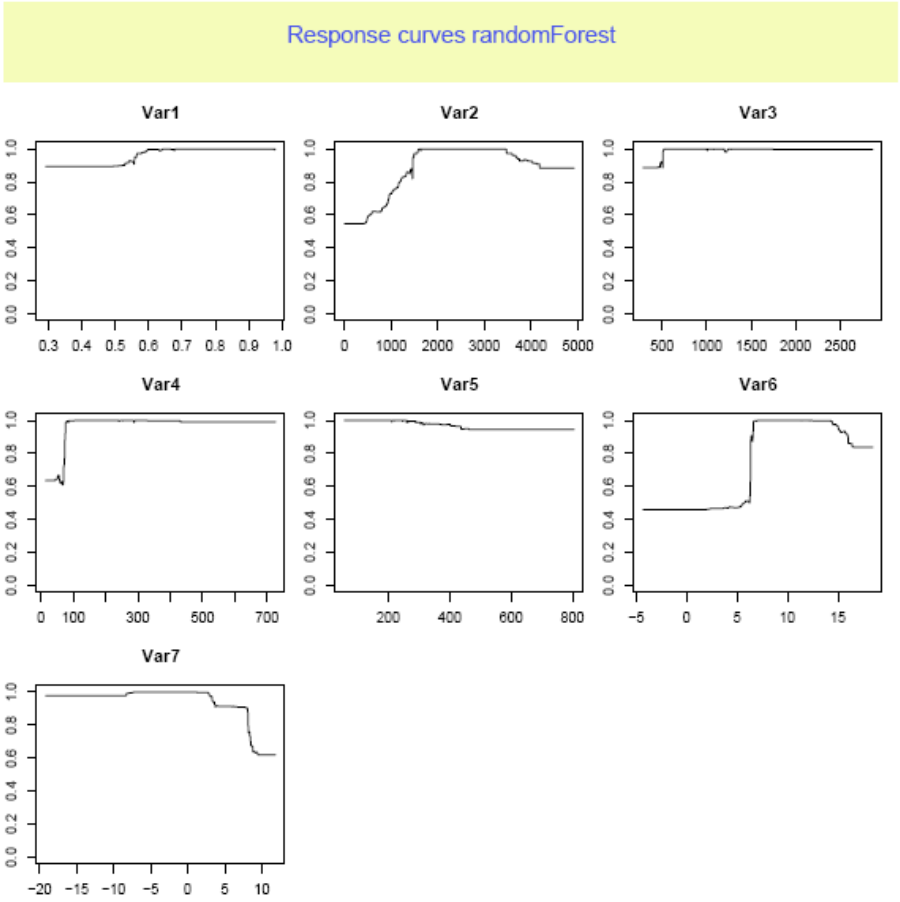
这给出第一种的 GLM 和 RF 例子。首先要载入模型（前面的命令已经导入内存中），在第一个参数的位置输入其名称，给出需要查看响应曲线的变量。

```
response.plot(Sp277_GLM_PA1, Sp.Env[4:10])
```

Response curves glm




```
response.plot(Sp277_RF_PA1, Sp.Env[4:10])
```



此时，N-1 个变量作为常量，各取其平均值，而我们感兴趣的变量包含 100 个点，位于最大值和最小值之间。预测结果的变异程度，通过这 100 个点，只表示我们所选变量的变异程度。所以，该图将模型对我们感兴趣的变量的响应可视化，而另外一些变量作为常量。

对所选的所有变量都将进行相应的处理。

0.5.13 对原始数据的预测

每个模型预测的各个种的数据存储在 pred 文件夹中。我们希望有一个表示所有模型对于所有种的预测结果的矩阵。

CurrentPred(GLM=T, GBM=T, GAM=T, CTA=T, ANN=T, SRE=T, MDA=T, MARS=F, RF=T,

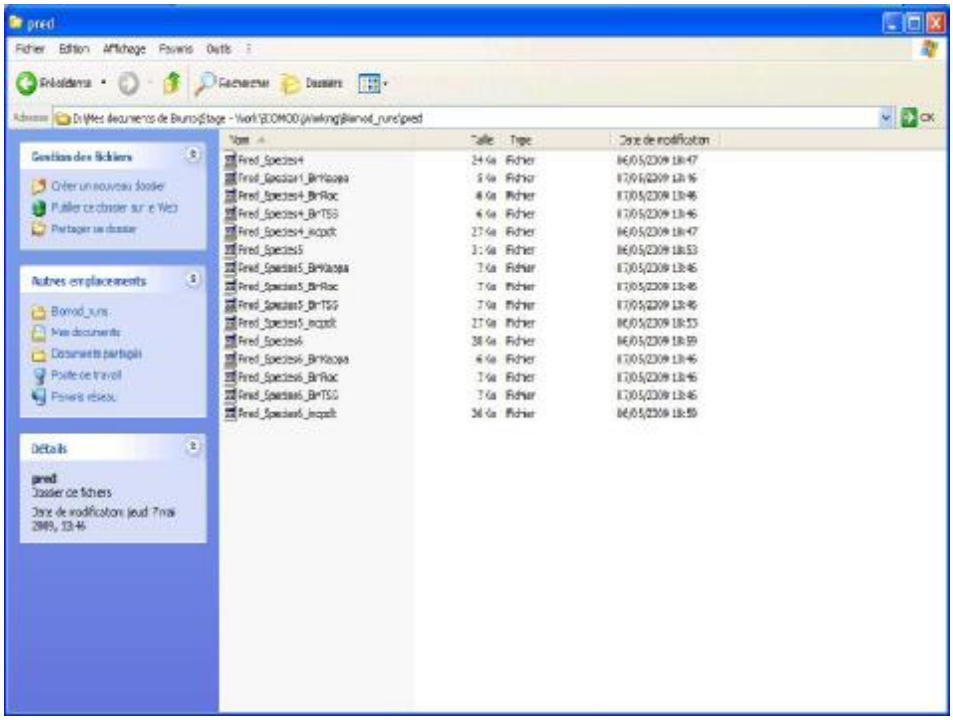
BinRoc=T, BinKappa=T, BinTSS=T, FiltKappa=T)

对于每个所选选择的模型(没有运行的模型将自动关闭)预测的对象，都将生成一个 Pred_Speciesname 的对象，如 Pred_Sp277(记住它们没有直接存储在工作空间，需要载入进来)。它们含有每个模型预测的物种出现的概率（记住其尺度是 0 到 1000）。

提取 Presence/absence 预测结果也许会有用。这需要设定 BinRoc,BinKappa,和/或 BinTSS 为 TRUE,则预测的结果都将给予 ROC， Kappa 或 TSS 转换为 presence/absence 数据。

同时将生成数据集：Pred Sp277 BinRoc, Pred Sp277 BinKappa, Pred Sp277

BinTSS, Pred Sp277 FiltKappa 等。没有扩展名的文件只能供 R 读取，而文本文件可以用于其它目的（如导入其它软件等）。



0.6 未来的情景

对于当前运行的所有模型，BIOMOD 都可以对物种潜在分布区，地区的土地利用情况，别的分辨率或时间的未来情形进行预测。BIOMOD 不会用到地理坐标，也不会在预测未来气候情景时对数据进行重新排序。用户应该保证所有数据出现的顺序一致，以保证对实际和预测的地图进行比较。

预测未来的情景，需要用到函数 Projection。

BIOMOD 给出了两种假设的未来气候数据库。分别称为 Future1 和 Future2。注意，列的名称应该与校准数据解释变量（或独立变量）的名称完全一致。

该函数的语法与之前函数的语法非常相近。首先加载新数据（例如气候变化情景），之后是输出结果名称的前半部分（Proj.name），未来情景预测中用到的模型。之后，用户可以选择预测的数据是否需要转换为 presence/absence 数据，或经（Kappa,ROC 或 TSS）临界值进行筛选。

Proj.name 选项非常重要，因为它关系到结果的存储，以及其他函数调用预测的数据。Project 函数生成一个同名的文件夹。在我们的例子中，它将在 pred 和 models 文件夹之后生成 proj.Future1 文件夹。每一个情景函数的运行都将产生一个新的文件夹。

```
> #like for calibrating the models, you can load your own data
> #Here we use the example file
> data(Future1)
> Projection(Proj = Future1[,4:10], Proj.name='Future1', GLM = T,
```

```
GBM = T, GAM = T,
   CTA = T, ANN = T, SRE = T, Perc025=T, Perc05=F, MDA =T, MARS = T,
RF = T,
   BinRoc = T, BinKappa = T, BinTSS = T, FiltRoc = T, FiltKappa = T,
FiltTSS = T,
   repetition.models=T)
```

查看 GLM 预测的结果：

```
   > load("proj.Future1/Proj_Future1_Sp277")
   > Proj_Future1_Sp277[740:760,,1,1]
   > load("proj.Future1/Proj_Future1_Sp277_BinRoc")
   > Proj_Future1_Sp277_BinRoc[740:760,,1,1]
```

0.7 模型优化

0.7.1 原始数据的预测

在运算过程中，BIOMOD 可以对模型进行直接比较。这样对于做出最优的预测将更为灵活。

函数 PredictionBestModel 将迭代的在每一次运行中，利用所选的方法（Roc,Kappa 或 TSS）检查哪些是最为准确的预测。要进行模型的优化，只需键入 T(TRUE) 或 F(FALSE)，就可进行模型优化。注意，如果已经用 Models 函数选取了所有的模型，则无需再对每个模型进行优化，而只针对感兴趣的模型即可。

该函数将生成 PredBestModelByX 为前缀的数据集，(X 对应于 Kapp,ROC 或 TSS 等检验的方法)，该数据集中将依据所选模型保存原始数据的预测数据。例如，第一种可以用 GLM 进行预测，第二种用 GAM 预测。所选的模型，预测的精度，相应的临界值，以及所选模型的敏感性、特异性都存储在新的数据集 BestModelByRoc 中。在一次模型最优化过程中，只能选择一种可以最优的评估方法如"Kappa"，或者全选。还有两个外的选项：由于前面的选项生成概率值，需要二元数据的用户应该进行转换，转换的方法为：Bin.trans=T. 此时，新数据集将依照所采用的评估方法产生，如 PredBestModelByRoc.BinRoc.

