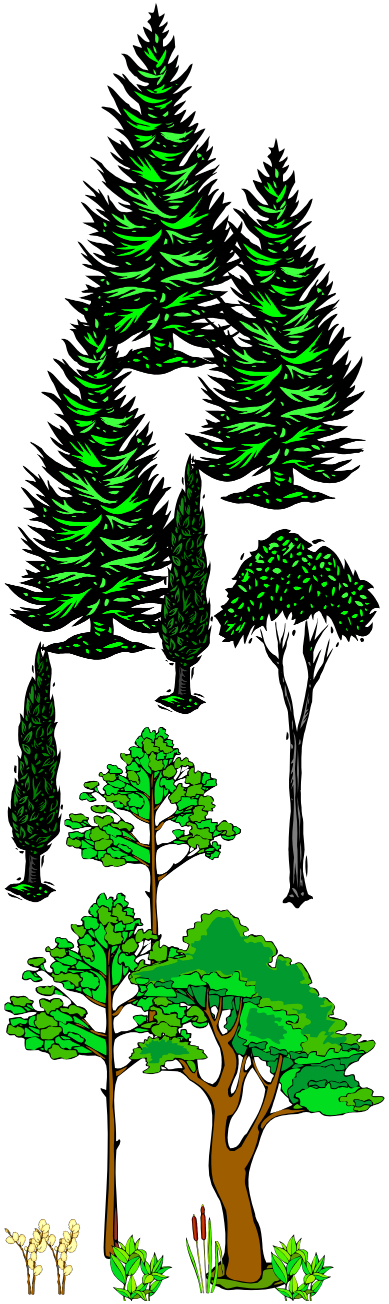


用R实现随机森林的分类与回归
Applications of Random Forest using R
Classification and Regression

李欣海

中科院动物所

邮件：lixh@ioz.ac.cn
主页：http://people.gucas.ac.cn/~LiXinhai
博客：http://blog.sciencenet.cn/u/lixinhai
微博：http://weibo.com/lixinhaiblog

Random Forest

http://jfbelisle.com/2011/03/an-introduction-to-data-mining-for-marketing-and-business-intelligence/

- Random Forest is an ensemble classifier that consists of many decision trees.
- It outputs the class that is the mode of the class's output by individual trees (Breiman 2001).
- It deals with “small n large p”-problems, high-order interactions, correlated predictor variables.

Breiman, L. 2001. Random forests. Machine Learning 45:5-32. (Being cited 6500 times until 2012)

http://jfbelisle.com/2011/03/an-introduction-to-data-mining-for-marketing-and-business-intelligence/

History

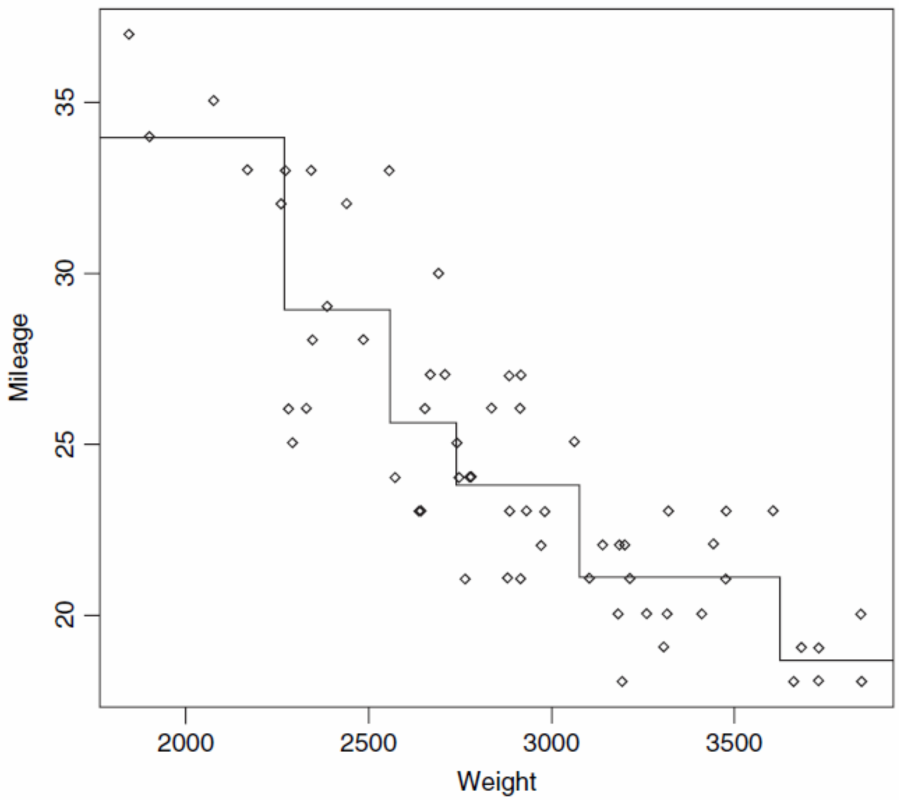
The algorithm for inducing a random forest was developed by Leo Breiman (2001) and Adele Cutler, and "Random Forests" is their trademark.

The term came from **random decision forests** that was first proposed by Tin Kam Ho of Bell Labs in 1995.

The method combines Breiman's "bagging" idea and the random selection of features, introduced independently by Ho (1995) and Amit and Geman (1997) in order to construct a collection of decision trees with controlled variation.

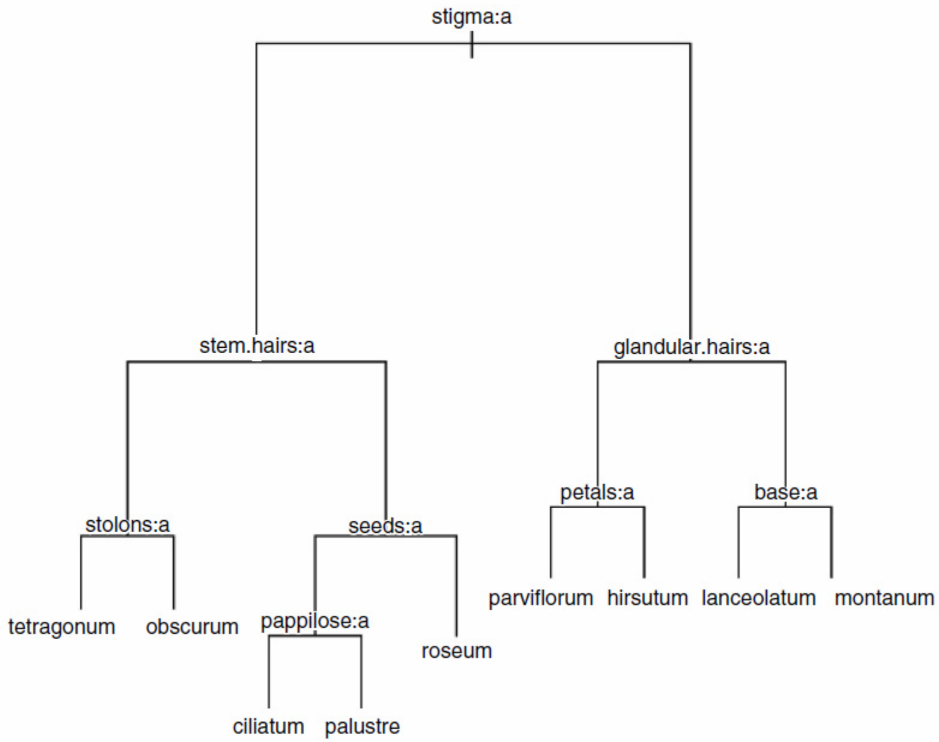
Tree models

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \varepsilon_i$$



Regression tree

(Crawley 2007 *The R Book* p691)



