广义线性模型
Generalized linear model

李欣海

中科院动物所

Generalized Linear Model

$$g(\mu) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdot\cdot\cdot + \beta_k x_k$$

GLM is an extension of general linear model that deals with ordinal and categorical response variables.

There are three components that are common to all GLMs (McCullagh & Nelder 1989) :

- Random component
- Systematic Component
- Link Function

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

Random Component:

The random component: refers to the probability distribution of the response Y.

Case 1. (Y_1, Y_2, . . ., Y_N) might be normal. In this case, we would say the random component is the normal distribution. This component leads to ordinary regression and analysis of variance models.

Case 2. If the observations are Bernoulli random variables (which have values 0 or 1), then we would say the link function is the binomial distribution. When the random component is the binomial distribution, we are commonly concerned with logistic regression models or probit models.

Case 3. Quite often the random variables Y_1, Y_2, . . ., Y_N have a Poisson distribution. Then we will be involved with Poisson regression models or loglinear models.

Systematic Component

The systematic component involves the explanatory variables x_1, x_2 , · · · , x_k.as linear predictors:

$$\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdot\cdot\cdot + \beta_k x_k$$

Link Function

The third component of a GLM is the link between the random and systematic components.

It says how the mean **μ** = E(Y) relates to the explanatory variables in the linear predictor through specifying a function g(μ):

$$g(\mu) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdot\cdot\cdot + \beta_k x_k$$

g(μ) is called the link function.

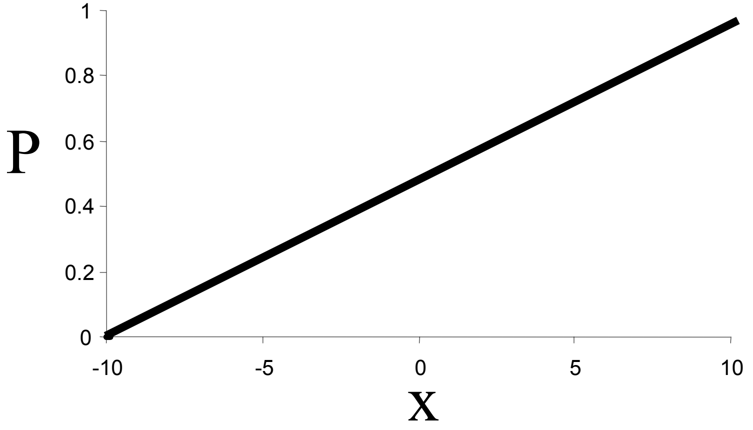
Generalized Linear Models

- The y_i's are allowed to have a distribution from the exponential family of distributions.
- The link function g(μ_i) is any monotonic function and defines the relationship between μ_i and $x_i\beta$.

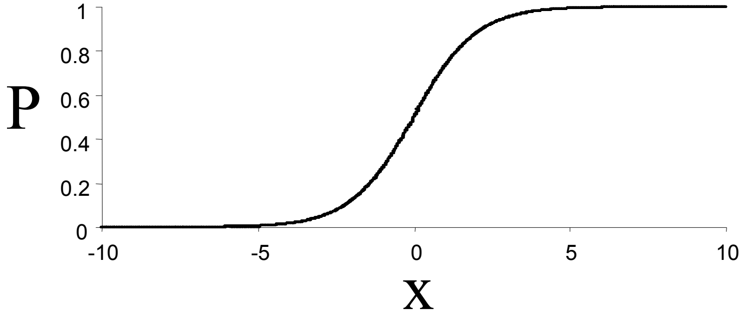
$$g(\mu_i) = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_k X_{ki}$$

Logistic regression

Dependent variable is binary

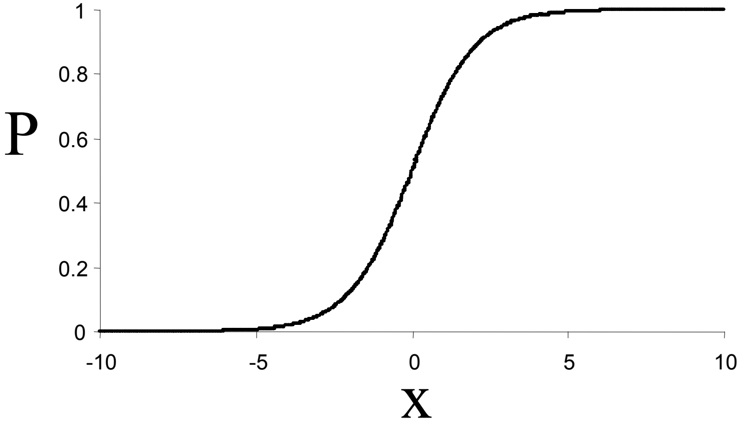


Linear function



Probit regression function

$$P(y_i = 1 | \ x_i) = p_i = \int_{-\infty}^{\alpha+\beta x_i} \frac{1}{\sqrt{2\pi}} \exp(-\frac{1}{2} t^2) dt$$



Logistic function

$$P(y_i = 1 | \ x_i) = p_i = \frac{1}{1 + e^{-(x_i)}}$$

$$P(y_i = 0 | \ x_i) = p_i = \frac{1}{1 + e^{(x_i)}}$$

Logit transformation

$$P(y_i = 1 \mid x_i) = p_i = \frac{1}{1+e^{-(x_i)}}$$

$$= \frac{e^{x_i}}{1+e^{x_i}}$$

$$1 - p_i = 1 - \frac{e^{x_i}}{1+e^{x_i}} \quad = \frac{1}{1+e^{x_i}}$$

$$Odds = \frac{p_i}{1-p_i} = e^{x_i}$$

$$\ln\left(\frac{p_i}{1-p_i}\right) = x_i$$

Model meanings – nest site use of birds

$$Odds = \frac{p_i}{1-p_i} = e^{x_i}$$

The response variable was the odds of a site having a nest, where odds are calculated as p/(1-p) and p is the proportion of sites have a nest. The statistical model was:

$Odds = exp(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots \beta_n X_n)$

where *n* is the number of explanatory variables. The log of the odds is known as the logit transform of *p*.

Advantages of Logit

- Properties of a linear regression model
- Logit between - ∞ and + ∞
- Probability (P) constrained between 0 and 1
- Directly related to odds of event

$$\ln\left(\frac{P}{1-P}\right)=\alpha+\beta x \qquad \frac{P}{1-P}=e^{\alpha+\beta x}$$

Assumptions

- Dependent variable is binary or dichotomous, vs. continuous dependent variables in linear regression.
- The cases are independent.
- The independent variables are not linear combinations of each other
- No linearity, the population means of the dependent variables at each level of the independent variable are not on a straight line.
- No homogeneity of variance, the variance of the errors are not constant.
- No normality, the errors are not normally distributed.

Example

- Risk of developing coronary heart disease (CD) by age (< 60 and > 60 years old)

CD	> 60 (1)	< 60 (0)
Present (1)	28	23
Absent (0)	11	72

Odds of disease among the old = 28/11
Odds of disease among the young = 23/72 **Odds ratio = 7.97**

R code

```
# Logistic regression
# Risk of developing coronary heart disease by age (<60 and >60 years old)
coronary1 <- data.frame(present = rep(1, 28), age = 'old')
coronary2 <- data.frame(present = rep(0, 11), age = 'old')
coronary3 <- data.frame(present = rep(1, 23), age = 'young')
coronary4 <- data.frame(present = rep(0, 72), age = 'young')
coronary <- rbind(coronary1, coronary2, coronary3, coronary4)
coronary <- rbind(coronary3, coronary4, coronary1, coronary2)
fit <- glm(present~age, data = coronary, family = binomial())
summary(fit)
```

```
Coefficients:
             Estimate  Std. Error  z value    Pr(>|z|)
(Intercept)   0.9343    0.3558     2.626     0.00865 **
ageyoung     -2.0755    0.4289    -4.839     1.31e-06 ***
```

```
Coefficients:
             Estimate  Std. Error  z value    Pr(>|z|)
(Intercept)  -1.1412    0.2395    -4.765     1.89e-06 ***
ageold        2.0755    0.4289     4.839     1.31e-06 ***
```

$$\ln\left(\frac{P}{1-P}\right)=\alpha+\beta_1\times Age=-1.1412+2.0755\times Age$$

Logistic Regression Model

	Coefficient	SE	Coeff/SE
Age	2.0755	0.4289	4.839
Constant	-1.1412	0.2395	-4.765

- ➢ β = increase in logarithm of odds ratio for a one unit increase in x
- Test of the hypothesis that β = 0 (Wald test)

$$\chi 2 = \frac{\beta^2}{\text{Variance } (\beta)} \qquad \text{(1 df)}$$

$$\text{Odds ratio} = e^{2.0755} = 7.97$$

$$\text{Wald Test} = 4.839^2 \text{ with 1df } (p < 0.05)$$

$$95\% \text{ CI} = e^{(2.0755 \pm 1.96 \text{ x } 0.4289)} = 3.4,\ 18.5$$

Interpretation of the coefficients in terms of the odds ratio – An Example

- Whether owning a car as a function of the income.
- 17 individuals, 14 own a car and 3 do not.

