Generalized linear model

- Popular generalized linear models
- Generalized linear mixed models
- Zero truncated generalized linear models
- Zero inflated generalized linear models
- Conditional logistic regression
- Multinomial logistic regression
- Model evaluation (Kappa, ROC, etc.)

Recall what is generalized linear models

- The y_i's are allowed to have a distribution from the exponential family of distributions.

- The link function $g(\mu_i)$ is any monotonic function and defines the relationship between μ_i and $x_i\beta$.

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

The exponential family (McCullagh & Nelder 1989)

We will assume that the observations come from a distribution in the exponential family with probability density function

$$f(y_i) = \exp\{\frac{y_i\theta_i - b(\theta_i)}{a_i(\phi)} + c(y_i, \phi)\}. \quad \text{(B.1)}$$

Here θ_i and ϕ are parameters and $a_i(\phi)$, $b(\theta_i)$ and $c(y_i, \phi)$ are known functions. In all models considered in these notes the function $a_i(\phi)$ has the form

$$a_i(\phi) = \phi/p_i,$$

where p_i is a known *prior weight*, usually 1.

The parameters θ_i and ϕ are essentially location and scale parameters. It can be shown that if Y_i has a distribution in the exponential family then it has mean and variance

$$\mathrm{E}(Y_i) = \mu_i = b'(\theta_i) \quad \text{(B.2)}$$

$$\mathrm{var}(Y_i) = \sigma_i^2 = b''(\theta_i)a_i(\phi), \quad \text{(B.3)}$$

where $b'(\theta_i)$ and $b''(\theta_i)$ are the first and second derivatives of $b(\theta_i)$. When $a_i(\phi) = \phi/p_i$ the variance has the simpler form

$$\mathrm{var}(Y_i) = \sigma_i^2 = \phi b''(\theta_i)/p_i.$$

The exponential family just defined includes as special cases the normal, binomial, Poisson, exponential, gamma and inverse Gaussian distributions.

The exponential family (McCullagh & Nelder 1989)

Example: The normal distribution has density

$$f(y_i) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\{-\frac{1}{2}\frac{(y_i-\mu_i)^2}{\sigma^2}\}.$$

Expanding the square in the exponent we get $(y_i-\mu_i)^2 = y_i^2 + \mu_i^2 - 2y_i\mu_i$, so the coefficient of y_i is μ_i/σ^2. This result identifies θ_i as μ_i and ϕ as σ^2, with $a_i(\phi) = \phi$. Now write

$$f(y_i) = \exp\{\frac{y_i\mu_i - \frac{1}{2}\mu_i^2}{\sigma^2} - \frac{y_i^2}{2\sigma^2} - \frac{1}{2}\log(2\pi\sigma^2)\}.$$

This shows that $b(\theta_i) = \frac{1}{2}\theta_i^2$ (recall that $\theta_i = \mu_i$). Let us check the mean and variance:

$$E(Y_i) = b'(\theta_i) = \theta_i = \mu_i,$$
$$\text{var}(Y_i) = b''(\theta_i)a_i(\phi) = \sigma^2.$$

Normal

$$f(y) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2}\left(\frac{y-\mu}{\sigma}\right)^2\right) \qquad -\infty < y < \infty$$

$$V(\mu) = 1$$

$$a(\phi) = \phi = \sigma^2$$

Inverse Gaussian

$$f(y) = \frac{1}{\sqrt{2\pi y^3}\sigma} \exp\left(-\frac{1}{2\mu^2 y}\left(\frac{y-\mu}{\sigma}\right)^2\right) \qquad y>0$$

$$V(\mu) = \mu^3$$

$$a(\phi) = \phi = \sigma^2$$

Gamma

$$f(y) = \frac{1}{y\Gamma(\nu)}\left(\frac{\nu y}{\mu}\right)^{\nu} \exp\left(-\frac{\nu y}{\mu}\right) \quad \text{for } y > 0$$

$$V(\mu) = \mu^2$$

$$a(\phi) = \phi = \nu^{-1}$$

Poisson

$$f(y) = \frac{\mu^y e^{-\mu}}{y!} \quad \text{for } y = 0, 1, 2, \ldots$$

$$V(\mu) = \mu$$

$$a(\phi) = \phi = 1$$

Binomial

$$f(y) = \binom{m}{r} \mu^r (1-\mu)^{m-r} \quad \text{for } y = r/m, r = 0, 1, 2, ..., m$$

