

AICcTab Function

```
AICcTab(models,nobs=10,mnames=c("Null","Spatial","Seven","Spatial*Seven","Global"),base=T,weight
s=T,delta=T,logLik=T)
```

- Function needs
 - List of models
 - * List_of_models<-list(model1,model2,model3,model4...)
 - nobs – number of observations
 - mnames (options) list of model names for the output
 - weight (T/F) – include model weights
 - delta (T/F) – include delta AIC
 - logLik (T/F) – include log likelihood

Models

- For AIC analyses, models can be just about anything
- Often, these are simple linear models (lm function)
 - * First_model<-lm(response_var~predictor1+predictor2)
- For this assignment, your predictor variable is log_resratio
 - Already log transformed
 - Represents each species ratio of stream to reservoir abundance
 - Your assignment is to use AIC to select a parsimonious model predicting this from species life history traits.



Log_resratio=2.56



Log_resratio=0.0002

Specifying Sample Size in AICcTab

```
AICcTab(models,nobs=10,mnames=c("Null","Spatial","Seven","Spatial*Seven","Global"),base=T,weight
s=T,delta=T,logLik=T)
```

	logLik	AICc	dLogLik	dAICc	df	weight
Global	10.7	-67.1	67.7	0.0	16	1
Seven	2.2	5.6	59.3	72.7	3	<0.001
Spatial*Seven	7.6	54.8	64.6	121.9	7	<0.001
Spatial	-25.7	67.5	31.3	134.6	4	<0.001
Null	-57.0	119.8	0.0	186.9	2	<0.001

```
AICcTab(models,nobs=29,mnames=c("Null","Spatial","Seven","Spatial*Seven","Global"),base=T,weight
s=T,delta=T,logLik=T)
```

	logLik	AICc	dLogLik	dAICc	df	weight
Seven	2.2	2.5	59.3	0.0	3	0.69
Spatial*Seven	7.6	4.1	64.6	1.6	7	0.31
Global	10.7	55.9	67.7	53.4	16	<0.001
Spatial	-25.7	61.1	31.3	58.6	4	<0.001
Null	-57.0	118.5	0.0	116.0	2	<0.001

Not the same in AICtab function

```
AICtab(models,nobs=10,mnames=c("Null","Spatial","Seven","Spatial*Seven","Global"),base=T,weights
=T,delta=T,logLik=T)
```

	logLik	AIC	dLogLik	dAIC	df	weight
Spatial*Seven	7.6	-1.2	64.6	0.0	7	0.7984
Seven	2.2	1.6	59.3	2.8	3	0.1994
Global	10.7	10.6	67.7	11.8	16	0.0022
Spatial	-25.7	59.5	31.3	60.7	4	<0.001
Null	-57.0	118.1	0.0	119.3	2	<0.001

```
AICtab(models,nobs=29,mnames=c("Null","Spatial","Seven","Spatial*Seven","Global"),base=T,weights
=T,delta=T,logLik=T)
```

	logLik	AIC	dLogLik	dAIC	df	weight
Spatial*Seven	7.6	-1.2	64.6	0.0	7	0.7984
Seven	2.2	1.6	59.3	2.8	3	0.1994
Global	10.7	10.6	67.7	11.8	16	0.0022
Spatial	-25.7	59.5	31.3	60.7	4	<0.001
Null	-57.0	118.1	0.0	119.3	2	<0.001