Income	Car owner
10	0
10	1
10	1
11	0
11	1
11	1
11	1
11	1
12	0
12	1
12	1
12	1
12	1
12	1
12	1
12	1
12	1

```
car1 <- data.frame(income = c(10:12), carowner = rep(0, 3))
car2 <- data.frame(income = rep(c(10:12), c(2, 4, 8)), carowner = rep(1, 14))
car <- rbind(car1, car2)
fit <- glm(carowner ~ income, data = car, family = binomial())
summary(fit)
```

Variables in the Equation					
	B	S.E.	Wald	df	Exp(B)
INCOME	**0.6931**	**0.8072**	**0.7372**	**1**	**2.0**
Constant	-6.2383	8.9794	0.4826	1	0.00195

Interpretation of the coefficients in terms of the odds ratio – An Example

$$\ln\left(\frac{P}{1-P}\right) = \alpha + \beta \times \text{income} = \alpha + 0.69 \times \text{income}$$

- $e^{\beta} = 2$

$$\frac{P}{1-P} = e^{\alpha} \times e^{0.69 \times income} = e^{\alpha} \times 2^{income}$$

- So: increasing the income by one unit increases the odds of owning a car by a factor of 2 (increase in 100%) so that:

(odds after increasing income)/ (odds before increasing income) = 2

- If we look at the data we can see that this model predicts perfectly:

	car owner				
income	**1**	**0**	**P(own)**	**P(not own)**	**Odds of Owning a car**
10	2	1	2/3=0.66	1/3=0.33	0.66/0.33=2
11	4	1	4/5=0.8	1/5=0.2	0.8/0.2=4
12	8	1	8/9=0.888	1/9=0.111	0.888/0.111=8

Marginal effect of a change in X

- ln[p/(1-p)] = α + βX + e

  The slope coefficient (β) is interpreted as the rate of change in the "log odds" as X changes … not very useful.

- We are also interested in seeing the effect of an explanatory variable on the probability of the event occurring

- p = 1/[1 + exp(-α - βX)]

  The marginal effect of a change in X on the probability is:

  əp/əX = β*p(1-p)*

$$= \beta \times \frac{1}{1+e^{-(\alpha+\beta X)}} \times \frac{1}{1+e^{(\alpha+\beta X)}}$$

Basically, the size of the ‘marginal effect’ will depend on two things:

- – β coefficient
- – The initial value of X

Marginal Effects: β xP(1-P)

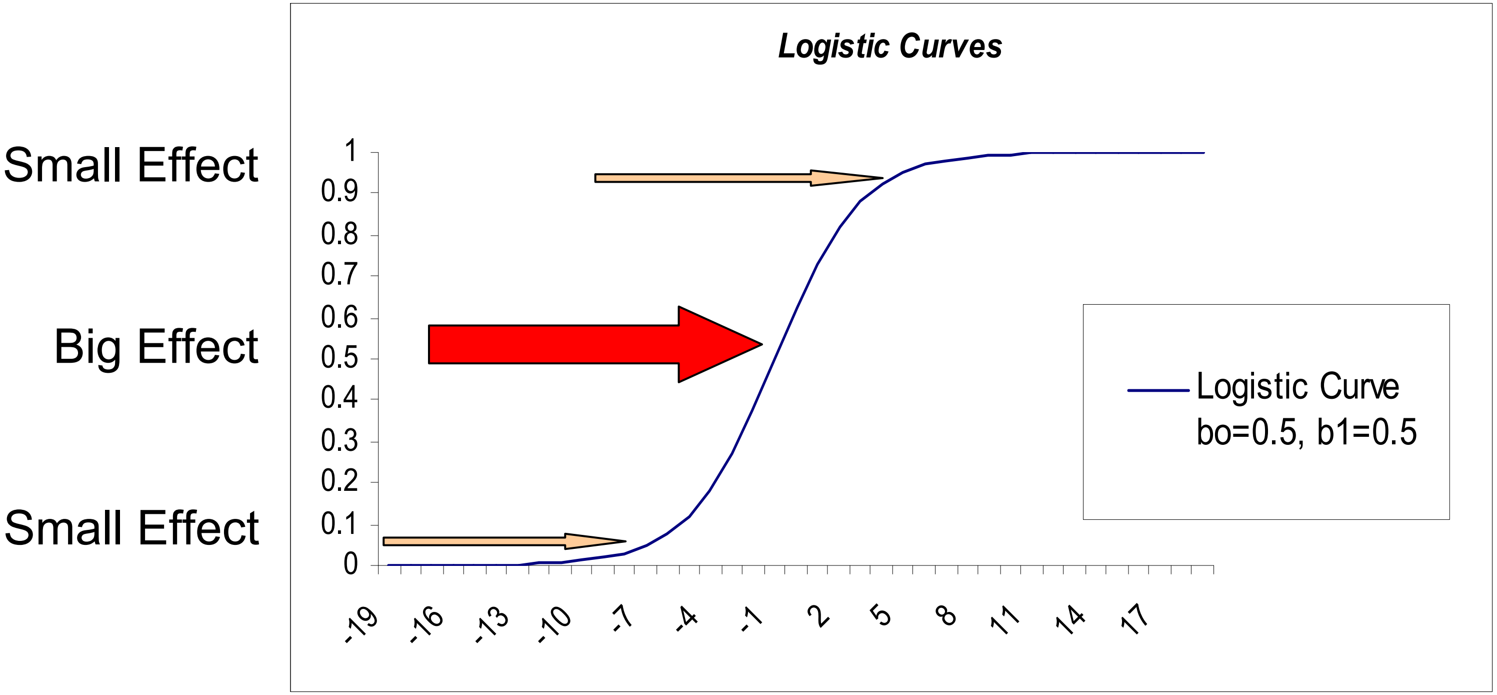
- Passing or failing an exam as a function of the number of hours of study
- Previous study indicated the estimates of α and β were:

$$\alpha = -5 , \quad \beta = 0.3$$

- So what's the effect of studying one more hour in the probability of the event occurring:

Initial hours of study	P	1-P	P(1-P)	Marginal effect
5	0.029	0.971	0.028	0.009
10	0.119	0.881	0.105	0.031
15	0.378	0.622	0.235	0.071
20	0.731	0.269	0.197	0.059
25	0.924	0.076	0.070	0.021
30	0.982	0.018	0.018	0.005

The importance of the initial value of X in the marginal effect



Starting the change from the central values of X will have a higher impact on the probability of the event occurring than starting from very low or very high values of X.

Some useful R codes

```
# Logistic regression
fit <- glm(carowner ~ income,  data = car,  family = binomial())

summary (fit) # display results
confint (fit) # 95% CI for the coefficients
exp(coef (fit)) # exponentiated coefficients
exp(confint (fit)) # 95% CI for exponentiated coefficients
pred = predict (fit,  type = "response") # predicted values (logit)
res = residuals (fit,  type = "deviance") # residuals
```

How to estimate model coefficients
Maximum likelihood estimation (MLE)

For one observation

$$P(y_i) = p_i^{y_i}(1-p_i)^{1-y_i}$$

Likelihood function

$$L(\theta) = \prod_{i=1}^{n} p_i^{y_i}(1-p_i)^{1-y_i}$$

Goodness of fit for the full model - likelihood ratio test (LR)

- We compare the value of the likelihood function in a model with the variables with the value of the likelihood function in a model without the variables. The test:

$$LR = -2L\hat{L}_0 - (-2L\hat{L}_S) \Rightarrow \chi^2{}_k$$

where $L\hat{L}_0$ is the log likelihood value of the null model (only intercept included); $L\hat{L}_S$ is the log likelihood value of the full model (taking into account of all variable parameters).

  - The statistic is distributed as χ^2 with as many degrees of freedom as coefficients we are restricting

```
# likelihood ratio test
fit.full  <- glm(present ~ .,          data = coronary,  family = binomial())
fit.null <- glm(present ~ NULL,  data = coronary,  family = binomial())
lrtest(fit.full,  fit.null)
```

Goodness of fit - Analogous R^2

$-2LL\hat{L}_0$ Refer to total sum of square

$-2LL\hat{L}_0 - (-2LL\hat{L}_S)$ Refer to regression sum of square

Likelihood ratio index (LRI):

$$\mathrm{LRI} = \left(\frac{-2LL\hat{L}_0 - (-2LL\hat{L}_S)}{-2LL\hat{L}_0}\right) = R^2 \qquad R^2{}_{adj} = \frac{R^2}{R^2{}_{\max}} = \frac{R^2}{1-(\hat{L}_0)^{2/n}}$$

```
# R code
library(Design) # required for lrm()
fit2 <- lrm(y ~ x1  +  x2,  data = data1)
fit2[[3]][10] # R square
```

Stepwise Regression base on Akaike's Information Criterion (AIC)

AIC = -2 ln (likelihood) + 2K

K = number of parameters in the model, including 1 for the constant and 1 for the error term