$$V(\mu) = \mu(1-\mu)$$

$$a(\phi) = \phi/m = 1/m$$

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())
```

Poisson Regression Model

$$\Pr(y) = \frac{\lambda^y e^{-\lambda}}{y!}$$

- The response variable, *y*, has a Poisson distribution conditional on the predictor variables.
- The default link function is the log.
- Therefore,

$$\log(E[y_i \mid x_i]) = \log(\mu_i) = x_i\beta$$

or

$$\mu_i = e^{x_i\beta}$$

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.

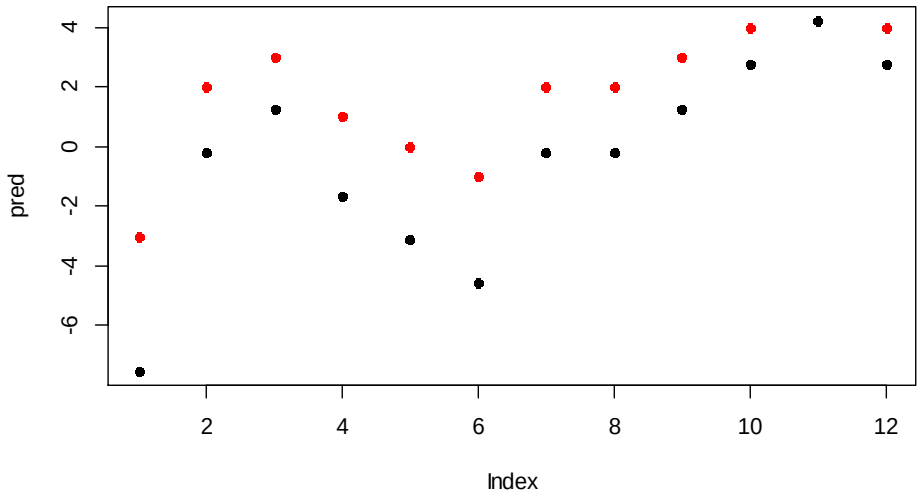
$\emptyset = \frac{D}{n-p}$ >> 1 means over dispersion (where D is deviance, n is sample size, p is the number of variables)(Zuur et al. 2009).

Residuals

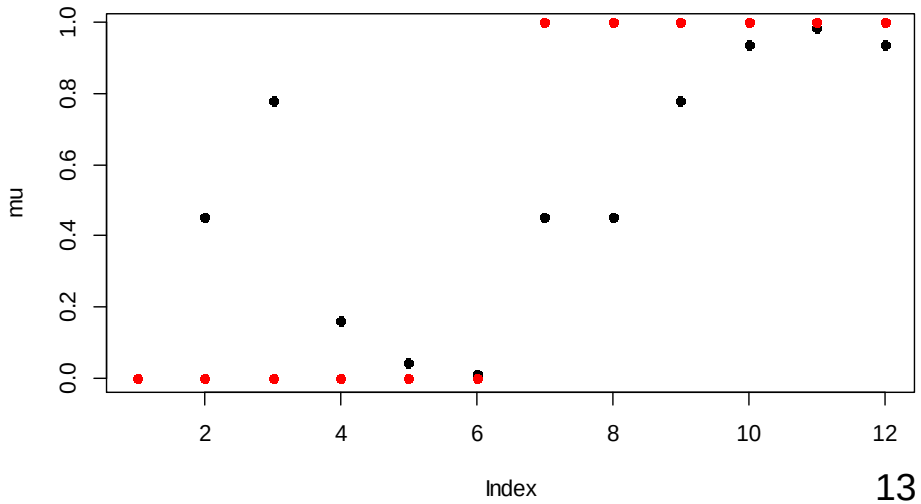
Predicted values in logistic regression

```
# sample data
y = c(0,0,0,0,0,0,1,1,1,1,1,1)
x = c(-3,2,3,1,0,-1,2,2,3,4,5,4)
fit = glm(y ~ x, family = 'binomial')
```

```
pred = predict(fit)
plot(pred, pch=16)
points(x, col="red", pch=16)
```

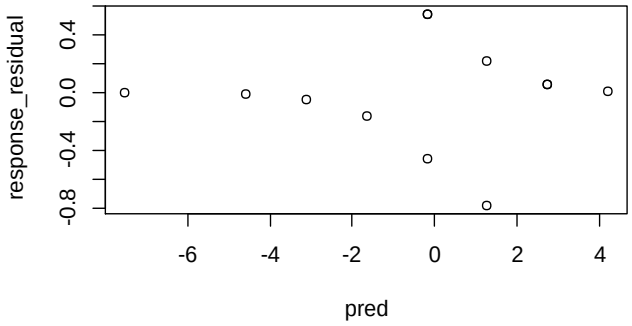