若用户想要保留临界值以上的概率值（低于临界值的概率均被更改为 0，而临界值以上的部分保留），则需要选择 Filt.trans=T。

在我们的例子中，可以比较用三种不同评价方法得到不同种的模型。我们同时将概率转换为 presence/absence 数据。

```
   > PredictionBestModel(GLM=T,GBM=T, GAM=T, CTA=T, ANN=T, MDA=T,
MARS=F, RF=T, SRE=T,
   method='all', Bin.trans = T, Filt.trans = T)
```

根据 TSS 统计量进行的多重比较

```
   > load("pred/BestModelByTSS")
   > BestModelByTSS
```

根据 ROC 进行的多重比较

```
   > load("pred/BestModelByRoc")
```

```
> BestModelByRoc
```

根据 Kappa 进行的多重比较

```
> load("pred/PredBestModelByKappa_Sp277")
> PredBestModelByKappa_Sp277[740:760,]
```

0.7.2 对未来情景的预测和其他地区的预测

ProjectionBestModel 函数根据（ROC,Kappa 或 TSS）对未来情景进行的预测，取决于在 PredictionBestModel 函数选择的最优模型。

ProjectionBestModel 的形式与 Projection 的形式一样。用户需要给出在 projection 函数运行中的气候数据集名称。与 Predic-tionBestModel 函数类似，用户也能将经过优化的未来情形转换为 presence-absence 数据或用临界值筛选，相应的，应该键入 Bin.trans=T 及 Filt.trans=T。

```
>     ProjectionBestModel(Proj.name='Future1',     Bin.trans=T,
Filt.trans=T, method='all')
> load("proj.Future1/Proj_Future1_BestModelByTSS")
> dim(Proj_Future1_BestModelByTSS)
> dimnames(Proj_Future1_BestModelByTSS)[-1]
```

未来情形的预测保存在三维数组中，其中第二维是多次运行，第三维是物种。

```
> Proj_Future1_BestModelByTSS[740:760,,"Sp277"]
> load("proj.Future1/Proj_Future1_BestModelByTSS_Bin")
> Proj_Future1_BestModelByTSS_Bin[740:760,,"Sp277"]
```

注意，在运行 ProjectionBestModel 之前，需要运行 PredictionBestModel。

可以查看 pred，proj.Future1 文件夹查看新生成的对象。

0.8 预测情景的综合

物种分布模型的困难之一在于，可用的方法数目众多，并且还在继续增加，使得并不精通的用户难以寻则他们所需的方法(Elith, J. et al. 2006, Heikkinen, R. et al. 2006)。近来的分析也表明，不同方法之间的差异可能会很大，这更增加了选择合适模型的困难。在处理物种分布独立的数据预测其分布区随气候变化的情形时尤其如此(Pearson, R. G. et al. 2006, Thuiller,W. 2004) 。解决的模型间变化的办法之一就是通过在一套初始条件，模型类型，模型参数及边界值上拟合一个综合情景，分析产生的分布区边缘的界限，整理成一致性的概率方法，而不是仅仅给出一个输出结果(Ara ujo, M. B. and New, M. 2007, Thuiller, W. 2007) 。BIOMOD 给出了这个平台，进行情景综合。

在 BIOMOD 中，进行模型的综合有几种方法。这里采用 Ensemble.Forecasting 函数，以及其他方式：

四种"committee"平均值 (每个元素权重相同)：

- 基于概率
- 基于 ROC 方法的二元情形预测
- 基于 Kappa 方法的二元情形预测
- 基于 TSS 方法的二元情形预测

基于模型评分的加权方法也可用

基于 01 情形的平均值将给出出现的平均值。例如，对给出的 TSS 方法选择的点，六种模型给出 1，而两种模型给出 0。相应的平均值应为 0.75。该值从二元情形中提取，因此不能确定一个优先的临界值。不过将结果转换为 01 值仍然是可行的。

同时计算模型给出的概率中值。中值更为可信，因为其受到极值的影响较小。无法进行加权计算，也不能依据已知数据确定一个临界值。

```
> Ensemble.Forecasting(Proj.name= "Future1", weight.method='Roc', PCA.median=T,
binary=T, bin.method='Roc', Test=F, decay=1.6, repetition.models=T)
```

该函数返回 R 内存中的一个数表。在我们的例子中称为 consensus Future1 results。其中包括综合预测的所有计算信息，例如应用到当前预测的每种方法的预测效果（在 Test=TRUE 时）,在赋予权重时的权重得分，PCA.median 方法选择的模型（如果设置为 TRUE）。预测的结果直接存储在硬盘的相应文件夹中。

选项：

reprtition.models: 可以开启或关闭该选项。如果选择了，函数将汇总每次模型每种方法运行的结果进行汇总，生成最终的预测结果。最终的汇总结果与本选项设定的 TRUE 或 FALSE 不一致。

weight.method: 依据各模型的表现进行排序。decay 给出重要值的相对重要性。默认的模型重要性衰减为 1.6；见下面的例子：

```
models GAM GBM GLM ANN RF MARS CTA MDA
score with Roc 0.96 0.92 0.90 0.88 0.87 0.75 0.72 0.68
decay of 1 0.125 0.125 0.125 0.125 0.125 0.125 0.125 0.125
decay of 1.2 0.217 0.181 0.151 0.126 0.105 0.087 0.073 0.061
decay of 1.6 0.384 0.240 0.150 0.094 0.059 0.037 0.023 0.014
decay of 2 0.502 0.251 0.125 0.063 0.031 0.016 0.008 0.004
```

可以输入任意值（必须大于 1），设定所需的判断强度。衰减率为 1 表明各元素的权重相同。

PCA.median： 该选项是获取模型贡献信息不是基于每个模型的表现。

对每个所选模型的预测概率进行 PCA。在当前的 BIOMOD 版本下，汇总的模型是与 PCA 第一主成分最相关的模型。可是，利用 PCA 选择时可以采用几种方法。可以用来选择一个模型，也可以选择多个非加权的汇总模型（与 PCA 第一轴关系最密切的模型），或者用于非加权的 PCA 不同轴关系密切的模型。这些方法可以在文献中（Thuiller 2004, Araujo 2005, Araujo 2006）找到。

当前版本的 BIOMOD 中不能生成这一选项的数据，所选择的模型在函数输出信息中。

binary: 如果本选项设置为真，则，汇总的预测结果将转换为二元格式。临界值的设定随着选定的方法不同而不同。

-mean on probabilities : 平均概率，基于平均临界值转换二元数据（给出三种概率-ROC,Kappa 或 TSS, 需要在 bin.method 中设定）。

-weighted mean on probabilities: 加权平均概率，基于加权的平均临界值转换为二元数据（采用同样的方法进行排序，例如 weighted.method 选项）

ROC-Kappa-TSS 平均： 设定值为 500 （等价于概率为 0.5），意味着超过一半儿的概率认为该种在某区域存在。

Test: 该选项将利用校正汇总校准模型数据检验模型的有效性。对该模型进行 ROC

曲线的检验，其结果存储在 test.results 中。

输出：

该函数对每个种都将运行。每个种都将生成一个对象。这些对象均为三维数组。

```
load("proj.Future1/consensus_Sp164_Future1")
dim(consensus_Sp164_Future1)
dimnames(consensus_Sp164_Future1)[-1]
```