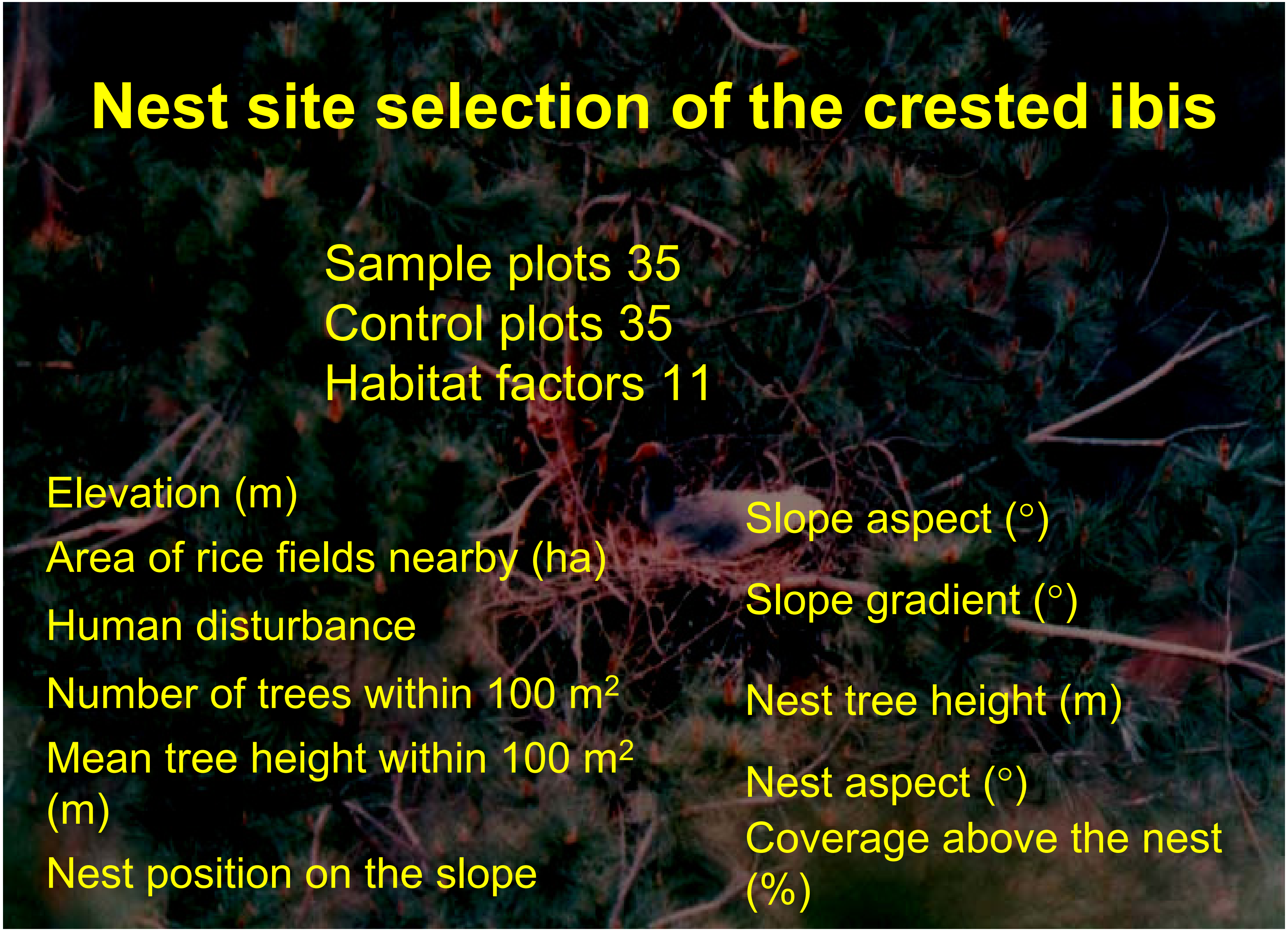
$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 \qquad \mathrm{K} = 6$$

For small samples (*n*/*K* < 40), use $\mathrm{AIC_c}$ for small sample size

$$\mathrm{AIC_c} = \mathrm{AIC} + \frac{2K(K+1)}{n-K-1}$$

# R code

```
step(fit) # Stepwise Regression
```



Nest site selection of the crested ibis

Sample plots 35
Control plots 35
Habitat factors 11

Elevation (m)

Area of rice fields nearby (ha)

Human disturbance

Number of trees within 100 m^2

Mean tree height within 100 m^2 (m)

Nest position on the slope

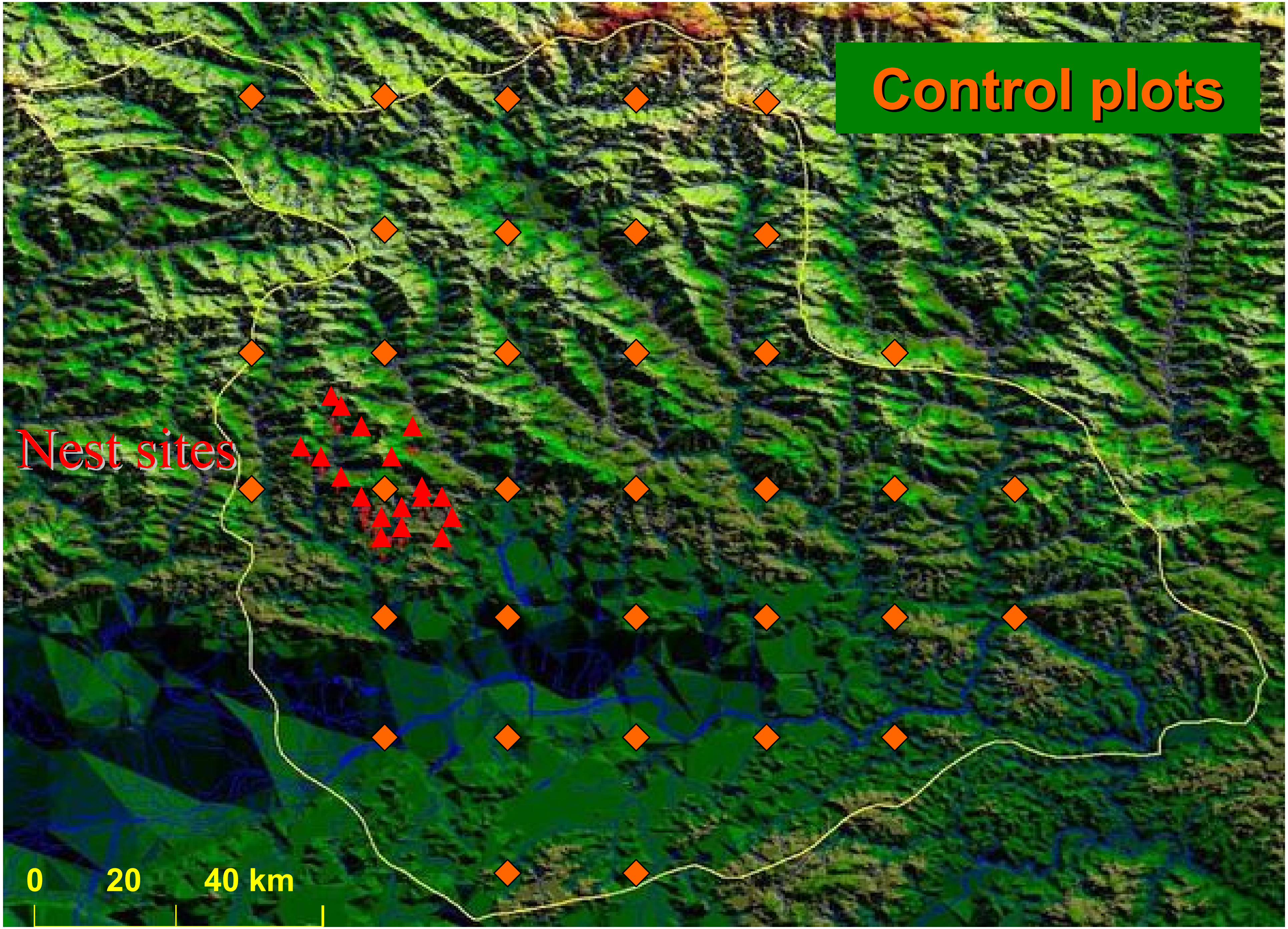
Slope aspect (°)

Slope gradient (°)