If you are using linear models (lm), you may not need to

```
nobs (integer) number of observations: required for type="BIC" or type="AICc" unless objects
have a nobs method

AICcTab(models, nobs=29, mnames=c("Null", "Spatial", "Seven", "Spatial*Seven", "Global"), base=T, weight
s=T, delta=T, logLik=T)

      logLik AICc  dLogLik dAICc df weight
Seven      2.2  2.5  59.3    0.0 3  0.69
Spatial*Seven 7.6  4.1  64.6    1.6 7  0.31
Global     10.7 55.9  67.7   53.4 16 <0.001
Spatial    -25.7 61.1  31.3   58.6 4  <0.001
Null      -57.0 118.5  0.0   116.0 2  <0.001

> nobs(spatial)
[1] 29

AICcTab(models, mnames=c("Null", "Spatial", "Seven", "Spatial*Seven", "Global"), base=T, weights=T, del
ta=T, logLik=T)

      logLik AICc  dLogLik dAICc df weight
Seven      2.2  2.5  59.3    0.0 3  0.69
Spatial*Seven 7.6  4.1  64.6    1.6 7  0.31
Global     10.7 55.9  67.7   53.4 16 <0.001
Spatial    -25.7 61.1  31.3   58.6 4  <0.001
Null      -57.0 118.5  0.0   116.0 2  <0.001
```

Regression Tree Object

```
node), split, n, deviance, yval
* denotes terminal node

1) root 29 86.7551000 1.531342e-17
2) env7< 0.4530161 14 9.0540510 -1.565784e+00
4) env7<= 0.2482159 6 0.8007901 -2.356327e+00 *
5) env7>= 0.2482159 8 1.6911960 -9.728763e-01 *
3) env7>= 0.4530161 15 11.3422800 1.461398e+00
6) env7< 0.6909247 7 1.3414610 6.710477e-01 *
7) env7>= 0.6909247 8 1.8022420 2.152955e+00 *

Vector assigning samples to groups
> regtree$where
[1] 3 3 3 3 3 3 4 4 4 4 4 4 4 4 6 6 6 6 6 6 7 7 7 7 7 7 7 7

> regtree$plits
      count  best  improve  index adj
env7    29  -1  0.7648976 0.4530161 0
spatial 29   3  0.6793786 1.0000000 0
env2    29  -1  0.1877116 0.2115776 0
env5    29  -1  0.1630396 0.4046268 0
env6    29  -1  0.1499714 0.6479728 0
env7    14  -1  0.7287656 0.2482159 0
spatial 14   3  0.6649992 2.0000000 0
env1    14  -1  0.2698882 0.6441778 0
env4    14  -1  0.2570974 0.8397628 0
env2    14  -1  0.2560837 0.9360351 0
env7    15  -1  0.7228333 0.6909247 0
spatial 15   3  0.6710280 0.0000000 0
env3    15  -1  0.2238585 0.7973027 0
env5    15  -1  0.2072037 0.3884382 0
env1    15  -1  0.1680820 0.7431955 0
```

List of the candidate splits.
 Improve = decrease in deviance
 Index = where the split is made

Fine Tuning

```
rpart(response~var1+var2+var3, data=envdata, control=rpart.control(minsplit=5, minbucket=10))
```

- minsplit – minimum number of samples in a split (5)
- minbucket – minimum samples in a terminal bucket (2)
- surrogatestyle – how to choose splits, penalties for missing data or missclassifications
- cp – complexity parameter. Degree or required improvement in group homogeneity (0.01)
 - With lower cp, you will grow a larger tree that can be pruned with cross validation

- Control of boosting
 - Learning rate (shrinkage) – contribution of each subsequent tree to overall model
 - Complexity – number of nodes and interactions
 - Number of trees

Journal of Animal Ecology

Journal of Animal Ecology, 2002, 71, 801–811

A working guide to boosted regression trees

J. Elmer*, J. R. Leathwick* and T. Hayes†

*School of Botany, The University of Melbourne, Parkville, Victoria, Australia 3101; †National Institute of Water and Atmospheric Research, PO Box 11710, Hamilton, New Zealand; and ‡Department of Statistics, Stanford University, CA, USA

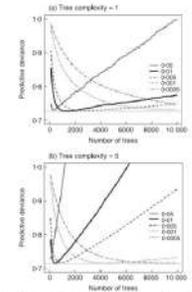


Fig. 2. The relationship between number of trees and predictive deviance for models fitted with the learning rates and two levels of tree complexity. Deviance was calculated from models fitted to a data set of 1000 trees, and predicted to a data set of 12 500 sites. The lowest predictive deviance achieved for each model is indicated by a dotted horizontal line; the line for learning rate achieving that minimum is shown in bold.