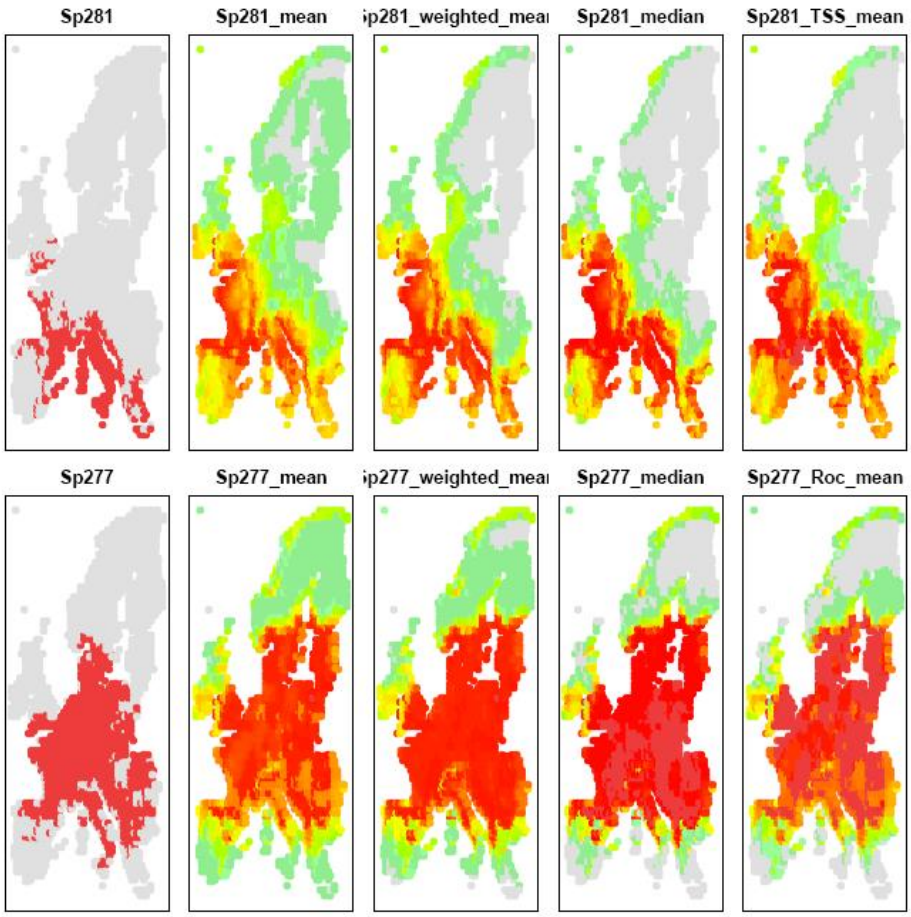
第二维是运行的重复次数，第三维是汇总方法。还有一个对象称为"Total consensus Future1"，为每个重复输出结果。

```
> load("proj.Future1/Total_consensus_Future1")
> dim(Total_consensus_Future1)
[1] 2264 3 6
> dimnames(Total_consensus_Future1)[-1]
[[1]]
[1] "Sp281" "Sp277" "Sp164"
[[2]]
[1] "prob.mean" "prob.mean.weighted" "median"
[4] "Roc.mean" "Kappa.mean" "TSS.mean"
```

当前，第二维是物种。现在来绘制一些其中一些预测结果：

```
> Total_consensus_Future1[1:20,,1]
Sp281 Sp277 Sp164
1 345.4 54.83 62.01
2 425.9 61.78 67.44
3 372.5 59.97 64.04
4 507.5 46.75 88.75
5 562.8 46.44 88.93
6 542.4 46.42 102.38
7 572.1 46.25 96.15
8 345.9 48.50 69.29
9 366.5 48.08 84.38
10 351.6 49.39 85.32
11 538.1 45.78 107.82
12 404.8 63.39 72.50
13 258.6 51.61 59.31
14 244.3 55.14 67.11
15 239.2 50.11 62.61
16 401.2 68.69 73.61
17 580.0 122.17 96.43
18 551.6 43.72 125.57
19 564.9 43.36 109.85
20 596.5 50.00 121.74
> data <- Total_consensus_Future1
> par(mfrow=c(2,5))
> par(mar=c(0.6,0.6,2,0.6))
> level.plot(DataBIOMOD[,8], CoorXY, show.scale=F, title='Sp281',
```

```
cex=0.5)
   > level.plot(data[,1,1], CoorXY, show.scale=F, title='Sp281_mean',
cex=0.5)
   >       level.plot(data[,1,2],       CoorXY,       show.scale=F,
title='Sp281_weighted_mean', cex=0.5)
   >       level.plot(data[,1,3],       CoorXY,       show.scale=F,
title='Sp281_median', cex=0.5)
   >       level.plot(data[,1,6],       CoorXY,       show.scale=F,
title='Sp281_TSS_mean', cex=0.5)
   > level.plot(DataBIOMOD[,9], CoorXY, show.scale=F, title='Sp277',
cex=0.5)
   > level.plot(data[,2,1], CoorXY, show.scale=F, title='Sp277_mean',
cex=0.5)
   >       level.plot(data[,2,2],       CoorXY,       show.scale=F,
title='Sp277_weighted_mean', cex=0.5)
   >       level.plot(data[,2,3],       CoorXY,       show.scale=F,
title='Sp277_median', cex=0.5)
   >       level.plot(data[,2,4],       CoorXY,       show.scale=F,
title='Sp277_Roc_mean', cex=0.5)
```



如果 binary 设定为真，则包含了汇总结果将为二元数据结果名称末尾将改为_Bin。

0.9 物种迁移

在预测未来情形的物种分布时，该函数允许物种能够进行迁移。该模型允许物种在当前的地理位置上，特性的圆圈内迁移。物种能够迁移的半径需要由用户来指定。该函数使用两个数据集：当前的物种分布及未来的物种分布（在没有迁移限制的情况下）。

数据中的经纬度可以用来计算允许的迁移距离。

注意： 使用该函数，则当前及未来的数据集都应该存在，并且拥有相同的坐标系和分辨率。

接下来需要设定迁移率。选项可用。用户可以设定对每个种的迁移率（必须给出数值），或者对每个种的迁移率给出不同的值（给出一个向量）。

对共同的迁移率，可以建立一个向量，该向量的行数对应着种数，每一行包含着对应种的最大可能迁移距离。

最后，给出限定迁移的预测名称。例如，Future1.Migration.1km.per.year

注意，迁移率可以用度来表示。例如，某个种的 10 年的最大迁移率为 1 分（约 1.6 千米）。如果我们预测其 50 年的分布变化：Rate=1x0.16667x5()

每十年最快速度为 3 分的（4.8 千米）,在 2080 年，应为 Rate=3x0.16667x8

```
   Projection(Proj = Sp.Env[,4:10], Proj.name='Current',
   GLM = T, GBM = T, GAM = T, CTA = T, ANN = T, SRE = T, Perc025=T,
Perc05=F, MDA =T, MARS = T,
   RF = T, BinRoc=T, BinKappa=T, BinTSS=T, FiltRoc=T, FiltKappa=T,
FiltTSS=T, repetition.models=T)

   Ensemble.Forecasting(Proj.name= "Current", weight.method='Roc',
PCA.median=T,
   binary=T,       bin.method='Roc',       Test=F,       decay=1.6,
repetition.models=T)
   load("proj.Future1/Total_consensus_Future1")
   load("proj.Current/Total_consensus_Current")
   Migration(CurrentPred = Total_consensus_Current[,,1], FutureProj
= Total_consensus_Future1[,,1],
   X=CoorXY[,1],         Y=CoorXY[,2],         MaxMigr=5*0.16667*8,
Pred.Save="Future1.Migration")


   Future1.Migration[740:760,]
```

0.10 物种空间变化率 Turnover

该函数根据一定时间段内的像素变化来进行物种的丧失、增加及空间变化率的估算。该函数有两套数据：物种当前的分布和未来的分布（例如考虑到物种迁移）。注意，当前和未来的预测必须是二元数据（出现或不出现数据）。最后，为空间变化率计算的结果指定一个名称。

在该函数存储的数据库中，包括 10 列。

第一列为相对数值。Disa 表示预测的从某一个点消失的物种。Stable0 是该点不含的物种数及没有迁移到的物种数。Stable1 表示给定的点上物种的数量，及在未来仍然保留的种数。Gain 表示该种当前不分布在本地区，但是在预测的情景中，将迁移到该点上。

PerLoss,PercGain 及 Turnover 与下面的百分率有关：

PerLoss=100×L/(SR)

PercGain=100×G/(SR)

Turnover=100×(L+G)/(SR+G)

SR 是当前的物种丰富度

CurrentSR 表示该点用当前模型预测的物种丰富度。

FutureSR0Disp 表示模型预测没有迁移的未来情形下物种丰富度。

FutureSR1Disp 表示模型预测具有迁移的未来情形下物种丰富度（取决于输入的数据，如果设定了迁移的话）。

```
> ProjectionBestModel("Current")
> load("proj.Future1/Proj_Future1_BestModelByRoc_Bin")
> load("proj.Current/Proj_Current_BestModelByRoc_Bin")
> Biomod.Turnover(CurrentPred =
Proj_Current_BestModelByRoc_Bin[,1,],
FutureProj = Proj_Future1_BestModelByRoc_Bin[,1,],
Turnover.Save= "Turnover.2050")
> Turnover.2050[740:760,]
```

0.11 物种分布区变化

该函数可以估算物种所占像素丧失、获得或者维持稳定的相对比例。

该函数将用到两个数据。物种当前分布及未来分布。注意，当前及未来预测必须是二元格式（出现或不出现）。最后，要设定物种分布区变化结果存储在什么对象。

```
Biomod.RangeSize(CurrentPred =
Proj_Current_BestModelByRoc_Bin[,1,],
FutureProj = Proj_Future1_BestModelByRoc_Bin[,1,],
SpChange.Save="SpChange.2050")
```

可以生成两个数据集的列表：Compt.By.Species 和 Diff.By.Pixel

Diff.By.Pixel 保存了每个种的有用信息。种处于每一列上，点 pixel 位于行上。对于每个种来说，每个像素有四个不同的值。

-2 该点上该种丧失。

-1 该点上该种稳定。

0 该点没有被该种占据，未来也不会迁移到该点上。

1 该点当前没有被占据，但是该种将会迁移到该点上。

SpChange.2050$Diff.By.Pixel[740:760,]

这个数据表可以用 GIS 软件绘制到地图上，以表示所关注种的格局变化。

Comt.By.Species 存储着每个种分布区变化的信息（以行表示）。前四列为相对值：Disa 表示预测的某个种丧失的点的数目。Stable0 表示表示物种当前没有占有的，并且未来也不会占有的点的数目。Stable1 表示该物种当前占据，并且将来仍然会占据的点数目。PercLoss，PercGain, SpeciesRangeChang 表示以下四个百分率：