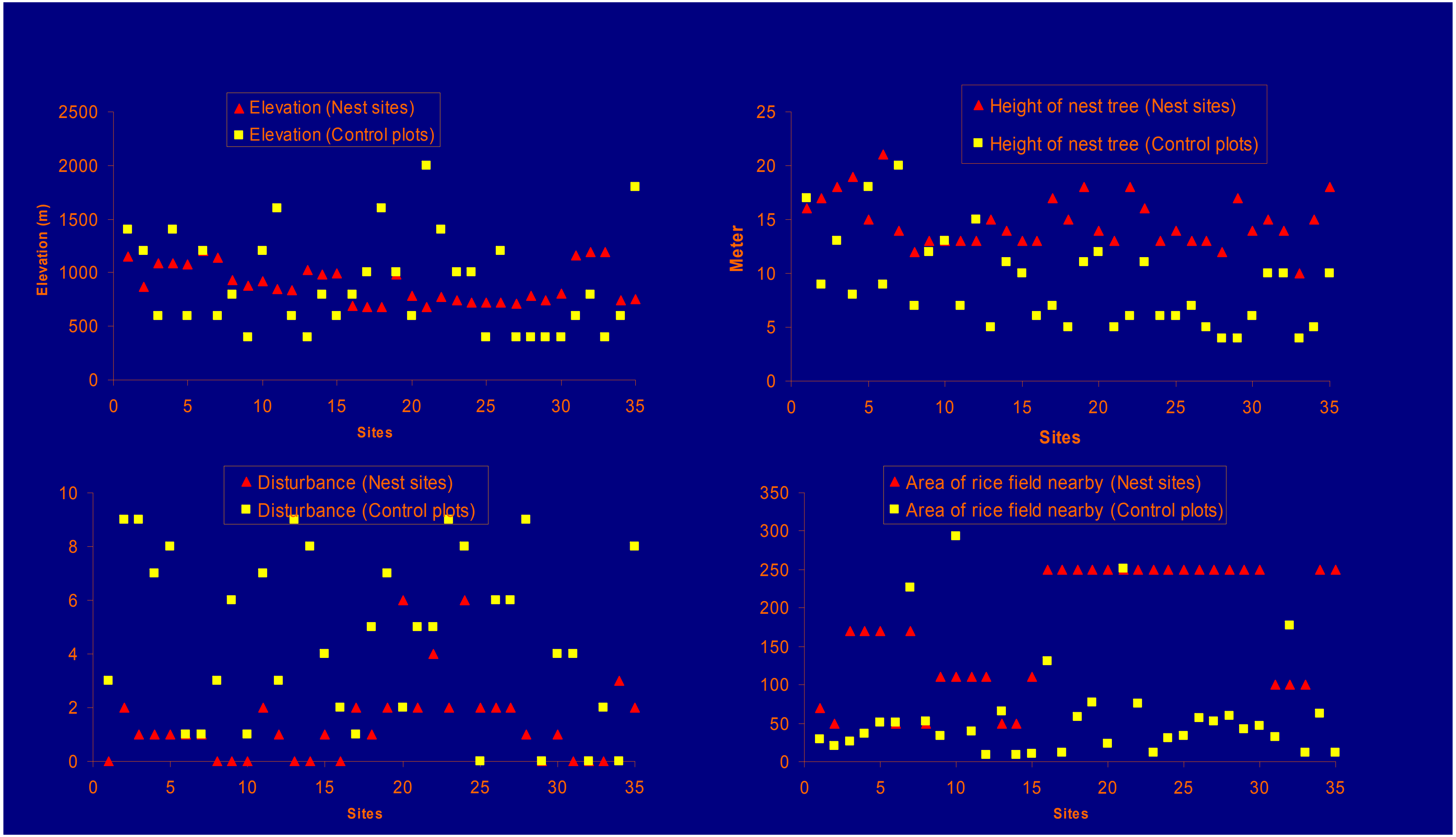
Nest tree height (m)

Nest aspect (°)

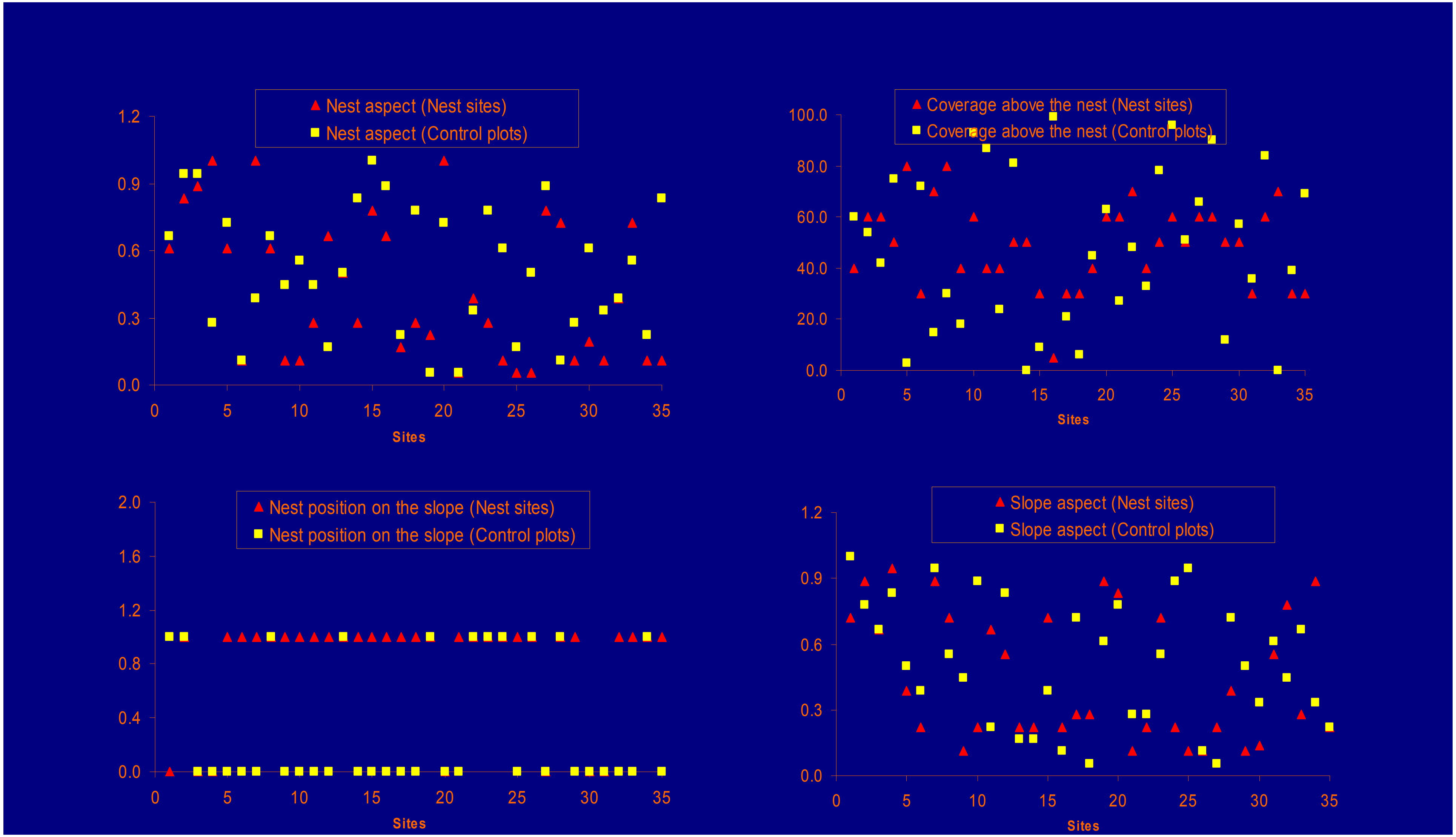
Coverage above the nest (%)