Classification tree

(Crawley 2007 *The R Book* p694)

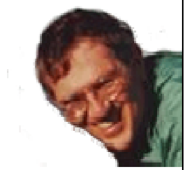
The statistical community uses irrelevant theory, questionable conclusions?


David R. Cox


Brad Efron

NO


Emanuel Parzen


Bruce Hoadley

YES

Statistical Science
2001, Vol. 16, No. 3, 199–231

Statistical Modeling: The Two Cultures

Leo Breiman

Abstract. There are two cultures in the use of statistical modeling to reach conclusions from data. One assumes that the data are generated by a given stochastic data model. The other uses algorithmic models and treats the data mechanism as unknown. The statistical community has been committed to the almost exclusive use of data models. This commitment has led to irrelevant theory, questionable conclusions, and has kept statisticians from working on a large range of interesting current problems. Algorithmic modeling, both in theory and practice, has developed rapidly in fields outside statistics. It can be used both on large complex data sets and as a more accurate and informative alternative to data modeling on smaller data sets. If our goal as a field is to use data to solve problems, then we need to move away from exclusive dependence on data models and adopt a more diverse set of tools.

http://med.stanford.edu/profiles/Trevor_Hastie/

Ensemble classifiers

Tree models are simple, often produce noisy (bushy) or weak (stunted) classifiers.

- Bagging (Breiman, 1996): Fit many large trees to bootstrap-resampled versions of the training data, and classify by majority vote.
- Boosting (Freund & Shapire, 1996): Fit many large or small trees to reweighted versions of the training data. Classify by weighted majority vote.
- Random Forests (Breiman 1999): Fancier version of bagging.

In general Boosting > Random Forests > Bagging > Single Tree (Trevor Hastie).

http://med.stanford.edu/profiles/Trevor_Hastie/

How Random Forest Works

- At each tree split, a random sample of m features is drawn, and only those m features are considered for splitting. Typically m = sqrt(p) or log(p), where p is the number of features.
- For each tree grown on a bootstrap sample, the error rate for observations left out of the bootstrap sample is monitored. This is called the out-of-bag (OOB) error rate.
- Random forests tries to improve on bagging by “de-correlating” the trees. Each tree has the same expectation.

(Trevor Hastie, p21 in *Trees, Bagging, Random Forests and Boosting*)

R Packages

randomForest **randomForest()**

Title: Breiman and Cutler's random forests for classification and regression
Version: 4.6-6
Date: 2012-01-06
Author: Fortran original by Leo Breiman and Adele Cutler, R port by Andy Liaw and Matthew Wiener.

Implementation based on CART trees for variables of different types.
Biased in favor of continuous variables and variables with many categories.

party **cforest()**
Based on unbiased conditional inference trees.
For variables of different types: unbiased when subsampling.