- Bag fraction – proportion of samples taken to do cross validation
- Published procedures to do stepwise model selection for determining the optimal parameters

- (1) Randomly divide available data into n subsets (from here on for this example, $n = 10$; this is a setting we commonly use).
- (2) Make 10 different training sets each comprising a unique combination of 9 subsets. Therefore for each training set there is a unique omitted subset that is used for testing.
- (3) Starting with a selected number of trees (m), say 50, develop 10 BRT models simultaneously on each training set, and test predictive performance on their respective omitted data. Both mean performance and standard errors are recorded.
- (4) Step forward and increase the m in each model by a selected and constant amount, and repeat step 3.
- (5) Repeat step 4 and after 10 steps start comparing the predictive performance of the 6th to 10th previous iterations against that of the current to 5th previous ones. Once the average of the more recent set is higher than the average of the previous set, the minimum has been passed.
- (6) Stop and record the minimum; this is the optimal m .

Fig. 4. Cross-validation method for identifying the optimal number of trees in a boosted regression tree model.

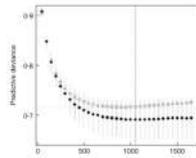


Fig. 4. Cross-validation (CV) model-fitting example. Data set of 1000 observations, with estimation of predictive accuracy from the CV mean on odd sites and SE on an independent estimate based on 13 500 excluded sites (replicate). Initial number of trees = 50, step size = 50. With learning rate of 0.01 and a complexity of 3, the step procedure identified the optimal number of trees as 1000, whereas with independent data the minimum was 950.

- Model to predict presence or absence of eel



Table 2. Summary of the relative contributions (%) of predictor variables for a boosted regression tree model developed with cross-validation on data from 1000 sites, using tree complexity of 5 and learning rate of 0.001.

Predictor	Relative contribution (%)
Segment I	24.7
USNative	11.3
Michael	11.1
DISEnt	9.7
LactCol	9.6
DIMbackslope	7.1
USRange	6.9
USRangeDeps	6.5
USRangeT	5.7
LogFlow	5.7
SegI and Flow	2.8
DISEdist	0.1

Table 1. Environmental variables used to model fish occurrence.

Variable	Description	Mean and range
Reach with presence	Weighted average of proportional cover of fish habitat: 1 = reach, 2 = reach, 3 = fish present, 4 = no reach, 5 = no reach	0.71, 1-7
Segment	Segment on temperature (°C)	16.1, 10.0-20.0
LogFlow	Water on temperature (°C), normalized with respect to logFlow	0.36, 0.1-1.1
LogFlowDeps	Segment on flow (m³ s⁻¹), rank was transformed	1.00, 1.0-4.0
DISEnt	Differences in water flow	15, 4.0-21.0
DISEnt	Volume of forest drainage (drainage, acres) dist	0.18, 0.1-1
DISEnt	Mean annual precipitation (mm)	11.1, 0.2-27
USRangeT	Average temperature in conditions (°C) computed with segment, normalized with respect to logFlow	0.36, -1.5-2.2
USRangeDeps	Dist on reach (meters) (0-100)	1.22, 0.1-3.84
USRange	Average slope to the system (calculated (°))	1.3, 0.0-5.0
USNative	Area with drainage from upstream	0.72, 0.1
Fishing method	Fishing method in the class (electric, net, spe, trap, minnow)	NA

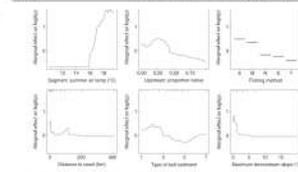


Fig. 5. Model variables that the best model selected for the eel. The graphs show the relationship of each variable with the presence/absence of eel. The x-axis is the variable and the y-axis is the response. The graphs show the relationship of each variable with the presence/absence of eel.