Control plots
Nest sites
0
20
40 km

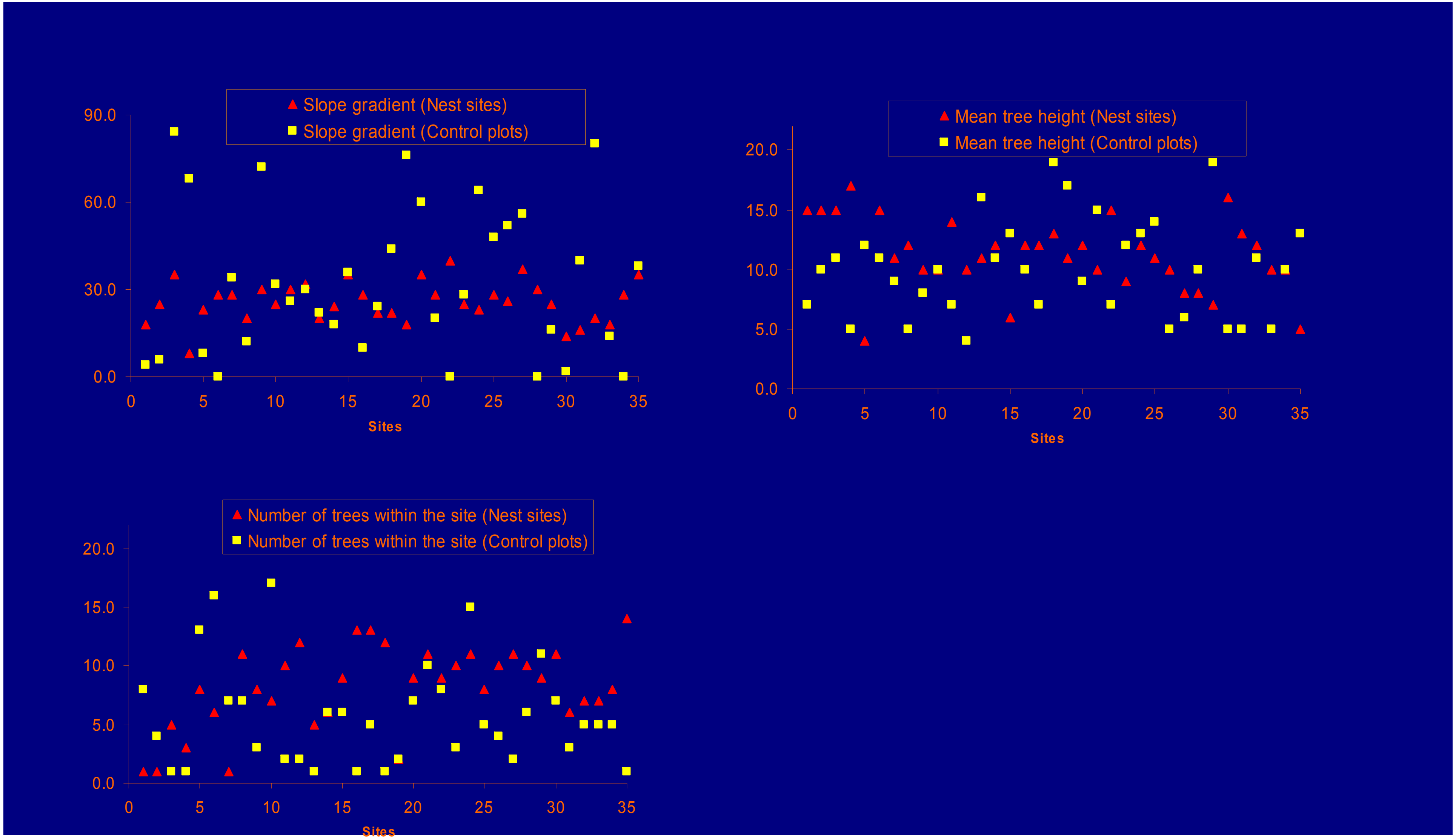
Source data 1



Source data 2



Source data 3



Correlation

The Pearson correlations between the 11 habitat variables measured at 35 nest sites of crested ibis in Yang county, Shaanxi province, China. Mean values and standard deviations (S.D.) are also shown.

Habitat variables	Correlation coefficients											Mean	S.D.
	1	2	3	4	5	6	7	8	9	10	11		
1. Elevation (m)	1	-0.72*	-0.48*	-0.70*	0.21	-0.02	0.39*	-0.38*	0.162	0.34*	0.21	894.00	176.53
2. Area (ha) of rice fields within 1 km^2		1	0.53*	0.49*	-0.23	-0.08	-0.23	0.23	0.05	-0.21	-0.12	11.62	5.40
3. Human disturbance			1	0.22	0.06	-0.154	0.15	0.38*	0.10	-0.02	0.08	1.40	1.52
4. Number of trees within 100 m^2				1	-0.37*	-0.00	-0.52*	0.34*	-0.33	0.012	-0.25	8.11	3.53
5. Mean tree height within 100 m^2 (m)					1	-0.24	0.23	-0.34*	0.32*	0.11	-0.06	11.23	3.06
6. Nest position on the slope						1	0.03	0.22	-0.21	-0.00	-0.07	2.03	0.45
7. Slope aspect (South = 1, North = 0)							1	-0.15	0.18	0.55*	0.06	0.45	0.29
8. Slope gradient (°)								1	-0.05	0.10	0.01	25.69	7.01
9. Nest tree height (m)									1	-0.08	-0.23	14.80	2.36
10. Nest aspect (South = 1, North = 0)										1	0.32	0.43	0.32
11. Coverage above the nest (%)											1	49.00%	16.53%

Stepwise logistic regression for modeling nest site selection of crested ibis in Yang County, Shaanxi Province, China.

Step	Habitat features	Selection coefficients	Standard Error	P value for model selection	AIC
1	Nest tree height (m)	0.94	0.38	<.0001	63.356
2	Human disturbance	-0.99	0.40	0.0001	50.475
3	Slope aspect	-5.82	3.25	0.0013	41.727
4	Area of rice fields nearby (ha)	0.35	0.19	0.0109	36.252
5	Nest position on the slope	3.73	2.30	0.0478	34.336
6	Mean tree height within 100 m^2 (m)	0.28	0.27	0.0320	31.924
7	Nest aspect	54.9285	31.5378	0.0112	26.048
8	Slope gradient (°)	-0.4080	0.3602	0.2866	23.226
9	Coverage above the nest	0.5201	0.5586	0.0841	24.322
10	Number of trees within 100 m^2	-0.006830	0.00616	0.1160	25.764
11	Elevation (m)	0.07670	0.1328	0.1450	27.275

Model equation

$$\begin{aligned}\text{logit (p)} = & -20.99 + 0.94 \times \text{nest tree height} \\ & - 0.99 \times \text{human disturbance} \\ & + 3.63 \times \text{nest position} \\ & + 0.35 \times \text{rice paddy area} \;+\; \ldots\end{aligned}$$

Probability of nest selection：

$P = e^{\text{logit (p)}} /(1 + e^{\text{logit (p)}})$

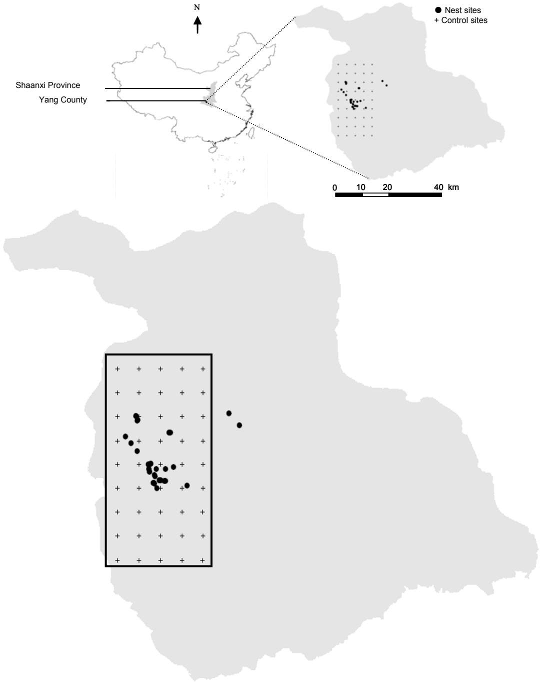
- R-Square 0.7380
- Max-rescaled R-Square 0.9840

李欣海, 马志军, 李典谟, 丁长青, 翟天庆, 路宝忠。2001。应用资源选择函数研究朱鹮的巢址选择。生物多样性, 9：352-358。

Scale dependant nest site selection of crested ibis

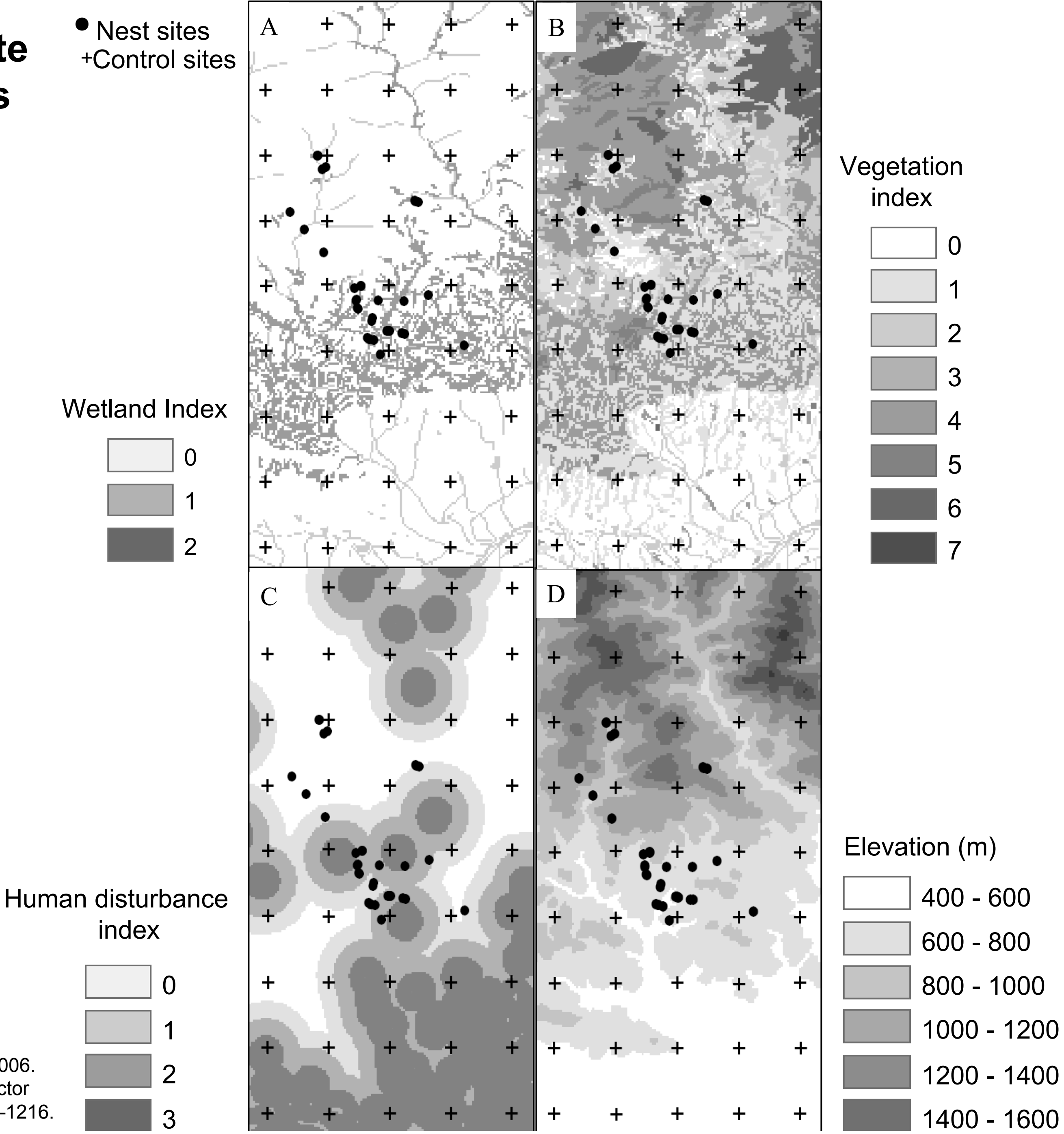
The maps of nest sites and control sites, and four layers in the GIS database:

A. Wetland index;
B. Vegetation index;
C. Human impact index;
D. Elevation.



Study area

Li, X. H., D. M. Li, Z. J. Ma, and D. C. Schneider. 2006. Nest site use of crested ibis: dependence of a multifactor model on spatial scale. Landscape Ecology 21: 1207–1216.