朱鹮的分布

Data

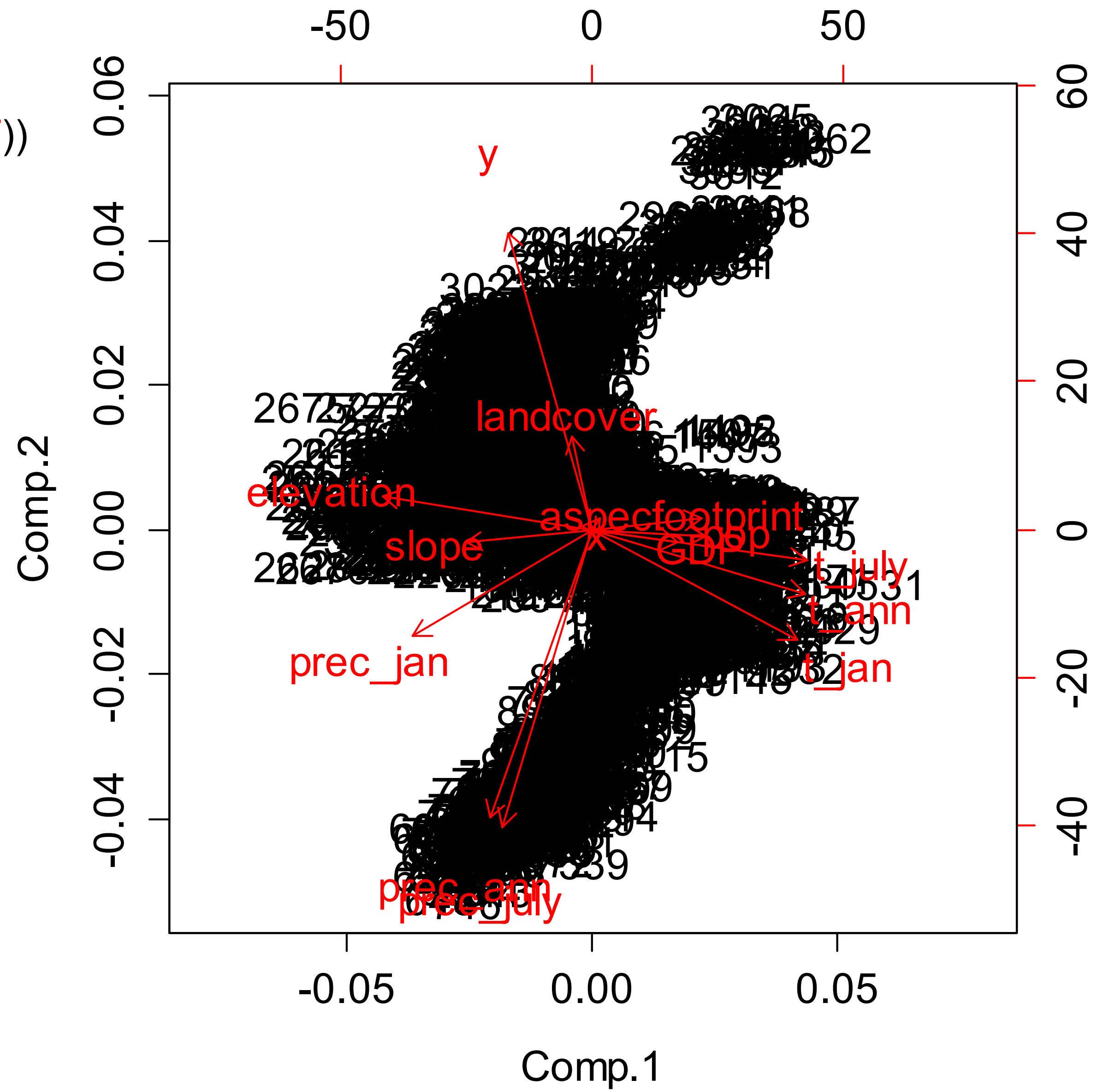
use	x	y	Elev	Aspect	Slope	Land cover	Pop	Foot print	GDP	prec_ ann	prec_ jan	prec_ july	t_ann	t_jan	t_july	year	Nest site
1	107.505	33.392	984	0.67	29.6	21	42.0	20	2.95	845	6	153	12.4	0.3	24.0	1981	金家河
1	107.548	33.409	1315	0.90	19.0	14	22.5	26	1.97	869	6	157	11.3	-0.6	22.7	1981	姚家沟
1	107.505	33.392	984	0.67	29.6	21	42.0	20	2.95	845	6	153	12.4	0.3	24.0	1982	金家河
1	107.548	33.409	1315	0.90	19.0	14	22.5	26	1.97	869	6	157	11.3	-0.6	22.7	1982	姚家沟
1	107.505	33.392	984	0.67	29.6	21	42.0	20	2.95	845	6	153	12.4	0.3	24.0	1983	金家河
1	107.548	33.409	1315	0.90	19.0	14	22.5	26	1.97	869	6	157	11.3	-0.6	22.7	1983	姚家沟
1	107.548	33.409	1315	0.90	19.0	14	22.5	26	1.97	869	6	157	11.3	-0.6	22.7	1984	姚家沟
1	107.405	33.406	1056	0.54	11.4	21	0.0	20	0.98	892	7	161	11.4	-0.5	22.9	1984	三岔河
1	107.405	33.406	1056	0.54	11.4	21	0.0	20	0.98	892	7	161	11.4	-0.5	22.9	1985	三岔河
1	107.548	33.409	1315	0.90	19.0	14	22.5	26	1.97	869	6	157	11.3	-0.6	22.7	1985	姚家沟
0	107.400	32.780	980	0.46	42.1	11	45.8	14	1.78	927	6	170	13.0	1.3	24.0	0	303
0	107.430	32.780	1553	0.97	29.6	14	171.8	32	4.76	887	5	162	13.0	1.3	24.0	0	304
0	107.460	32.780	1534	0.51	25.7	14	12.7	14	1.78	886	5	162	14.0	2.15	25.2	0	305
0	107.490	32.780	996	0.72	29.4	14	76.1	20	2.97	886	5	162	12.4	0.8	23.4	0	306
0	107.520	32.780	1144	0.16	9.3	14	29.3	20	1.78	956	6	175	12.4	0.8	23.4	0	307
0	107.550	32.780	915	0.91	20.7	11	214.7	20	5.95	956	6	175	11.6	0.15	22.5	0	308
0	107.580	32.780	930	0.13	35.7	22	153.2	29	4.76	993	7	181	11.6	0.15	22.5	0	309
0	107.610	32.780	873	0.40	31.9	11	66.4	29	2.97	931	6	171	12.7	1.1	23.8	0	310
0	107.640	32.780	1147	0.50	35.5	11	46.8	20	2.38	1041	7	189	12.7	1.1	23.8	0	311
0	107.670	32.780	1699	0.89	21.1	14	20.5	20	1.78	1060	8	192	10.4	-0.8	21.2	0	312

```
> table(ibis$use)

   0    1
2538  560
```

```
ibis$use        <- as.factor(ibis$use)
ibis$landcover <- as.factor(ibis$landcover)
```

Multicollinearity is a pain
Variables in the two-principal-component space

```
biplot(princomp(ibis[,2:16], cor=T))
```



Variance inflation factors (VIF) to detect multicollinearity

```
library(car)
vif(lm(pop ~ x+y+elevation+slope+aspect+footprint+GDP+
            prec_ann+prec_jan+prec_july+t_ann+t_jan+t_july,
            data=ibis))
```

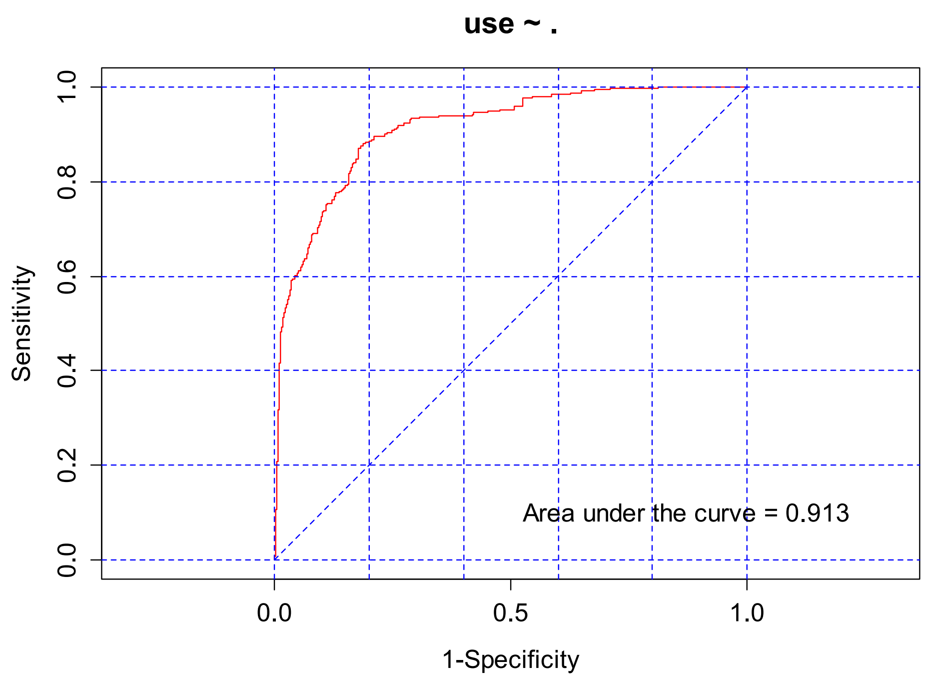
x	6.86
y	82.67
elevation	9.82
aspect	1.01
footprint	1.38
GDP	1.32
prec_ann	207.97
prec_jan	6.24
slope	1.50
t_ann	15170.35
prec_july	170.89
t_jan	4309.17
t_july	5132.72