-CurrentRangeSize 表示模拟的某个种当前的分布区大小（占有的点数目）。

-FutureRangeSize0Disp 表示在没有迁移的情况下，某个种的分布区大小。

-FutureRangeSize1Disp 表示在考虑物种迁移的情况下，某个种分布区的大小（取决于数据的输入，是否设定了 Migration 选项）。

SpChange.2050$Compt.By.Species

0.12 其他函数

这一节展示的是一系列与 BIOMOD 的功能并不直接相关的一些函数。这些函数可以用于任何数据，将会在后续的分析中需要它们。因此，不需要运行 BIOMOD 来使用它们。

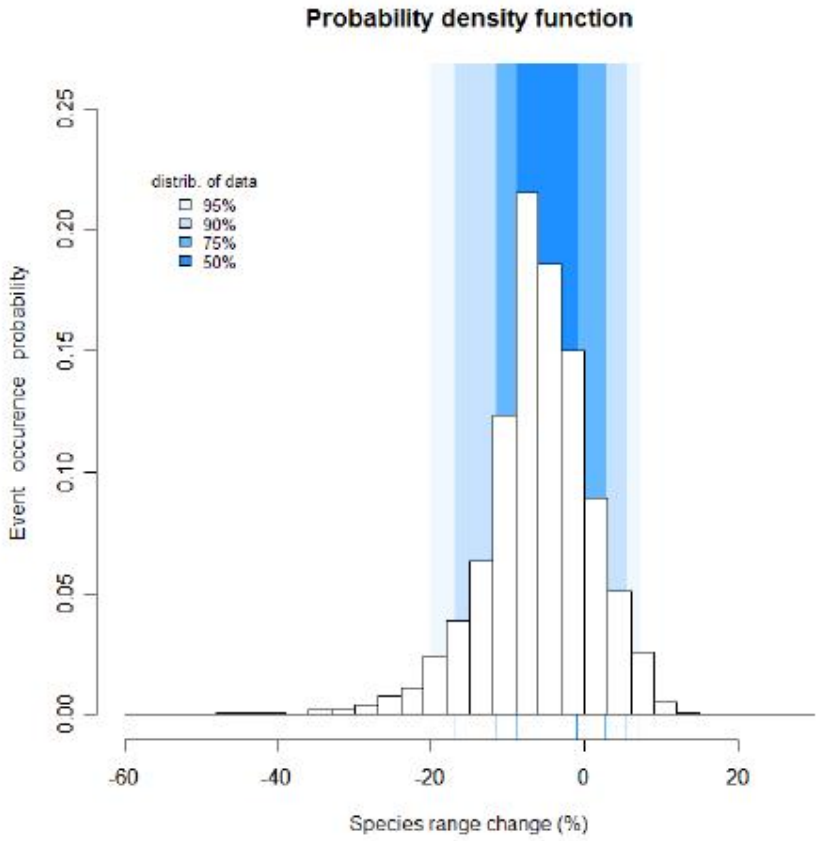
但是，不要将这些命令行拷贝到 R 控制台中并试着运行它们。因为这将返回一个错误的结果。下面的命令是一个例子。

0.12.1 概率密度函数

在预测中利用大量的参数模拟将不可避免的引入一定程度的变异，特别是在进行未来情形预测的时候。这个函数可以从总体上查看每个种的未来预测，并给出每个分布区变化的似然估计。

未来的分布区变化是以当前的百分率为基础计算的。例如，如物种当前占据了 100 个格子，而预测在未来将占据 120 个格子，分布区变化将为 20%。

```
ProbDensFunc(initial=Sp.Env[,9],      projections=Proj[,1:120],
distrib=T, cvsn=T, groups=gp, resolution=5)
```



initial： 为二元向量（0，1）表示物种分布的当前状况，用于分布区变化计算时的参照。

projection: 在预测中所有值，每一列为一种预测。确定保持预测的结果矩阵与intial 向量的顺序一致（第一行 为点 1，第二行为点二，以此类推）

distrib: 如果为真，将数据浓缩 50%，75%，90%和 95%的最优化方法将被选定，并绘制在地图上。

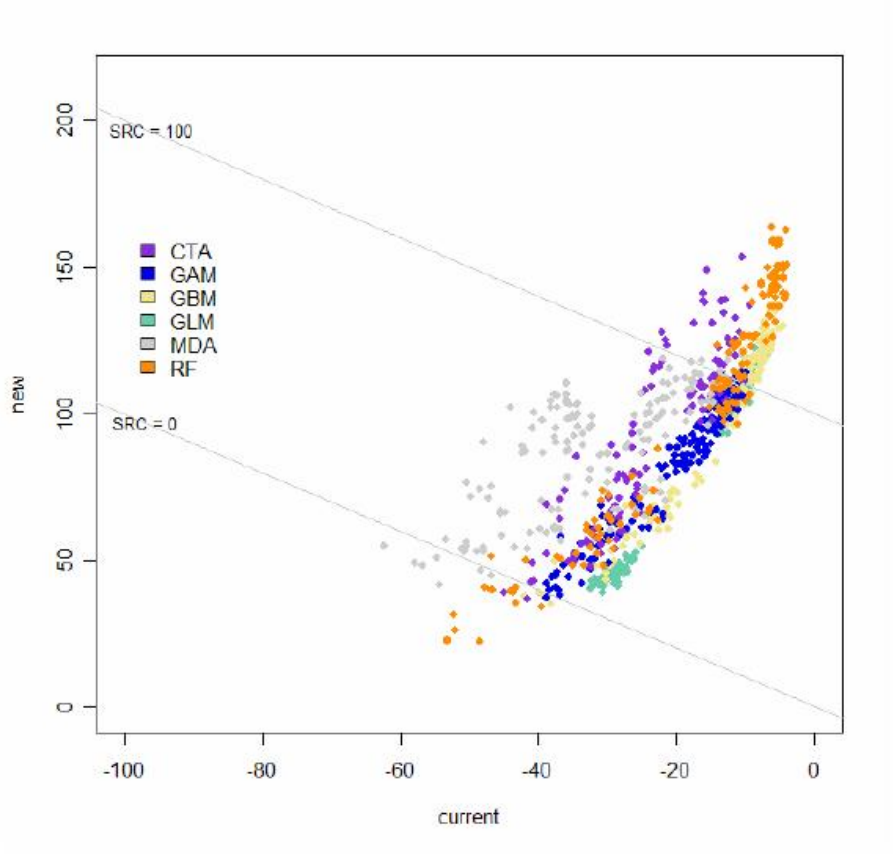
Resolution： 在图中设定预测值等级。默认为 5。

NOTE: 修改分辨率将直接影响到概率值。分辨率上升，将使预测量增加，而使总的预测分出的块数更多。概率只是针对某一特定的等级，而不是针对某一次预测。

cvsn: current vs new 的缩写，如果设定为真，则分布区变化将用两种方式计算：物种当前占有栅格数将丧失的百分率，及与当前物种分布区相比较，当前没有占据的，但是预测将会占据的栅格数,也就是“新”栅格。

上面的例子中，物种在未来将占据 120 个适宜的栅格，但是当前占据了 100 个。这可能是由不同事件引起的。一种情况下，当前 100 个栅格保留，而新占据了 20 个栅格，组成 120 个栅格。另一种可能是原有的 100 个栅格全部丧失了，而未来情况下，重新占据了另外的 120 个栅格。

这两种情况下，SRC 计算结果相同，但是第一种情况没有显示出多少生存对策的信息（在未来情形下，现有分布区的种群仍将生长的很好，物种甚至将进一步发现并定植到新的分布区）。但第二种情况表明，物种的迁移作用很强，种群倾向于生长在适宜的环境中。用这种方法可以将两种情况的 in-between 概率区分开来。



这里，每个点，是一个预测结果。例如，最左边的点给出如下信息：约 60%的现有点将丧失，而将占有 50%的新的点。SRC 曲线将其简化为物种分布区减少-10%。见这个简单的值没有反映出当前所有信息：没有显示出当前生境的丧失情况，而该信息将导致不同的管理对策。

两条线表示 SRC 曲线为 0（即适宜生境没有一点儿变化），和 100%（物种分布区范围将增加一倍）。沿着这些曲线，你可以有多种可能的组合给出同一个值（-10+10=0，-40+40=0；……）。

图中另一信息就是颜色。它们显示出利用当前的例子，模型给出的不同预测结果（见下文 group 说明）。在 groups 矩阵中拥有的行数越多，获得的图形也将越多。

groups：对某一类群的预测结果，给出不同的可视化结果。需要用到一个矩阵，该矩阵每一列表示一次预测解说，每一行表示每个参数的详情。例如，现有 9 种预测结果，3 个模型和三个阈值，此时，矩阵的形式如下所示：

```
[,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9]
[1,] "GAM" "GAM" "GAM" "CTA" "CTA" "CTA" "RF" "RF" "RF"
[2,] "Roc" "Kappa" "TSS" "Roc" "Kappa" "TSS" "Roc" "Kappa" "TSS"
```

或

```
[,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9]
[1,] "GAM" "CTA" "RF" "GAM" "CTA" "RF" "GAM" "CTA" "RF"
[2,] "Roc" "Roc" "Roc" "Kappa" "Kappa" "Kappa" "TSS" "TSS" "TSS"
```