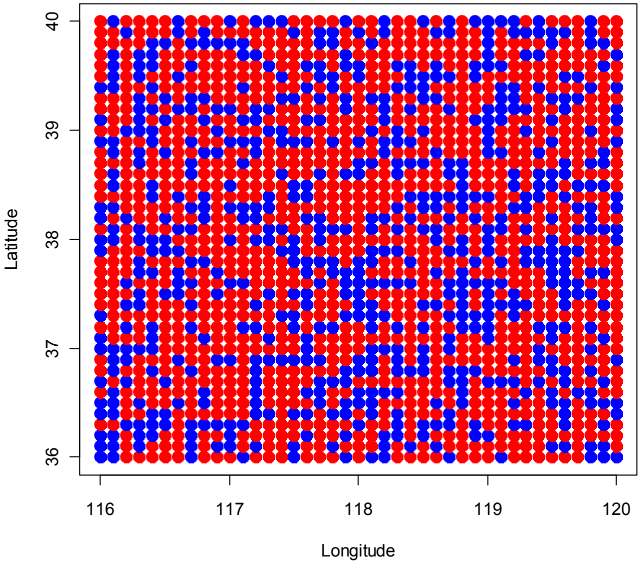


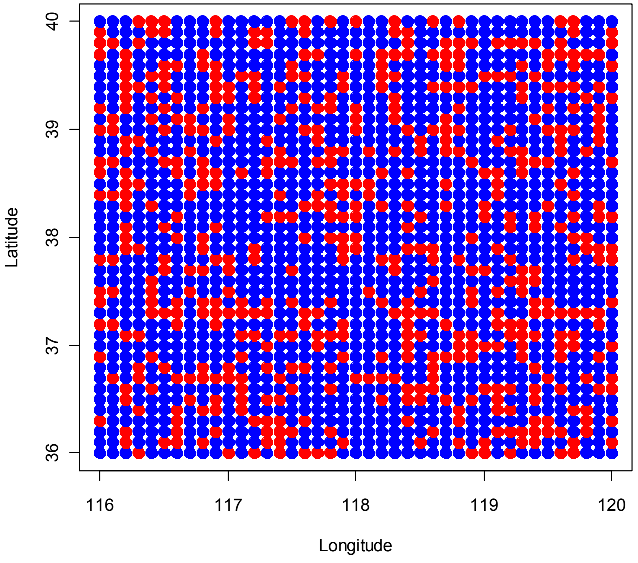
R code for logistic regression
– an example for species habitat prediction in future climate conditions

```
#Generate a dataset of species occurrences and control sites
x <- seq(116, 120,  by = 0.1) #longitude
y <- seq(36,     40,  by = 0.1) #latitude
Geo <- expand.grid(x, y) #switch a grid to a two-column table
names(Geo) = c('Lon', 'Lat')
use <- sample(0:1, length(Geo$Lat), rep = TRUE) #present/absent
elev <- rnorm(length(Geo$Lat), 1000, 200) #elevation
aspect <- sample(1:100, length(Geo$Lat), rep = TRUE)/100 #slope aspect of nest tree
temperature <- rnorm(length(Geo$Lat), 20, 5)*1000/elev #temperature negatively associated with elevation
distr <- cbind(Geo, use, elev, aspect, temperature) #the database
head(distr) #show the structure of database
fit <- step(glm(use ~ elev + aspect + temperature, data = distr, family = binomial())); summary(fit)
logit.current <- predict(fit,  type = "response") # predicted values
hist(logit.current)
p.current <- exp(logit.current)/(1 + exp(logit.current)); hist(p.current) # p values of species occurrences
coplot(p.current ~ elev | Lon*Lat,  data = distr,  overlap = 0,  number = c(8, 8))
result <- cbind(distr, p.current); head(result)

#plot current distribution
attach(distr); x11(); plot(116:120,  36:40,  type = "n",  xlab = 'Longitude',  ylab = 'Latitude')
for (i in 1:length(distr$Lat)){
if(p.current[i]>0.62) points(Lon[i],  Lat[i],  col = "red",  cex = 1.5,  pch = 19)
if(p.current[i]<0.62) points(Lon[i],  Lat[i],  col = "blue",  cex = 1.5,  pch = 19)}
```

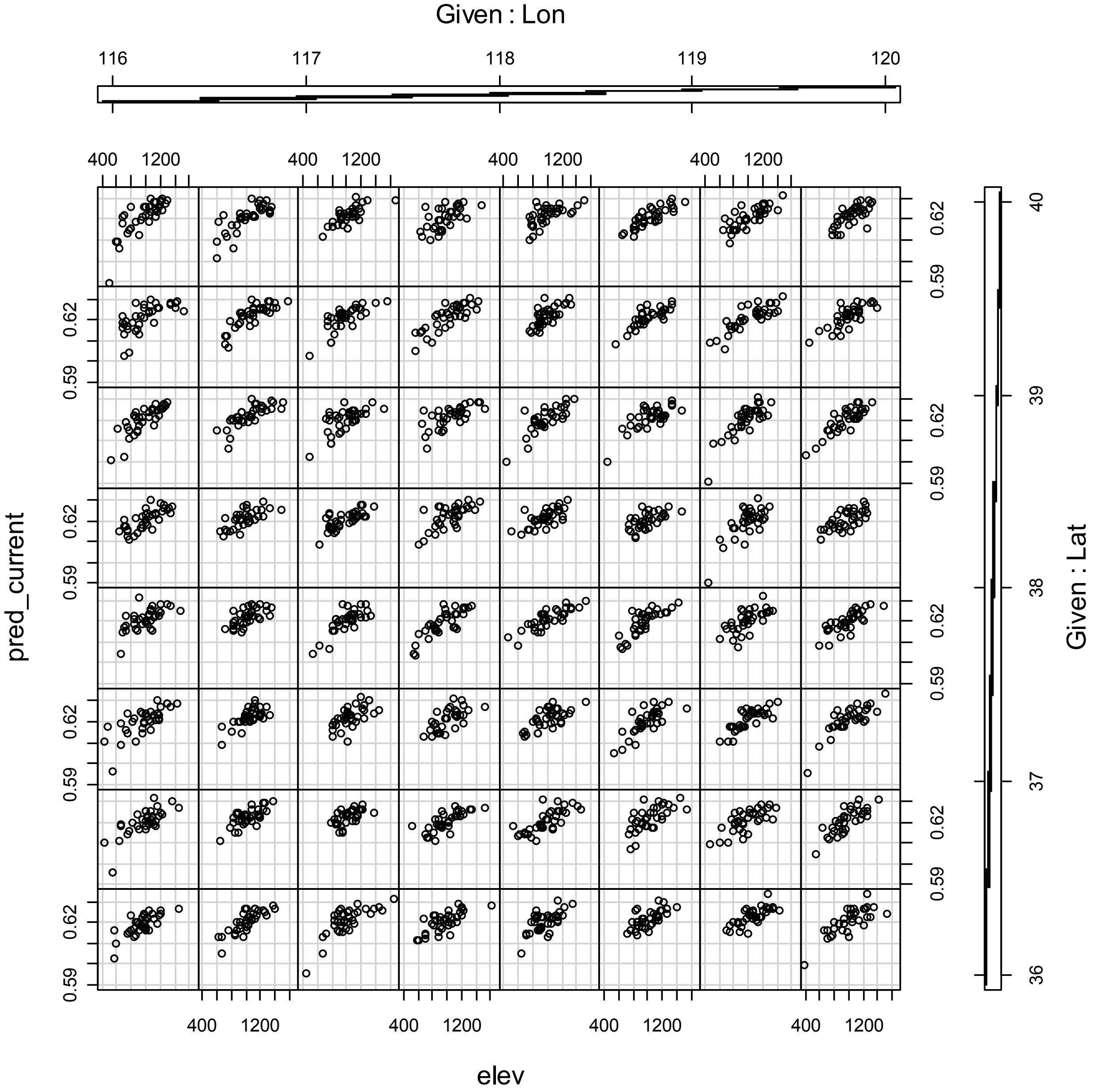
R code for logistic regression – an example for habitat prediction in future

```
#predict future distribution
distr.future <- distr
distr.future$temperature <- distr.future$temperature + 2 # a global warming scenario
logit.future <- predict(fit , newdata = distr.future); hist(logit.future)
p.future <- exp(logit.future)/(1 + exp(logit.future)); hist(p.future) # p values of species occurrences

#plot future distribution
x11(); plot(116:120,  36:40,  type = "n",  xlab = 'Longitude',  ylab = 'Latitude')
for (i in 1:length(distr.future$Lat)){
if(p.future[i]>0.5) points(Lon[i],  Lat[i],  col = "red",  cex = 1.5,  pch = 19)
if(p.future[i]<0.5) points(Lon[i],  Lat[i],  col = "blue",  cex = 1.5,  pch = 19)}
```





coplot(p.current ~ elev | Lon*Lat, data = distr, overlap = 0, number = c(8, 8))