Logistic regression

```
fit <- glm(use ~ ., data=ibis, family=binomial())
summary(fit)
step(fit)
```

```
Step: AIC=1710.33
use ~ x + y + elevation + aspect + slope +
      landcover + pop + footprint + GDP + prec_ann +
      prec_jan + prec_july + t_ann + t_jan
```

```
library(epicalc)
lroc(fit, title=TRUE,
     auc.coords=c(.5,.1))
```



randomForest()

```
library(randomForest)
ibis <- ibis[,-c(17,18)]
RF <- randomForest(ibis[,-1], ibis[,1],
        prox=TRUE, importance=TRUE)
summary(RF)
```

use	x	y	Elev	Aspect	Slope	Land cover	Pop	Foot print	GDP	prec_ ann	prec_ jan	prec_ july	t_ann	t_jan	t_july	year	Nest site
1	107.505	33.392	984	0.67	29.6	21	42.0	20	2.95	845	6	153	12.4	0.3	24.0	1981	金家河
1	107.548	33.409	1315	0.90	19.0	14	22.5	26	1.97	869	6	157	11.3	-0.6	22.7	1981	姚家沟
1	107.505	33.392	984	0.67	29.6	21	42.0	20	2.95	845	6	153	12.4	0.3	24.0	1982	金家河
1	107.548	33.409	1315	0.90	19.0	14	22.5	26	1.97	869	6	157	11.3	-0.6	22.7	1982	姚家沟
1	107.505	33.392	984	0.67	29.6	21	42.0	20	2.95	845	6	153	12.4	0.3	24.0	1983	金家河
1	107.548	33.409	1315	0.90	19.0	14	22.5	26	1.97	869	6	157	11.3	-0.6	22.7	1983	姚家沟
1	107.548	33.409	1315	0.90	19.0	14	22.5	26	1.97	869	6	157	11.3	-0.6	22.7	1984	姚家沟
1	107.405	33.406	1056	0.54	11.4	21	0.0	20	0.98	892	7	161	11.4	-0.5	22.9	1984	三岔河
1	107.405	33.406	1056	0.54	11.4	21	0.0	20	0.98	892	7	161	11.4	-0.5	22.9	1985	三岔河
1	107.548	33.409	1315	0.90	19.0	14	22.5	26	1.97	869	6	157	11.3	-0.6	22.7	1985	姚家沟
0	107.400	32.780	980	0.46	42.1	11	45.8	14	1.78	927	6	170	13.0	1.3	24.0	0	303
0	107.430	32.780	1553	0.97	29.6	14	171.8	32	4.76	887	5	162	13.0	1.3	24.0	0	304
0	107.460	32.780	1534	0.51	25.7	14	12.7	14	1.78	886	5	162	14.0	2.15	25.2	0	305
0	107.490	32.780	996	0.72	29.4	14	76.1	20	2.97	886	5	162	12.4	0.8	23.4	0	306
0	107.520	32.780	1144	0.16	9.3	14	29.3	20	1.78	956	6	175	12.4	0.8	23.4	0	307
0	107.550	32.780	915	0.91	20.7	11	214.7	20	5.95	956	6	175	11.6	0.15	22.5	0	308
0	107.580	32.780	930	0.13	35.7	22	153.2	29	4.76	993	7	181	11.6	0.15	22.5	0	309
0	107.610	32.780	873	0.40	31.9	11	66.4	29	2.97	931	6	171	12.7	1.1	23.8	0	310
0	107.640	32.780	1147	0.50	35.5	11	46.8	20	2.38	1041	7	189	12.7	1.1	23.8	0	311
0	107.670	32.780	1699	0.89	21.1	14	20.5	20	1.78	1060	8	192	10.4	-0.8	21.2	0	312

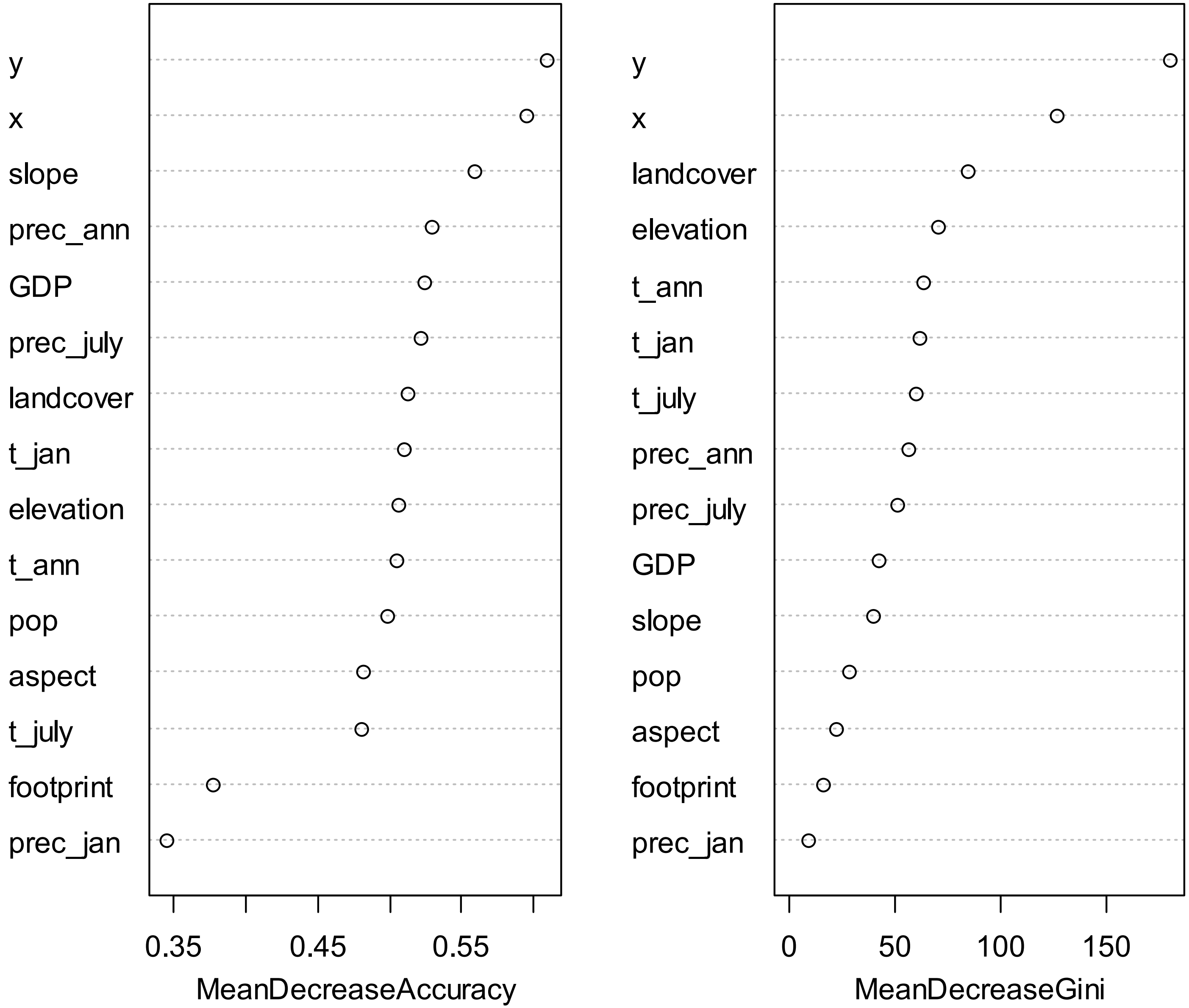
```
                Length  Class     Mode
call                  5 -none-    call
type                  1 -none-    character
predicted          3098 factor    numeric
err.rate           1500 -none-    numeric
confusion             6 -none-    numeric
votes              6196 matrix    numeric
oob.times          3098 -none-    numeric
classes               2 -none-    character
importance           60 -none-    numeric
importanceSD         45 -none-    numeric
localImportance       0 -none-    NULL
proximity       9597604 -none-    numeric
ntree                 1 -none-    numeric
mtry                  1 -none-    numeric
forest               14 -none-    list
y                  3098 factor    numeric
test                  0 -none-    NULL
inbag                 0 -none-    NULL
```