需要记住的是，该矩阵表示在 projection 选项中选取的参数。需要将矩阵排列的顺序与 projection 的顺序相同。

不确定性估计

本函数可以对预测中的不确定性进行估计。PDF 图使得可视化估算，也可以对每个参数的变化的估计值进行计算。

uncertianty: 如果设置为真，可以计算 groups 选项的每个参数的变化量。3 个或更少的参数（如在 group 矩阵有 3 行），将给出一个数据表。这里是 3 个参数的例子：9 种模型，3 个阈值，5 种未来气候情形。输出的结果是 R 的标准模式，这里给出的仅表是为了更好的阅读。

	Roc	Kap	Prev	Sc1	Sc2	Sc3	Sc4	Sc5
GLM	0.047	0.050	0.046	0.106	0.107	0.109	0.113	0.109
GBM	0.072	0.092	0.070	0.115	0.118	0.119	0.119	0.118
GAM	0.068	0.074	0.067	0.100	0.098	0.098	0.100	0.097
CTA	0.168	0.167	0.175	0.175	0.184	0.182	0.185	0.181
ANN	0.205	0.225	0.173	0.196	0.206	0.197	0.210	0.200
MDA	0.138	0.136	0.140	0.139	0.154	0.149	0.144	0.150
MARS	0.329	0.271	0.211	0.387	0.368	0.379	0.366	0.374
RF	0.108	0.122	0.095	0.164	0.172	0.168	0.169	0.173
Roc	NA	NA	NA	0.265	0.275	0.273	0.266	0.278
Kap	NA	NA	NA	0.234	0.266	0.249	0.233	0.263
Prev	NA	NA	NA	0.314	0.317	0.313	0.316	0.318

可以发现四个框： model/threshold（左上方）。model/(scenarios)(右上方)，threshold/scenarios (右下方) 和 threshold/threshold (左下方，NA 值的意思是数据不可用)

我们先看一下矩阵左上角的第一个值。阅读方法如下：预测未来情形只用了 GLM 和 ROC 评价方法。这给出了 5 种预测，每一个是一种情形。每一行给出了预测之间的标准误。矩阵中的值表示不同行间的标准误。表示不同情形下的变化。

可以发现，在不同的阈值评价方法下，不同情形的影响多少接近常数（如每一行三个值之间），但是在模型之间变化较大（每一列的前 8 个值）。不同的情形对 MARS 模型的影响最大，在预测时可能会给出迥异的结果。对于每个模型，评价阈值的方法似乎具有更大的影响。

注意：

时刻谨记标准差会受到样本量的影响。样本数量越大，降低其差异的可能性就越大。同时，例如，利用极端气候情景做出的预测将产生更大的变化，而不仅仅是几个近中间的情形。在解释这些值的时候要小心。

一个重复的例子

ProbDensFunct 函数的帮助文件给出了完整的例子。用部分数据对模型进行校准后，对一半的模型进行了 20 次预测，来评估预测中的变化量。只对 Sp163 进行了分析。请参见帮助文件，查看数据准备，之后才能正确的运行函数。

```
example(ProbDensFunc)
```

在你的 R 中，该函数将生成一系列表示所得预测的变化相关的图。

0.12.2 Pseudo-absences

大部分模型需要 presence 和 absence 数据，才能确定特定种的适宜生长环境。但是有些数据集并不含有 absence 数据，而只有出现的数据，此时就需要建立虚拟的 absence。例如，对于鸟类来说，确定 abesence 是十分令人头疼的事情。假定的 absence 称为 pseudo-absence，这些信息没有经过野外调查的预测。

Pseudo-absence 可以是没有物种分布记录的，同时环境条件可能引起 absence 的任意点。如果输入模型的数据中 absence 的数据较多，会强烈的影响到模型对物种分布和气候关系的判定。不仅如此，运行这样的庞大的数据将会耗费更多的时间。

此外，指定的 absence 可能实际上是有分布的，(特别在样本不完整的情况下)，因此，pseudo-absence 数据给出的物种气候关系的估计是错误的。所以，在运行模型之前，我们将用多种方式将选择的 pseudo-absence 的不良影响去除。

如下面例子所示，利用 pseudo.abs 函数：

```
pseudo.abs(coor=data[,1:2],  status=data[,3],  strategy='per',
env=data[,4:16], distance=10000, plot=F,
species.name= 'Sp1', acol='grey80', pcol='red', add.pres=T)
```

coor: 具有两列数据的矩阵，给出 presence 点的坐标，及整个潜在 absence。

status: 包含 1-0 数据信息的向量。没有设定为 1 的点都默认赋值为 0，也就是认为是 absence。

strategy: (见下图中的例子)

-random: absence 将从潜在 absence 中随机选取。

-per: 代表 presence 总周长。

-perind: 与 per 选项相同，但是对每个 presence 计算周长。这种方法需要所需距离的信息（distance 选项）

-sre: 具有物种倾向于分布的环境的数据点（基于 SRE 模型）

这样的点将不作为 pseudo-absences 的候选点。选择本选项的时候，必须选定 env 参数。

distance: 只适用于 perind。其单元为 coor 数据之一。

env. 是“sre”选项所需的。该矩阵给出一组变量作为环境信息。

species.name: 结果所保存的种名将由此选项给出，这个参数结尾应该再加一个点儿。例如，假如给出的参数为 larix,选用的模型为 sre，之后，输出的结果将存储在 larix.sre 文件中。

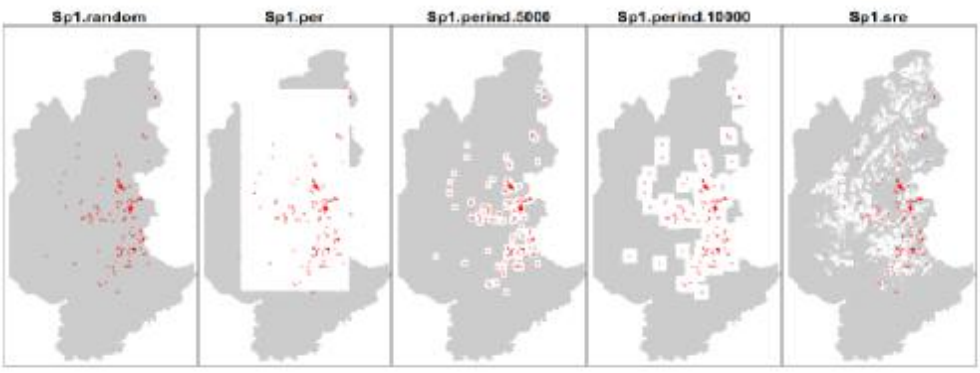
nb.points:有限的随机选取一些点的选项。默认为（nb.points=NULL）根据方法选取的不同，保持着所有可能的 absences.

add.pres: 如果为 TRUE,则输出结果将包括精确的信息。

plot: 获取 presence 和 absence 输出结果

acol 和 pcol : 分别设定不出现和出现的颜色

4 中可用方法的例子表示的是法国阿尔卑斯山地区 Larix decidua Miller 的分布情况。出现为红色，每种方法选择的 pseudo-absences 为灰色。



怎样正确的使用 pseudo.abs 函数的输出

该函数的输出结果，包含某种方法从原始 presence-absence 数据集选出的 absence 的行（如果 add.pres 设定为真，也将包括 presence）。注意，如果使用了 nb.points 选项，将产生有限个 absence。下面是如何正确使用输出结果：

假如你的原始数据名为 fulldata, 而你希望用 sre 方法选择 pseudo-absences。运行 pseudo.abs 函数：

```
pseudo.abs(coor=data[,1:2],  status=data[,3],  strategy='sre',
env=data[,4:16],
species.name= 'first.species', add.pres=T)
```

运行之后将生成一个名为“first.species.sre”的对象，其中包含所有可能的 absece，也包括 presences（因为在设定了该选项）。新的数据可以用如下方式处理：

```
new.data.set<-fulldata[first.species.sre,]
```

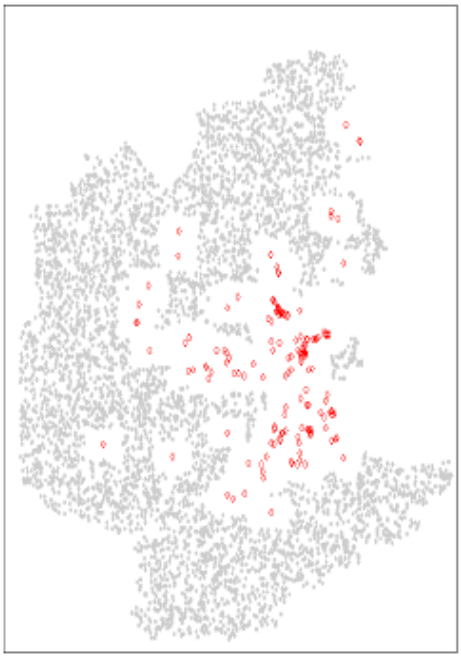
读取原始数据集相应的行，建立一个新的数据集，并起一个新名字。如果想从 absence 方法选取的点中选取 5000 个点（假设你拥有更多的点），或者你不想保留 presences，接下来要做的与设定合适的值相同。

例如：

```
pseudo.abs(coor=data[,1:2], status=data[,3], strategy='perind',
distance=10000, plot=T,
species.name= 'Sp1', nb.points=5000, add.pres=T)
```

此时你的数据将如下所示：

Sp1.perind.10000



0.13 模型描述

0.13.1 GLM- 广义线性模型

广义线性模型比经典多元回归的形式要自由，能提供变量的误差分布，而不仅仅是正态和方差不恒定。如果相应变量与预测变量的关系并非线性，则需要进行转换，在-functions 或 hierarchical 模型等的多项式转换中，可以允许偏态的和双峰的响应变化。随之而来的缺点是，物种和环境变量之间的关系需要事先了解。不仅如此，很多情况下，GLM 在趋近于回归面的时候，灵活性往往不够。为了选取最为简约的模型，BIOMOD 应用了自动逐步模型筛选。Splus 的 stepAIC 函数能够为模型增加新的变量，并检验其对模型的贡献，删除对拟合结果影响不显著的变量，从而对模型进行优化。模型的拟合度的统计量可以是 Akaike Information Criterioin(AIC)或 Bayesian Information Criteria(BIC) 逐步回归能够去除冗余变量，减少共线性（虽然不是总是）

可以运行三种广义线性模型

Y 1 = X1 + X2 + X3 + (X1*X2) + (X2*X3)

GLM Quad: Used linear, 2nd and 3rd order.