```
mu = exp(pred)/(1+exp(pred))
plot(mu, pch=16)
points(y, col="red", pch=16)
```

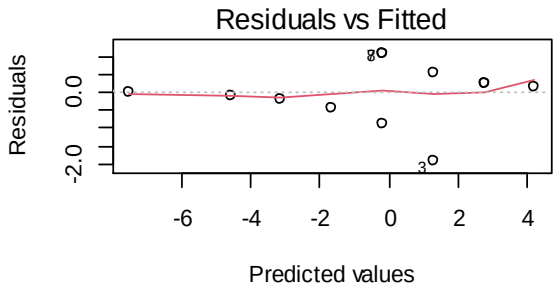


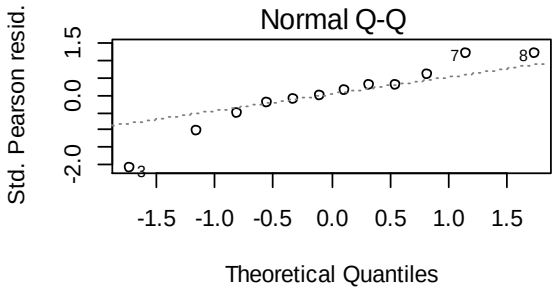
Response residuals

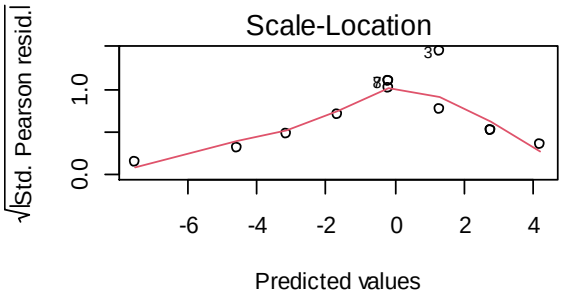
```
resid(fit, type="response")
(response_residual = y - mu) # same
plot(pred, response_residual)
```

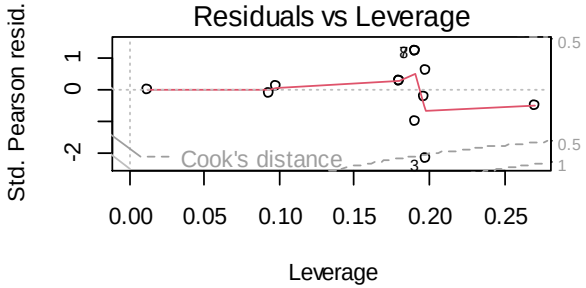