importance(RF)

```
imp <- importance(RF)
impvar <- imp[order(imp[, 3], decreasing=TRUE),]; impvar
```

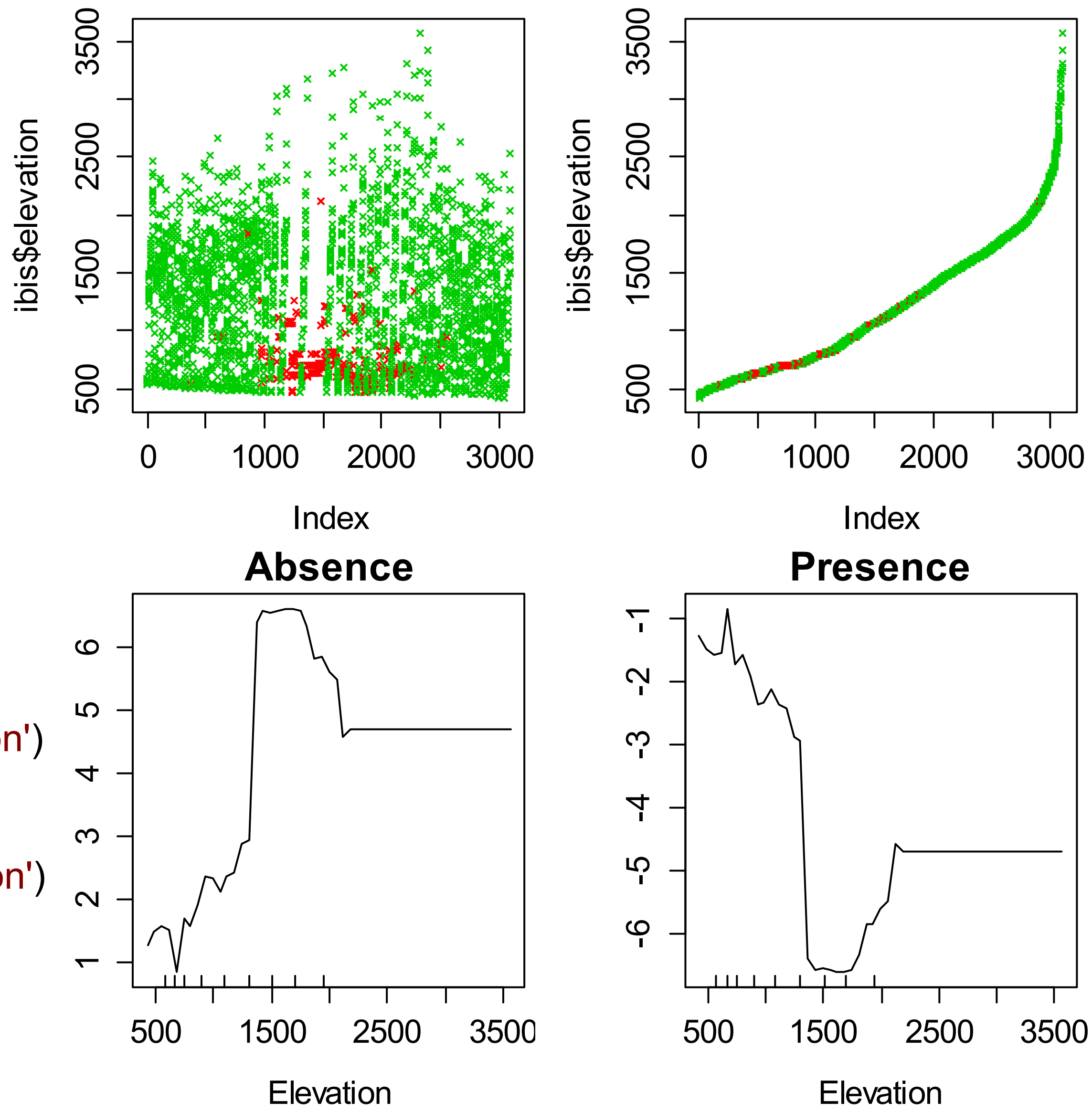
	0	1	MeanDecreaseAccuracy	MeanDecreaseGini
y	0.544507	1.3917066	0.6081245	179.819331
x	0.455705	1.3866419	0.594095	126.056688
slope	0.169391	1.2966797	0.5587788	39.304633
prec_ann	0.324172	1.2131164	0.5293486	55.937209
GDP	0.393354	1.1838283	0.5233282	42.083964
prec_july	0.366197	1.166394	0.521239	51.248009
landcover	0.250204	1.1888627	0.5125181	84.391889
t_jan	0.406161	1.1651366	0.5096002	61.646713
elevation	0.380496	1.1895061	0.5053193	69.867269
t_ann	0.383974	1.150548	0.5048529	63.471821
pop	0.423507	1.0874984	0.4981203	28.297228
aspect	-0.0799	1.1987937	0.4809901	22.272998
t_july	0.352301	1.1058594	0.4796287	59.880941
footprint	-0.05375	0.9224068	0.3773902	15.344756
prec_jan	0.228913	0.8092592	0.344224	8.766087

varImpPlot(RF)