Y 1 = X1 + X12 + X13 + X22 + X33

GLM Poly: Use ordinary polynomial terms.

Y 1 = f(X1 + X12 + X13) + f(X2 + X22 + X23) +

如果选择 GLM，只需设定 GLM=T 即可。

如果希望选择 Polynomial terms，键入 TypeGLM="poly"，或二次， 键入 TypeGLM="quad"，或利用线性的，键入 TpleGLM="simple"。 如果希望用 AIC 作为选择标准，只需键入 Test="AIC",希望使用 BIC 的时候，键入“BIC”。

Key reference.

McCullagh, P. and Nelder, J.A. (1989) Generalized linear models Chapman and Hall.

Key reference in ecology/biogeography.

Austin, M.P. and Meyers, J.A. (1996) Current approaches to modelling the environmental niche of

eucalypts: implication for management of forest biodiversity. Forest Ecology and Management, 85,95-106.

Elith, J., Graham, C.H., Anderson, R.P., Dudik, M., Ferrier, S., Guisan, A., Hijmans, R.J., Huettman,

F., Leathwick, J.R., Lehmann, A., Li, J., Lohmann, L., Loiselle, B.A., Manion, G., Moritz, C., Nakamura, M., Nakazawa, Y., Overton, J.M., Peterson, A.T., Phillips, S., Richardson, K., Schachetti

Pereira, R., Schapire, R.E., Sober on, J., Williams, S.E., Wisz, M., and Zimmermann, N.E. (2006) Novel methods improve predictions of species' distributions from occurrence data. Ecography, 29,129-151.

Guisan, A. and Thuiller, W. (2005) Predicting species distribution: o
ering more than simple habitat models. Ecology Letters, 8, 993-1009.

Guisan, A. and Zimmermann, N.E. (2000) Predictive habitat distribution models in Ecology. Eco-logical Modelling, 135, 147-186.

Thuiller, W., Ara ujo, M.B., and Lavorel, S. (2003) Generalized models versus classification tree analysis: a comparative study for predicting spatial distributions of plant species at di erent scales. Journal of Vegetation Science, 14, 669-680.

0.13.2 GAM 广义相加模型

该模型最近被应用到生态学领域，用来处理物种对环境因子的响应问题。GAMs 无需像 GLMs 那样需要很容易出问题的步骤假定曲线的形状或特性参数的响应函数，比 GLM 的能力更强。这类模型应用一类称为“smoother”的方程，将数据分段后转换为平滑曲线。GAMs 在数据形式更为复杂，难以用标准线性或非线性模型拟合的，或者没有确定应该用哪种特定模型的时候更为有用。其思想是将响应变量对某一个环境变量作图，在简约的前提下，计算用来尽可能趋近的拟合数据的平滑曲线。该算法对于每个变量绘制一个平滑曲线，之后对结果相加。BIOMOD 应用 3 次插值平滑算法，取决于分段，它们是自由度小于或等于 3 的多项式集合。相邻数据区段用不同的多项式进行拟合，因此可将所有的点连起来。与 GLM 类似，BIOMOD 利用自动逐步回归选取每个种最显著的变量。

Y = s(X1,4)+s(X2,4)+s(X3,4)

用户需要选定自由度。默认情况下，自由度为 4.之需要键入 Spline=4. 换言之，自由度为 4 与多项式的自由度 3 接近。

Key reference.

Hastie, T.J. and Tibshirani, R. (1990) Generalized additive models Chapman and Hall, London.

Key reference in ecology/biogeography.

Austin, M.P. and Meyers, J.A. (1996) Current approaches to modelling the environmental niche of eucalypts: implication for management of forest biodiversity. Forest Ecology and Management, 85,95-106.

Elith, J., Graham, C.H., Anderson, R.P., Dudik, M., Ferrier, S., Guisan, A., Hijmans, R.J., Huettman,

F., Leathwick, J.R., Lehmann, A., Li, J., Lohmann, L., Loiselle, B.A., Manion, G., Moritz, C., Nakamura, M., Nakazawa, Y., Overton, J.M., Peterson, A.T., Phillips, S., Richardson, K., Schachetti

Pereira, R., Schapire, R.E., Sober on, J., Williams, S.E., Wisz, M., and Zimmermann, N.E. (2006)

Novel methods improve predictions of species' distributions from occurrence data. Ecography, 29, 129-151.

Guisan, A. and Thuiller, W. (2005) Predicting species distribution: o
ering more than simple habi- tat models. Ecology Letters, 8, 993-1009.

Guisan, A. and Zimmermann, N.E. (2000) Predictive habitat distribution models in Ecology. Ecological Modelling, 135, 147-186.

Thuiller, W., Ara ujo, M.B., and Lavorel, S. (2003) Generalized models versus classification

tree

analysis: a comparative study for predicting spatial distributions of plant species at di erent scales. Journal of Vegetation Science, 14, 669-680.

Yee, T.W. and Mitchell, N.D. (1991) Generalized additive models in plant ecology. Journal of Vegetation Science, 2, 587-602.

0.13.3 CTA 分类树分析

该算法提供了与回归计算不同的思路。与 GAM 类似，无需对预测变量和响应变量之间的事先假设。该方法包括经过对响应变量的分析，将预测变量确定的空间递归的划分为尽量同质的类群。树建立过程中，采用一个简单规则，基于简单的解释变量，重复的将数据分组。每次数据分组中，数据均被分为两组，每一组尽量同质。算法会搜寻每个组内尽可能小的方差。每个节点的异质性可以解释成为高斯模型的偏差（回归树）或多项模型（分类树）。结果表示的是偏差函数和成本复杂性参数。最好的树是偏差最小和叶数目最少的权衡。BIOMOD 应用 rpart 包进行分类树计算。为了控制树的长度，程序建立子树的巢式序列，根据方差解释量，递归的将次要的分组去除。BIOMOD 采用交叉验证（X-fold cross-calidation ）的步骤权衡叶的数目和方差解释量。用户可以指定交叉验证的次数。

如果想使用分类树分析模型，只需键入 Tree=TRUE. 之后设定交叉验证的次数 CV.tree=10。

交叉验证的次数没有最优化。次数越高，需要的内存也越多。

主要参考文献：

Key reference.

Breiman, L., Friedman, J.H., Olshen, R.A., and Stone, C.J. (1984) Classification and regression trees Chapman and Hall, New York.

Key reference in ecology/biogeography.

De'Ath, G. and Fabricius, K.E. (2000) Classification and regression trees: a powerful yet simple technique for ecological data analysis. Ecology, 81, 3178-3192.

Thuiller, W., Vaydera, J., Pino, J., Sabat e, S., Lavorel, S., and Gracia, C. (2003) Large-scale environmental correlates of forest tree distributions in Catalonia (NE Spain). Global Ecology and Biogeography, 12, 313-325.

Vayssi eres, M.P., Plant, R.E., and Allen-Diaz, B.H. (2000) Classification trees: an alternative nonparametric approach for predicting species distributions. Journal of Vegetation Science, 11, 679-694. 80

0.13.4 ANN 人工神经网络

前馈神经网络为广义线性回归函数提供了新的灵活方式。它们并非线性回归模型，但是具有为数众多的参数，因此极为灵活；灵活到可以趋近任何平滑函数。ANN 的准确性主要取决于两个参数：权重的衰减（weight decay）程度和隐元的数量（hidden unit）。BIOMOD

应用 nnet 包进行相应计算。不同的运行将产生不同的结果，最优化的权重衰减和在隐层中“元”的数量的（或者等于变量的数目（参见 Wierenga et Kluytmans, 1999）或变量数目的 75%）是经过 N 重交叉验证确定的（默认为 3）。用户可以选择交叉验证的次数。注意，ANN 运算需要的时间长，要尽量避免额外的交叉验证。

如果要用到 ANN 模型，只需键入 ANN=T。之后，设定交叉验证的次数为 3， CV.ann=3。

Key reference.

Ripley, B.D. (1996) Pattern Recognition and Neural Networks Cambridge. Key references in ecology/biogeography

Lek, S., Delacoste, M., Baran, P., Dimopoulos, I., Lauga, J., and Aulagnier, S. (1996) Application of neural networks to modelling nonlinear relationships in ecology. Ecological Modelling, 90, 39-52.

Luoto, M. and Hjort, J. (2005) Evaluation of current statistical approaches for predictive geomorphological mapping. Geomorphology, 67, 299-315.