Example of a Generalized Linear Model - General Linear Model

- The response variable is continuous.
- The distribution is normal.
- The link function is the identity function.

$$g(\mu) = \mu$$

```
fit <- glm(y ~ x1  +  x2,  data = data2,  family = gaussian)
```

Example of a Generalized Linear Model - Logistic Regression

- The response variable is discrete.
- The distribution is binomial.
- The link function is the logit.

$$g(\mu) = \ln[\mu/(1-\mu)]$$

```
fit <- glm(y ~ x1  +  x2,  data = data2,  family = binomial())
```

Example of a Generalized Linear Model – Negative Binomial Distribution

- The response variable is a count.
- The distribution is a negative binomial distribution.
- The link function is the natural logarithm.

$$g(\mu) = \ln(\mu)$$

```
library(MASS)
fit <- glm.nb(y ~ x1  +  x2,  data = data2)
```

Example of a Generalized Linear Model - Poisson Regression

- The response variable is a count.
- The distribution is a Poisson distribution.
- The link function is the natural logarithm.

$$g(\mu) = \ln(\mu)$$

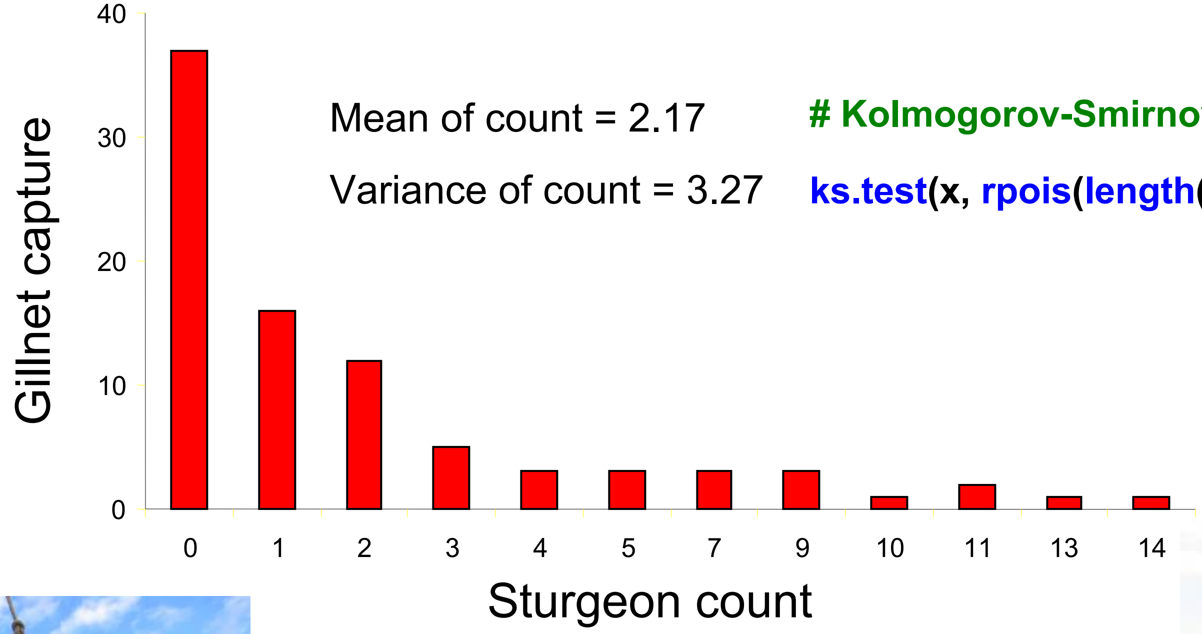
```
fit <- glm(y ~ x1  +  x2,  data = data1,  family = poisson())
```

Over dispersion

Generalized linear models (GLMs) are simple, convenient models for count data, but they assume that the variance is a specified function of the mean.

Over dispersion is a phenomenon that occurs occasionally with binomial and Poisson data. For Poisson data, it occurs when the variance of the response *Y* exceeds the Poisson variance. .

Habitat use of shortnose sturgeon - gill net capture

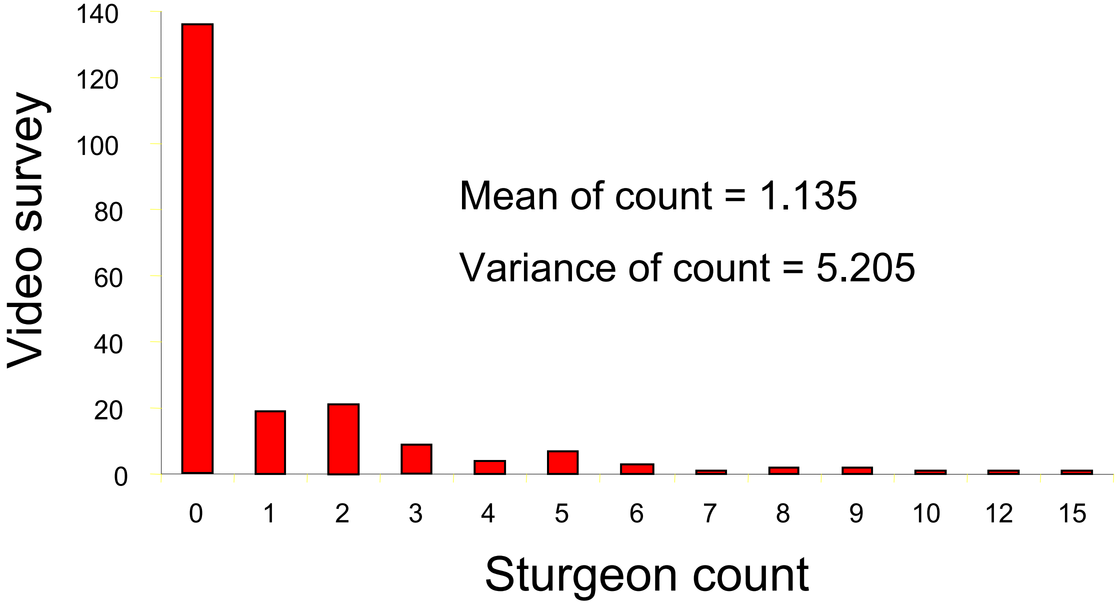


Mean of count = 2.17

Variance of count = 3.27

```
# Kolmogorov-Smirnov test for Poisson distribution
ks.test(x, rpois(length(x), mean(x)))
```

Habitat use of shortnose sturgeon - overwintering)



Mean of count = 1.135

Variance of count = 5.205

GLM for habitat use of shortnose sturgeon

Fish count data (gill net capture)

```
glm (count.gillnet ~ depth  +  temperature + salinity + substrate +
      velocity,  data = data.count.gillnet , family = poisson())
```

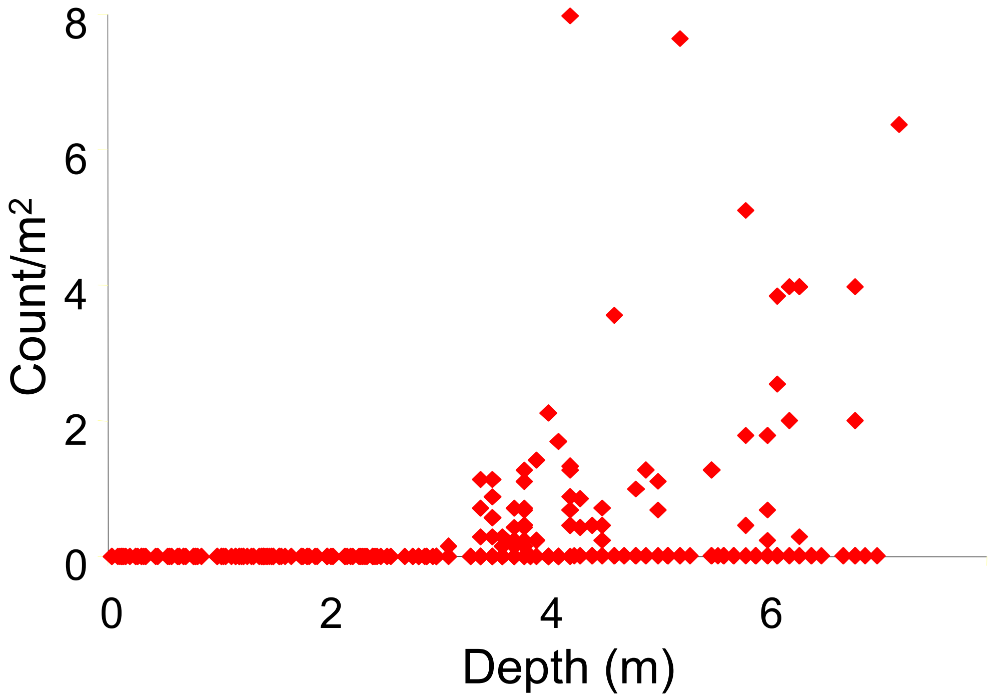
Fish count data (underwater video survey)