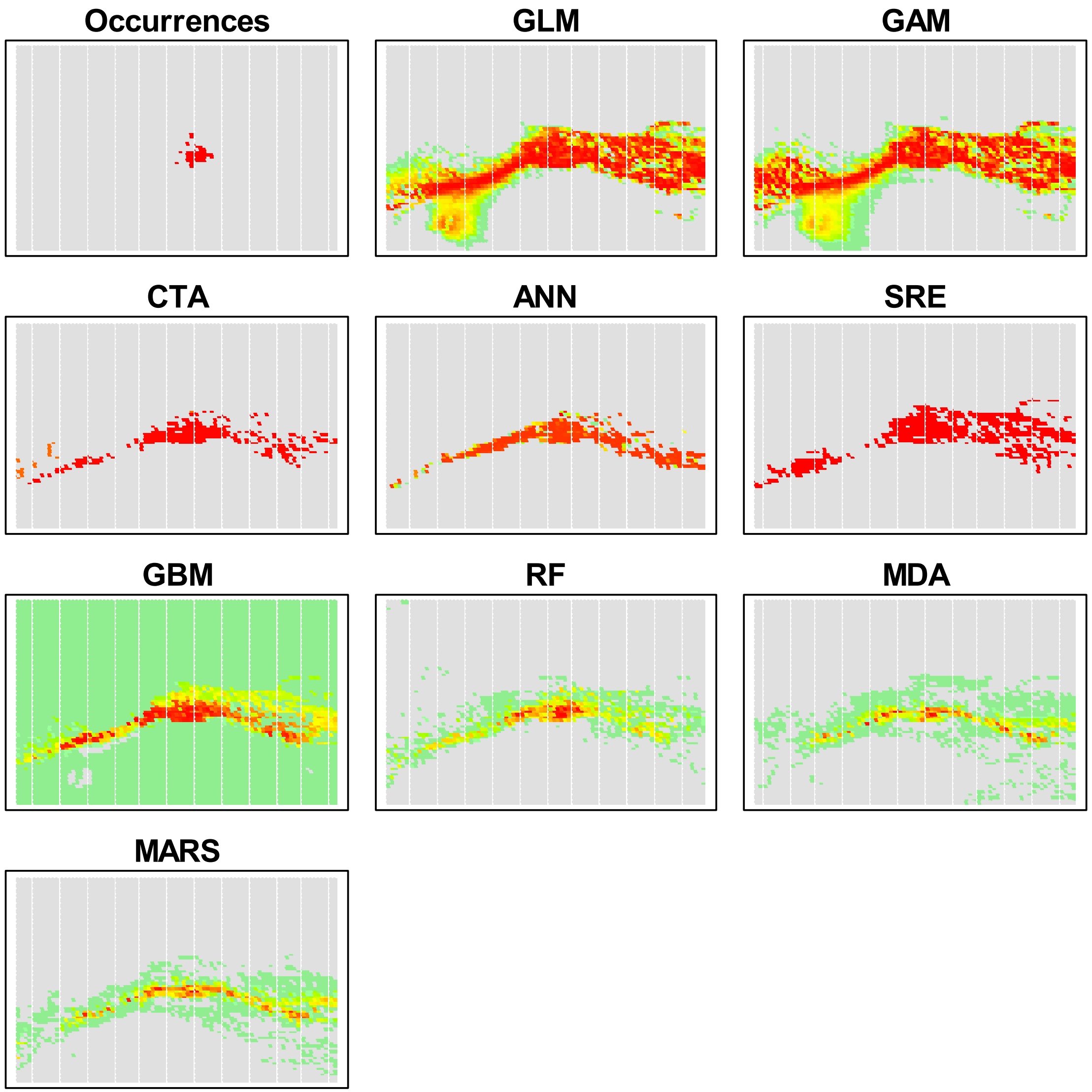
partialPlot: partial dependence of elevation

```
ibis=ibis[order(ibis$x),]
plot(ibis$elevation,
     col=4-as.numeric(ibis$use),
     cex=0.5, pch=4)

ibis=ibis[order(ibis$elevation),]
plot(ibis$elevation,
     col=4-as.numeric(ibis$use),
     cex=0.5, pch=4)
```

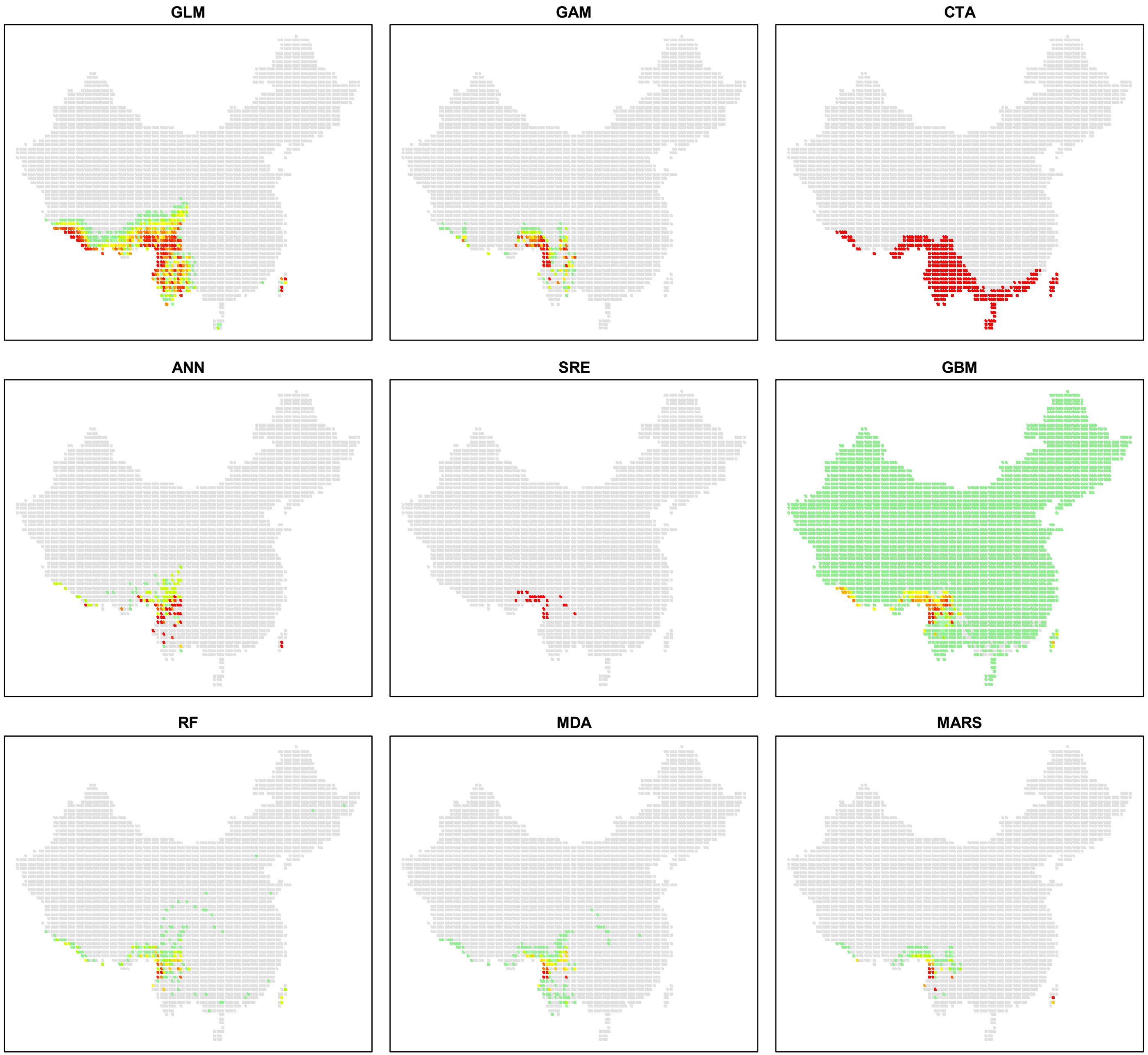



```
partialPlot(RF, ibis, elevation, "0",
            main='Absence',xlab='Elevation')

partialPlot(RF, ibis, elevation, "1",
            main='Presence',xlab='Elevation')
```