Moisen, G.G. and Frescino, T.S. (2002) Comparing ve modelling techniques for predicting forest characteristics. Ecological Modelling, 157, 209-225.

Pearson, R.G., Dawson, T.P., Berry, P.M., and Harrison, P.A. (2002) SPECIES: A Spatial Evalua- tion of Climate Impact on the Envelope of Species. Ecological Modelling, 154, 289-300.

Segurado, P. and Ara ujo, M.B. (2004) Evaluation of methods for modelling species probabilities of occurrence. Journal of Biogeography, 31, 1555-1568.

0.13. 5 MDA -混合判别分析

MDA 是基于混合模型的一种分类方法（经过监督的）。是著名的线性判断模型的推广。用混合的正态分布获取每个等级的密度估计。MDA 可用 MDA 包运行。大多数情况下，简单的 Gaussian 模型来模拟一个等级，如 LDA,受限制太多。MDA 是高斯模型的混合体。在最优化尺度过程中，可以用不同的分类模型。R-BIOMOD 利用 mars 来提高模型的预测能力。

Key reference.

Hastie, T., Tibshirani, R and Buja, A. (1994) Flexible Disriminant Analysis by Optimal Scoring, JASA, 1255-1270.

Hastie, T. J., Buja, A., and Tibshirani, R. (1995) Penalized Discriminant Analysis. Annals of Statistics.

Hastie, T. and Tibshirani, R. (1996) Discriminant Analysis by Gaussian Mixtures. JRSSB.

Key references in ecology/biogeography

Manel, D., Dias, J. M., Buckton, S. T. and Ormerod, S. J. (1999) Alternative methods for predicting species distribution: an illustration with Himalayan river birds. Journal of Applied Ecology. 36, 734-747.

0.13.6 MARS- 多元适应回归样条函数（ Multivariate Adaptive Regression Splines）

线性过程的主要假设是，各水平的解释变量的系数保持不变，在时间序列模型中，不随时间变化。MARS 模型提供了非常有用的分析方式，因为其推测模型解释变量参数在等级有不同的最优化值。有为数众多的理论根源在不同的场合与这种可能性一致，包括能源、财政、经济、社会科学和制造业等等。MARS 的方法是 Friedman (1991)年引入的，可以系统的根据解释变量不同的等级，确认和估计模型的参数。模型参数的临界点或阈值取决于样条函数结点，可以认为与分段回归类似。MARS 的优越之处在于，样条函数结点是通过运算自动确定的。此外，变量间的复杂非线性互作也可以确定。MARS 程序在处理大量右手变量（right-hand variables）和低等互作（low-order interaction）时，会显示出非常强大的功能。转换模型的方程，即 X 变量的模型倾角突然改变时，是 MARS 模型的一个特例。MARS 步骤可以在不同情况下，有清晰断点时，确定和拟合模型，这种情况在系数函数的潜在概率密度改变或变量间发生复杂的相互作用时发生。

BIOMOD 使用 Trevor Hastie 和 Robert Tibshirani 编写的 mda 程序包。对于每个预测变量及预测变量间的相互作用，MARS 自动选择所需的平滑度。可以认为，在预测未来情景时，变量选择没有被考虑，但是需要确定互作的最大程度。采用保守的方法时，只有两个水平的互作被包含在 R-BIOMOD 中（也可以方便的更改）。

这里无需设定特定的参数。理解更为深刻的用户可以参见其函数本身。

Key reference.

J. Friedman, \Multivariate Additive Regression Splines". Annals of Statistics, 1991

Key references in ecology/biogeography

Elith, J., Graham, C.H., Anderson, R.P., Dudik, M., Ferrier, S., Guisan, A., Hijmans, R.J., Huettman, F., Leathwick, J.R., Lehmann, A., Li, J., Lohmann, L., Loiselle, B.A., Manion, G., Moritz, C., Naka- mura, M., Nakazawa, Y., Overton, J.M., Peterson, A.T., Phillips, S., Richardson, K., Schachetti Pereira, R., Schapire, R.E., Sober on, J., Williams, S.E., Wisz, M., and Zimmermann, N.E. (2006) Novel methods improve predictions of species' distributions from occurrence data. Ecography, 29, 129-151.

Luoto, M. and Hjort, J. (2005) Evaluation of current statistical approaches for predictive geomorphological mapping. Geomorphology, 67, 299-315.

Moisen, G.G. and Frescino, T.S. (2002) Comparing ve modelling techniques for predicting forest characteristics. Ecological Modelling, 157, 209-225. 83

0.13.7 GBM Generalised Boosting Models (或 Boosting regression trees, BRT)

解释改编自 Greg Ridgeway

Boosting:解释 在 GLM 试图寻解释物种分布于生态预测变量最优化关系时，boosting 方法拟合了大量相对简单的模型，而且这种方法对响应作出的预测结果更为稳健。BIOMOD

中用到的算法是 boosted regression tree (BRT, Friedman 2001, Ridgeway 1999)。每个模型都包括一个简单的分类或回归树，例如，将预测变量空间基于一定的分类法则，依照响应变量，递归的划分成尽量同质的组。采用一个解释变量的简单规则递归将数据分组，从而得到回归树。每次划分组，数据都将划分为两个互不兼容的组，每个组内尽量同质。普通的广义线性模型具有如下形式：算法通过各种优化程序估计 j(最常用的是极大似然估计)。而基于广义线性模型的扩展的特殊情形，如广义相加模型（GAM），也采用相同的形式：h(x)是非参数函数（如:函数条）。这种方法解决了 hjs，之后利用标准程序（常规的最小二乘回归 OLS）寻找 j。回归树同样具有这种形式，其中 hjs 表示指示函数的 x 落入特定的盒子（box）内，而？作为末端节点的平均值。回归树不会提前选择 hjs 或 J,它们的值是通过递归的划分算法估计到的。GBM 利用回归树的形式给出 hj。拟合程度是不断增加的，因此 hi(x)是一个最优化树，h2(x)是预测 h1(x)残差的最优化树，以此类推（Friedman, et al. 2000）。通过这种方式，BRT 在完善最终模型时采用了一种迭代的方法，对树进行逐步叠加，在对数据的权重重新计算时，强调了之前表现差的树。

在 BIOMOD 中，用户可以指定交叉验证的次数，来最优化树的数目，从而在预测新的、独立点时，将模型的预测准确性最大化，避免不必要的模型复杂性。用户同时需要指定树拟合的最大数目。没有一个在运算前确定树数目的方法。2000 到 5000 之间是一个好的折中方案。更为重要的是，BRT 可以获得模型中每个变量的相对重要值。BIOMOD 应用置换的方法，将每个预测变量随机置换，计算相应的模型预测能力的变化。

详情参见：

http://www.salford-systems.com/friedmankdd.php

www.i-pensieri.com/gregr/ModernPrediction/L9boosting.pdf

BIOMOD 应用 Greg Ridgeway 编写的 gbm 程序包。该包可以运行 generalized boosted modelling 的工作。这种算法与 Friedman 的 Gradient Boosting Machine (Friedman, 2001)接近。interaction depth 和 learing rate 分别设置为 4 和 0.001（可以方便的更改）。

Key reference.

Friedman, J.H. (2001) Greedy function approximation: A gradient boosting machine. Annals of Statistics, 29, 1189-1232.

Friedman, J.H., Hastie, T.J., and Tibshirani, R. (2000) Additive logistic regression: a statistical view of boosting. Annals of Statistics, 28, 337-374.

Ridgeway, G. (1999) The state of boosting. Computing Science and Statistics, 31, 172-181.

Key references in ecology/biogeography

Elith, J., Graham, C. H., Anderson, R. P., Dudik, M., Ferrier, S., Guisan, A., Hijmans, R. J., Huettman, F., Leathwick, J. R., Lehmann, A., Li, J., Lohmann, L., Loiselle, B. A., Manion, G., Moritz, C., Nakamura, M., Nakazawa, Y., Overton, J. M., Peterson, A. T., Phillips, S., Richardson,

K., Schachetti Pereira, R., Schapire, R. E., Sober on, J., Williams, S. E., Wisz, M. and Zimmermann,

N. E. (2006) Novel methods improve predictions of species' distributions from occurrence data. Ecography. 29, 129-151.

Leathwick, J.R., Elith, J., Francis, M.P., Hastie, T.J., and Taylor, P. (2006) Variation in demersal sh species richness in the oceans surroundings New Zealand: an analysis using boosted regression trees. Marine Ecology Progress Series, In press.