Habitat use of age 0 Alabama shad in the Pascagoula River drainage, USA

Similar question - predicting presence/absence based on environmental data. AIC approach.

Table 1. Candidate models used in AIC_c model selection for all four response variables (presence/absence, CPUE, mean length and mean condition). Third order and higher interaction terms were excluded from all models. # indicates the number of model parameters.

Model	Number	Variables	#	Hypothesis - presence/absence, abundance, size and condition of Alabama shad are best explained by
Null	1	None	2	None of the measured variables.
Temporal	2	Month	5	Differences at the fine temporal scale (monthly), large temporal scale (years), or a combination of both.
Spatial	4	Year	4	
Physicochemical	5	Habitat	4	Differences at the fine spatial scale (channel, bank and sandbar habitat).
Temporal - spatial	6	River	4	Large spatial scale (Dichotomous, Pascagoula and Leaf Rivers) or a combination of both.
Physicochemical	7	Habitat*River	8	
Temporal - physicochemical	8	PCI	3	Sample differences in physicochemical variables. PCI is primarily current velocity, Secchi depth and conductivity. PCI is primarily temperature, DO and pH.
Temporal - spatial	9	PCI	3	
Temporal - physicochemical	10	PCI*PCI	5	
Temporal - spatial	11	Month*Habitat	10	Fine scale temporal (monthly) and habitat variability.
Temporal - spatial	12	Year*River	9	Large scale temporal (yearly) and river variability.
Temporal - spatial	13	Year + Month*Habitat	12	Fine spatial scale and large temporal scale variability.
Temporal - spatial	14	Month + River*Year	12	Large spatial scale and fine temporal scale variability.
Temporal - physicochemical	15	Month*PCI	9	Fine scale temporal and one physicochemical axis variables.
Temporal - physicochemical	16	Month*PCI2	9	
Temporal - physicochemical	17	Year*PCI	7	Large scale temporal and one physicochemical axis variables.
Temporal - physicochemical	18	Year*PCI2	7	
Temporal - physicochemical	19	Month*PCI*PCI2	14	Large or fine scale temporal and both physicochemical axis variables.
Temporal - physicochemical	20	Year*PCI*PCI2	11	
Spatial + physicochemical	21	Habitat*PCI	7	Fine spatial scale and one physicochemical axis variables.
Spatial + physicochemical	22	Habitat*PCI2	7	
Spatial + physicochemical	23	River*PCI	7	Large spatial scale and one physicochemical axis variables.
Spatial + physicochemical	24	River*PCI2	7	
Spatial + physicochemical	25	Habitat*PCI*PCI2	11	Large or fine scale spatial and both physicochemical axis variables.
Spatial + physicochemical	26	River*PCI*PCI2	11	
Global	27	All	28	Combination of all temporal, spatial and physicochemical variables.

Habitat use of age 0 Alabama shad in the Pascagoula River drainage, USA

Table 2. Intergratable models, AIC_c statistics and weights (w_j) for all four response variables. Only models with a weight >10% of the best model were interpreted and listed. Model numbers match list in Table 1.

Response variable	#	Variables	AIC _c	ΔAIC _c	w _j
Presence/absence	7	Habitat*River	217.50	0.00	0.282
	26	Habitat*PCI*PCI2	218.50	1.00	0.173
	25	Habitat*PCI	218.80	1.40	0.143
Abundance (CPUE)	9	PCI	219.20	1.80	0.116
	14	River*PCI2	221.70	4.30	0.034
	22	Year*River	865.00	0.00	0.602
Mean length	7	Habitat*River	868.80	3.00	0.132
	13	Year + Month*Habitat	348.70	0.00	0.392
	14	Month + River*Year	348.20	0.50	0.310
Condition	27	All	350.40	1.70	0.169
	14	Month + River*Year	-23.30	0.00	0.934