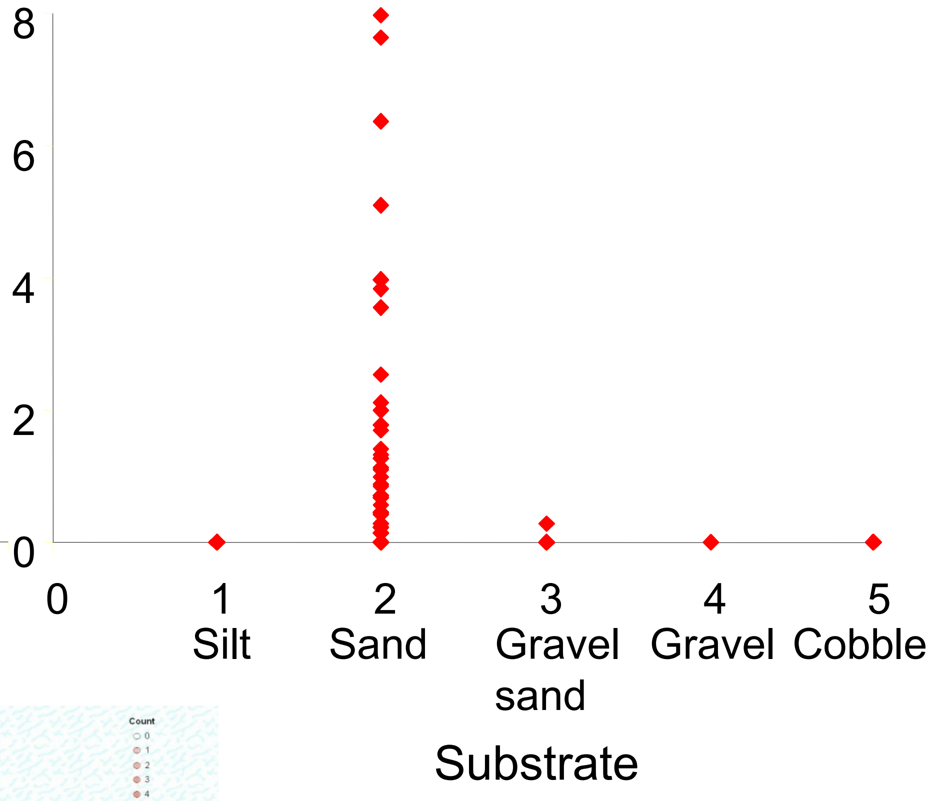
```
glm.nb (count.video ~ depth + substrate,  data = data.count.video)
```

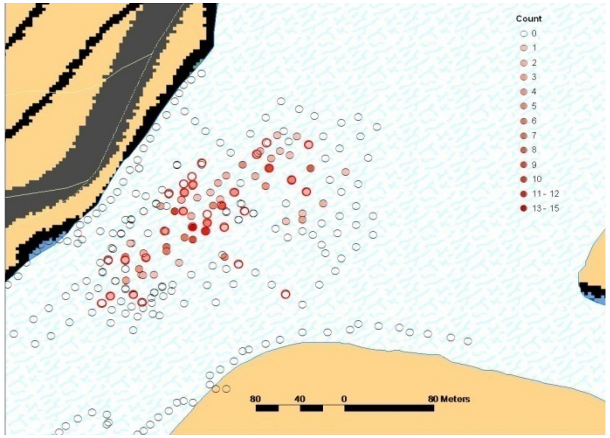
Fish tracking data (sonar telemetry)

```
glm (present.tracking ~ depth  +  temperature + salinity + substrate
     + velocity,  data = data.tracking, family = binomial())
```

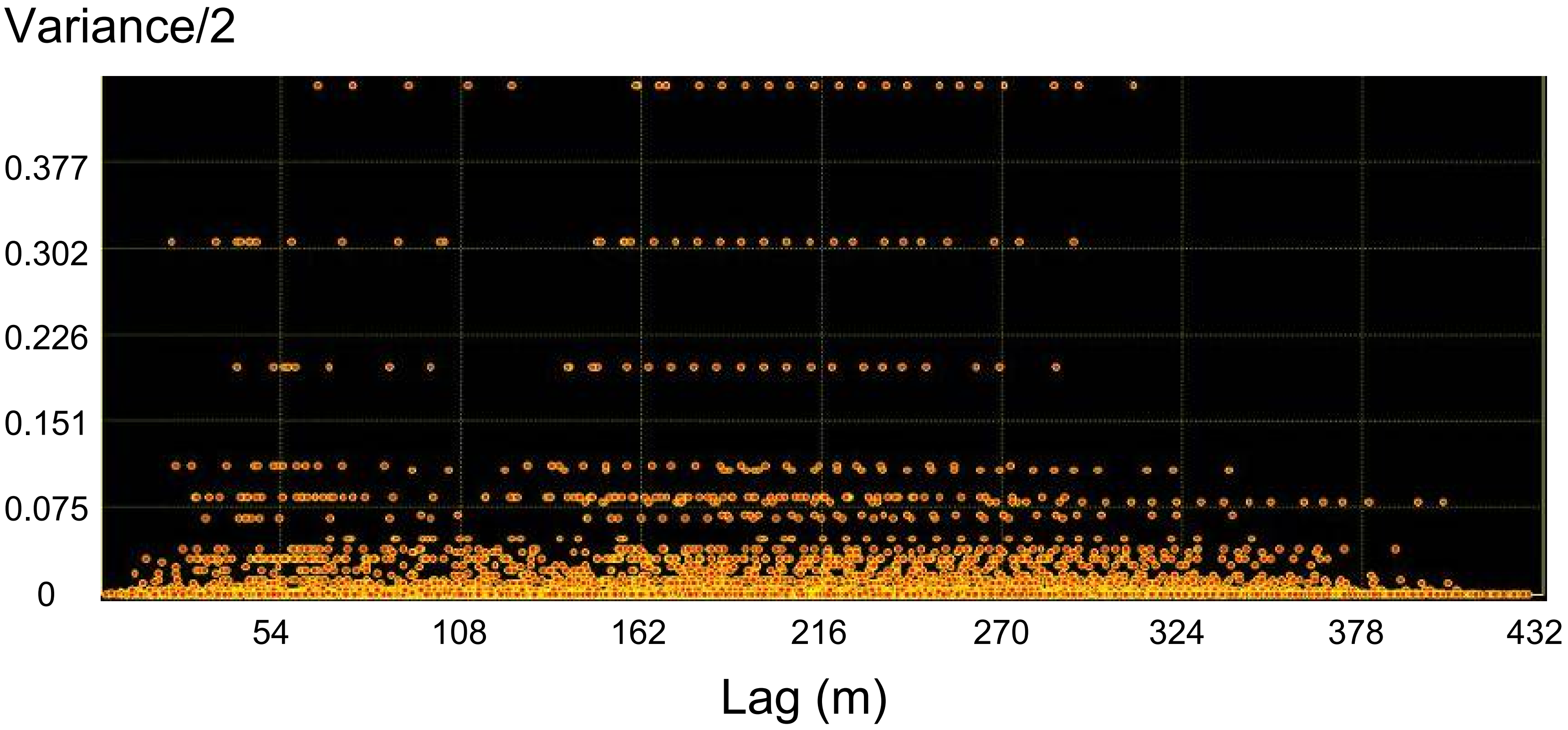
Association with habitat variables in winter







Check spatial autocorrelation using ArcGIS: Semivariogram cloud (density)



Results

$$\log(\mu) = -2.91 + 0.75Depth - 0.8Substrate$$

where $\mu = E(density)$

Parameter	DF	Estimate	Standard Error	Wald 95% Confidence Limits		Chi-Square	Pr > ChiSq
Intercept	1	-2.9109	0.6391	-4.1634	-1.6584	20.75	<.0001
depth	1	0.7547	0.1128	0.5336	0.9758	44.77	<.0001
Substrate	1	-0.8001	0.3107	-1.4090	-0.1911	6.63	0.0100

Conclusion

- Shortnose sturgeon concentrated within two ha
- On the flat sandy substrate
- At the depth of 3.1-6.9 m
- Population abundance is about 4836±140

Li, X. H., Matthew K. Litvak, and John E. Hughes Clarke. 2007. Overwintering habitat use of shortnose sturgeon: defining critical habitat using a novel underwater video survey and modeling approach. Canadian Journal of Fisheries and Aquatic Sciences 64: 1248-1257

谢谢!