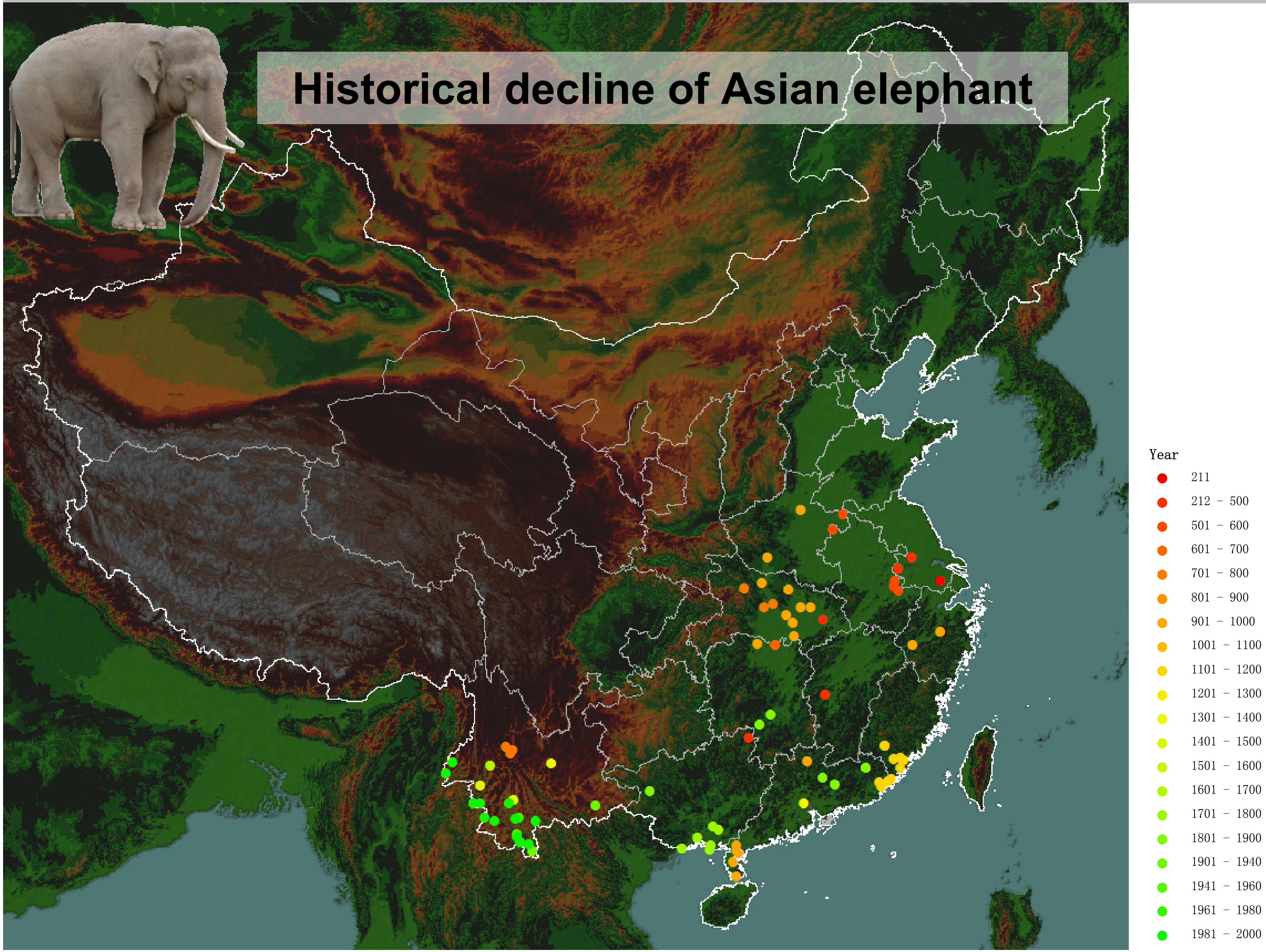
Model comparison



Predicted current suitable habitat of crested ibis using the models in BIOMOD
(The warm color areas are the suitable areas)



Predicted current suitable habitat of black snub-nose monkey using BIOMOD (The warm color areas are the suitable areas)