Thuiller, W., Midgley, G.F., Rouget, M., and Cowling, R.M. (2006) Predicting patterns of plant species richness in megadiverse South Africa. Ecography, 29, 733-744

0.13.8 随机森林- Breiman 和 Cutler 用于分类和回归的随机森林

randomForest 模型应用 Breiman 的随机森林算法（基于 Breiman 和 Cutler 的初始 Fortran 代码）用来进行分类和回归。该算法通过调用由 Andy Liaw 和 Matthew Wiener 编写的 random Forest 包实现。

随机森林是通过对大量的分类树的计算得到的。在输入向量中对一个新对象分类，将输入向量放到每个树下。每棵树将给出其分类信息，所有的树对其分类信息进行打分。随机森林将在所有树中搜寻得分最高的打分结果。

每棵树的建立如下所示：

如果给出训练集合的情形为 N,从原始数据开始，将样本的 N 个情形可替换的随机给出。这个样本就是建树的训练数据。如果有 M 个待输入的变量，令 m<<M，使 m 对于每个节点，在 M 中随机选取 m 个变量，对节点分组。随着森林的生长，m 值为常数。每棵树尽最大可能性的生长。没有剪枝过程。

在随机森林的原始文献中，随机森林的误差取决于两个因素：

森林中两棵树的相关性。随着相关性的增大，森林的误差也增大。

森林中每棵树的强度。误差率低的树是更好的分类结果。每棵树的分类能力的增加降低了森林误差率。

m 的减小同时减小树的相关性，但是也减少其强度。m 增加的增加也同时增加两者。两者之间 m 最优的取值范围通常相当宽泛。利用 oob 误差率（见下文）：可以快速找到 m 值。这只是随机森林敏感的参数，是可以调整的。

随机森林的特征

可以有效的处理大量数据。

在不去除变量的情况下，处理数以千记的输入变量。

给出变量在分类中的重要值估计

在森林建立过程中生成一个内部无偏的总体误差估计

对于样本量不等的数据集，可以平衡误差。

提供了一种检验变量相互作用的实验方法。

随机森林如何工作？

为了理解和应用各种选项，进一步了解随机森林是如何计算的是非常有用的信息。取决两个数据对象的选项，是随机森林产生的。当当前树的训练数据有放回的抽样时，约 1/3 的样本没有被抽到。这些 oob（out of bag）数据在树被添加到随机森林时，用来得到分类误差的无偏估计。 同时用来估算变量的重要值。

The out of bag (oob) 误差估计：

在随机森林中，欲获得检验数据误差的无偏估计时无需进行交叉验证或者进行外部的检验。检验是在内部完成的，即在运算过程中完成的，过程如下：每棵树用原始数据生成不同的 bootstrap 样本。大约 1/3 的树去除 bootstrap 样本，也将不会用到数以千记的树的建立中。将每个剩下来的树放到数以千记的树中，获得分类信息。通过这种方法，1/3 的树的每一颗都将获得其分类信息。在运行结束后，将每次 n 为 oob 的 j 设定为级别数。所有的情况 j 中与真实级别 n 不等次数的比例，就是 oob 误差估计。

变量重要性：

在森林中生长的每棵树，记下 oob 的次数，记录其分类正确的次数。将 oobs 的树变量 m 随机置换，记录次数。提取其 m 变量置换的 oob 数据在原始 oob 数据中正确分类的得分。森林中所有树的平均值是变量 m 重要值的得分。如果该得分在树与树之间是独立的，则用标准计算方法可以计算其标准差。通过对一定数量的数据集计算树间的相关性得分很低，因此，我们用经典的方法计算标准误，将其得分除以标准误，以得到 z 值，在假设正态的前提下，得到 z 值的显著性。对每种情况，考虑 oob 所有树。提取置换变量 m 的 oob 数据在原始 oob 数据总正确分类得分的百分率。

BIOMOD 采用 500 棵树（在 Models 函数中可以直接更改），藉此提取每个变量的重要值。

Key References. Breiman, L. (2001), Random Forests, Machine Learning 45(1), 5-32. Breiman,

L (2002), \Manual On Setting Up, Using, And Understanding Random Forests V3.1.

Key References in ecology/biogeography.

Elith, J., Graham, C.H., Anderson, R.P., Dudik, M., Ferrier, S., Guisan, A., Hijmans, R.J., Huettman, F., Leathwick, J.R., Lehmann, A., Li, J., Lohmann, L., Loiselle, B.A., Manion, G., Moritz, C., Nakamura, M., Nakazawa, Y., Overton, J.M., Peterson, A.T., Phillips, S., Richardson, K., Schachetti Pereira, R., Schapire, R.E., Sober on, J., Williams, S.E., Wisz, M., and Zimmermann, N.E. (2006) Novel methods improve predictions of species' distributions from occurrence data. Ecography, 29, 129-151.

Prasad, A.M., Iverson, L.R., and Liaw, A. (2006) Newer classi cation and regression tree techniques: bagging and random forests for ecological prediction. Ecosystems, 9, 181-199.

0.13.9 SRE 表面分布区分室模型

这是简单那的表面分布区分室模型，与 BioClim 模型类似。分室模型可以通过物种出现记录点每个变量的最大和最小值来确定。每个包含所有位于这些最大值和最小值之间的变量被包括其中。这是模拟物种分布或生物群系的最简单方法。Perc025 和 Per05 允许对所选的预测变量设定更宽的百分率。它可以将异常值即物种分布的极端值移除（在分室外缘附近的值）。

Key reference.

Busby JR (1991) BIOCLIM - a bioclimate analysis and prediction system. In: Margules CR, Austin MP, editors. Nature Conservation: Cost E
ective Biological Surveys and Data Analysis. Canberra, Australia: CSIRO. pp. 64-68.

Key References in ecology/biogeography.

Beaumont LJ and Hughes L (2002) Potential changes in the distribution of latitudinally restricted Australian butter
y species in response to climate change. Global Change Biology 8:954-971.

0.14 预测成效的表示

BIOMOD 采用三种评价步骤，分别成为 ROC 曲线，True Skill Statistic 和 Kappa statistic。每一个都可以单独运算，但是建议同时运算所有的以便进行比较。

统计模型经常可以用实际与预测的 confusion 矩阵的一致性程度进行评估。可以从该矩阵推断出四部分：

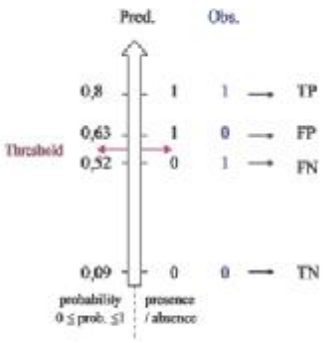
-敏感性 （实际正值部分）

-特异性 （实际负值部分）

-假正值部分

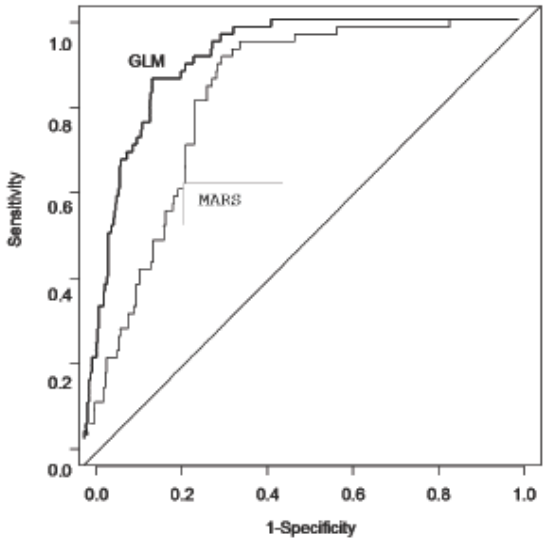
-假负值部分

在样本中，敏感性可以描述通过样本中正值的点与正确预测的正值点（存在）的比例。特异性是样本负值点与正确预测的负值点（不存杂）的比例。假阳性或假阴性分别为 1-特异性和 1-敏感性。为了生成这样的模型，由于多种模型给出的结果是物种在当地存在的概率，必须要设定一个概率的阈值以确定一个点或一个网格是否会被占据或没有占据。



		Predicted		
		Absence	Presence	Total
Observed	Absence	TN	FP	TN + FP
	Presence	FN	TP	FN + TP
	Total	TN + FN	FP + TP	

相对操作特征曲线（Relative operating characteristic curve ）：这个曲线并不取决于阈值。ROC 曲线是应用多个阈值表示假阳性比例（1-特异性）和敏感性关系的图形方法。如果所有的预测都是随机的，关系为 45 度。好的模型预测是将带有低的（1-特异性值）的最大敏感性曲线刻画的。例如，当曲线经过图形的左上角时。45 度线和曲线之间的面积作为判断的标志，也就是说，是判断否用模型对某一种的出现或不出现进行了正确的划分。曲线下的面积称为（AUC）。在下面的例子中，GLM 的得分高于 MARS,从而认为更为可信。



Cohen's Kappa： 该统计量表示两个定性变量非随机符合的程度（二元数据是其一种特殊形式）。Kappa 是基于错误分类矩阵，此时需要计算概率阈值。为此，BIOMOD 计算了 01 之间所有阈值。Kappa 值越高，表明结果越好。此时描述了最可能符合的程度。

The Hanssen-Kuiper Skill Score(KSS)或 True skill statistics (TSS): 该统计量传统上被用在天气预报的准确性评估上，通过比较将天气预报中正确的预报次数减去随机猜想的次数与最优化预报作比较。

对于 2*2 的 confusion 矩阵，TSS 可以定义为

TSS=sensitivity + specificity -1

与 Kappa 相似，TSS 考虑到了遗漏平均误差，随机猜想的成功可能性。该值位于-1 到+1 之间，+1 表明符合的极好，值为 0 或更小表示结果不比随机要好。但是，与 Kappa 相比，TSS 不会受到普遍性的影响。TSS 也不受到校正数据样本量的影响，两种表现一致的方法具有相同的 TSS 得分。TSS 是校正数据中给定 presence 和 absence 的比例的 Kappa 的特殊形式。

模型预测准确性的分类等级：

Accuracy	AUC	Kappa/TSS
Excellent or high	0.9 – 1	0.8 – 1
Good	0.8 – 0.9	0.6 – 0.8
Fair	0.7 – 0.8	0.4 – 0.6
Poor	0.6 – 0.7	0.2 – 0.4
Fail or null	0.5 – 0.6	0 – 0.2