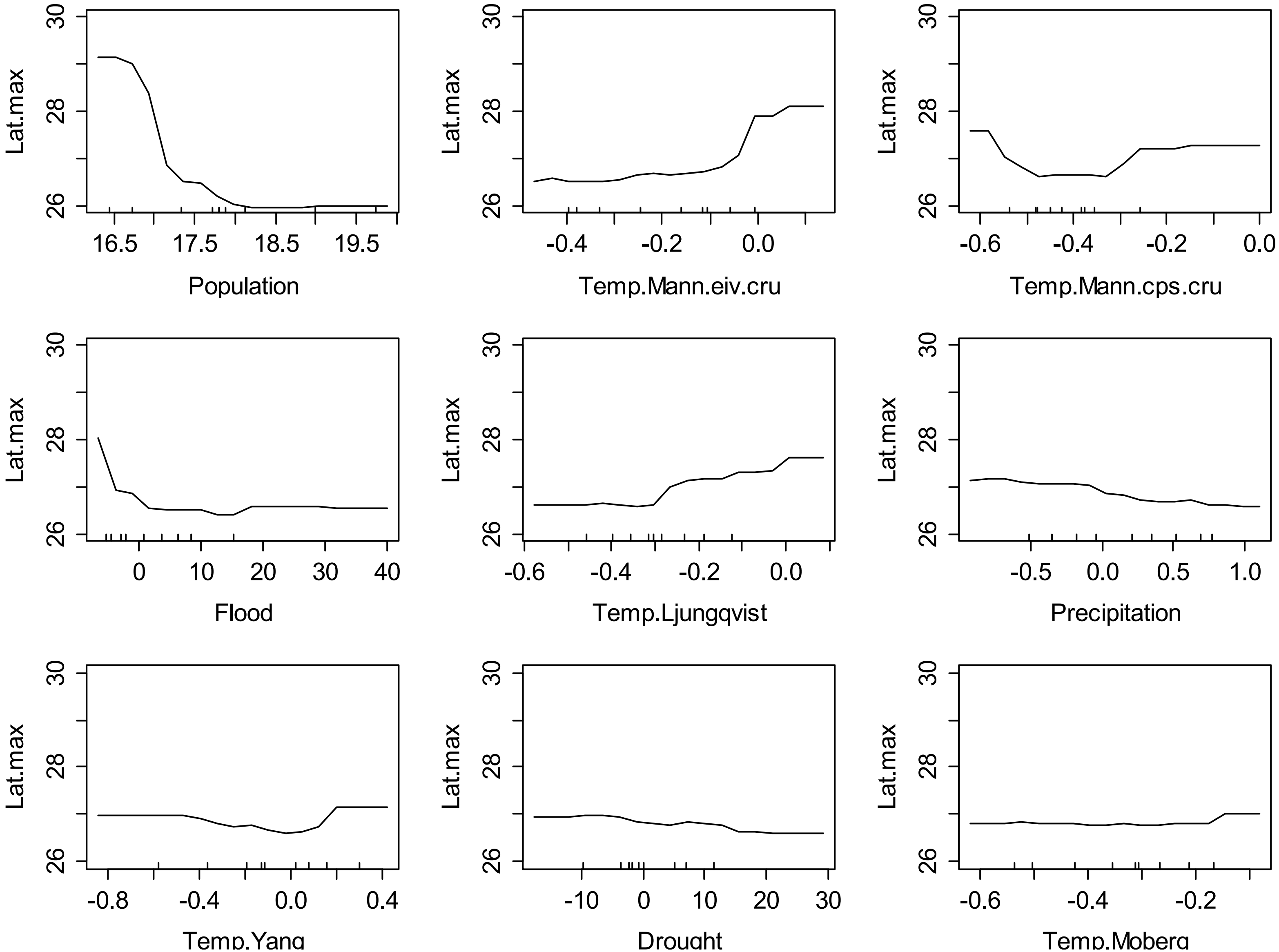

Historical decline of Asian elephant
Year
211
212 – 500
501 – 600
601 – 700
701 – 800
801 – 900
901 – 1000
1001 – 1100
1101 – 1200
1201 – 1300
1301 – 1400
1401 – 1500
1501 – 1600
1601 – 1700
1701 – 1800
1801 – 1900
1901 – 1940
1941 – 1960
1961 – 1980
1981 – 2000

Variables associated with species range

lat.max	Temp. Yang	Temp. Ljungqvist	Temp. Moberg	Temp. Mann.cps	Temp. Mann.eiv	Precipitation	Drought	Flood	Population
34.42	-0.52	-0.41	-0.49	-0.59	-0.22	-0.07	-0.10	-5.32	16.44
31.57	-0.41	-0.21	-0.35	0.00	-0.02	-0.35	-2.12	-6.37	16.44
29.53	0.28	-0.23	-0.25	-0.21	-0.11	-0.78	-3.76	-2.19	17.74
31.87	0.35	-0.25	-0.29	-0.27	0.02	-0.34	-3.24	-5.07	16.31
28.93	-0.13	-0.09	-0.31	-0.62	-0.16	0.11	-2.38	-1.79	16.93
34.79	0.19	0.09	-0.13	-0.38	0.14	0.67	7.30	22.57	16.60

```
RF <- randomForest(lat.max ~ ., data=dd, ntree=1000, importance=TRUE)
imp <- importance(RF)
impvar <- rownames(imp)[order(imp[, 1], decreasing=TRUE)] #sort importance

# Plot partial effects
op <- par(mfrow=c(3, 3),mar=c(4,4,2,2))
for (i in seq_along(impvar)) {
    partialPlot(RF, dd, impvar[i], xlab=impvar[i], #Partial effects
    ylab='Longitude', ylim=c(26,30), main='')
}
```

Partial effect of variables on maximum latitude



Popular species distribution models, history, complexity levels, popularity in climate change studies, types of species data, and reference papers

Models	History	Complexity	Popularity	Species data‡	Reference
Generalized linear model (GLM)	1972	low	33/66	p/a or abundance	(Nelder & Wedderburn 1972)
Generalized additive model (GAM)	1986	medium	28/86	p/a or abundance	(Hastie & Tibshirani 1986)
Multivariate Adaptive Regression Splines (MARS)	1991	medium	13/56*	p/a	(Friedman 1991)
Mixture discriminant analysis (MDA)	1996	medium	4/9	p/a	(Hastie & Tibshirani 1996)
Classification and Regression Tree (CART)	1984	medium	16/23	p/a	(Breiman *et al.* 1984)
Generalized Boosting Models (GBM)	1999	medium	0/14	p/a	(Friedman *et al.* 2000)
Random Forest	1995	high	26	p/a	(Breiman 2001a)
Artificial neural networks (ANN)	1943	high	96/75	p/a	(Hopfield 1982)
Genetic Algorithm for Rule Set Production (GARP)	1999	high	3/48	p	(Stockwell & Peters 1999)
Maximum entropy method (Maxent)	2006	high	5/125	p	(Phillips *et al.* 2006)
Hierarchical modeling	1996	low	13	p/a or abundance	(Wikle 2003)

WIKIPEDIA
The Free Encyclopedia

The advantages of Random Forest

- For many data sets, it produces a highly accurate classifier
- It handles a very large number of input variables
- It estimates the importance of variables in determining classification
- It generates an internal unbiased estimate of the generalization error as the forest building progresses
- It includes a good method for estimating missing data and maintains accuracy when a large proportion of the data are missing
- It provides an experimental way to detect variable interactions
- It can balance error in class population unbalanced data sets
- It computes proximities between cases, useful for clustering, detecting outliers, and (by scaling) visualizing the data
- Using the above, it can be extended to unlabeled data, leading to unsupervised clustering, outlier detection and data views
- Learning is fast

The disadvantages of Random Forest

- Random forests are prone to overfitting for some datasets. This is even more pronounced in noisy classification/regression tasks.
- Random forests do not handle large numbers of irrelevant features as well as ensembles of entropy-reducing decision trees.
- It is more efficient to select a random decision boundary than an entropy-reducing decision boundary, thus making larger ensembles more feasible. Although this may seem to be an advantage at first, it has the effect of shifting the computation from training time to evaluation time, which is actually a disadvantage for most applications.


Try Random Forest!
Phone took at Hailuogou, Sichuan
province in June 2007 by Xinhai Li