```
par(mfrow=c(2,2)); plot(fit)
```



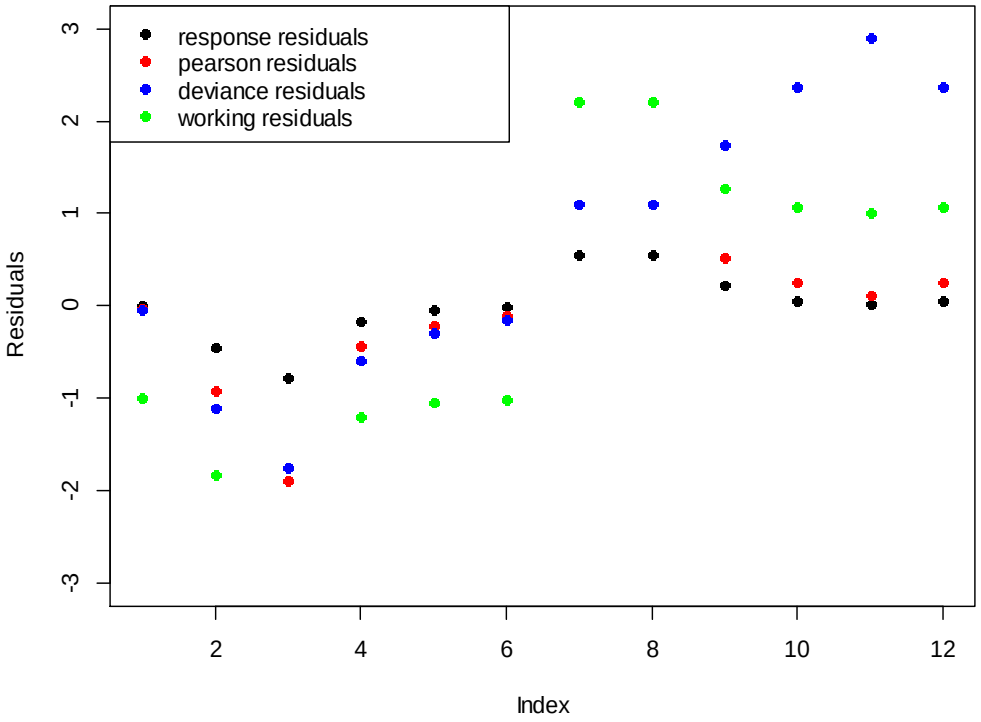






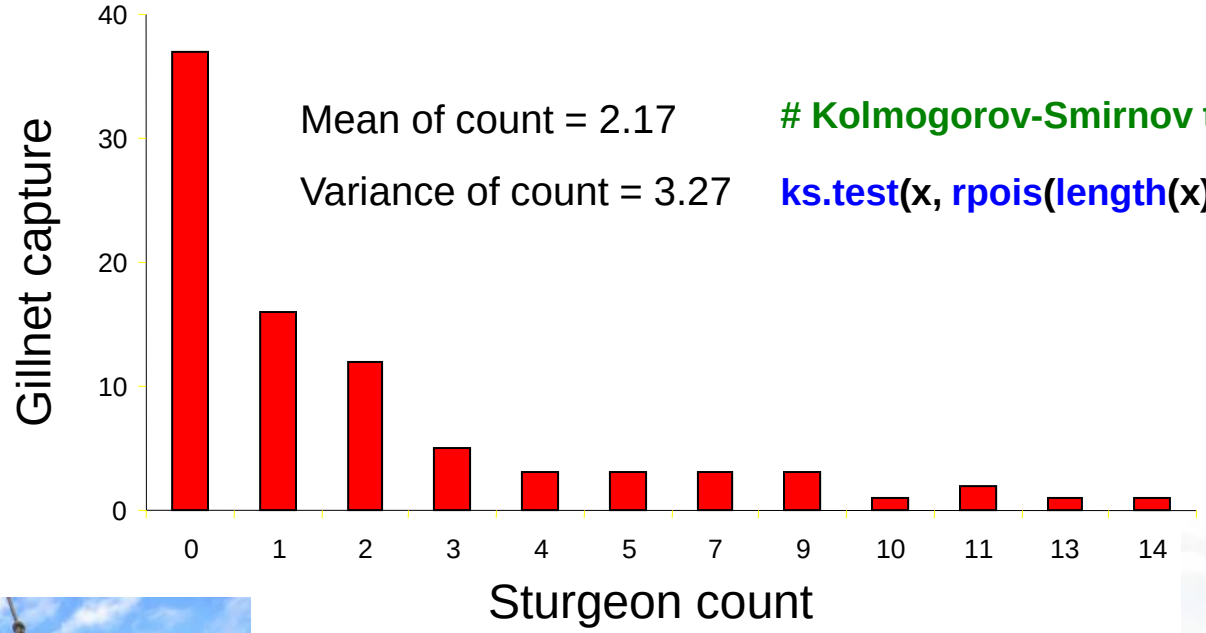
Pearson residuals, deviance residuals, and working residuals

```
plot(response_residual, col="black", pch=16, ylim=c(-3,3), ylab="Residuals")

# manually calculating the pearson residuals
resid(fit, type="pearson")
pearson_residual = (y-mu) / sqrt(mu*(1-mu)) # same
points(pearson_residual, col="red", pch=16)

# manually calculating the deviance residuals
resid(fit, type="deviance")
deviance_residual = sqrt(-2*log(1-mu))*sign(y-mu) # same
points(deviance_residual, col="blue", pch=16)

# manually calculating the working residuals
resid(fit, type="working")
working_residual = (y-mu) / (mu*(1-mu)) # same
points(working_residual, col="green", pch=16)

legend("topleft", legend=c("response residuals",
       "pearson residuals","deviance residuals","working residuals"),
       col=c("black","red","blue","green"), pch=16)
```



Cases of generalized linear models

Habitat use of shortnose sturgeon - gill net capture

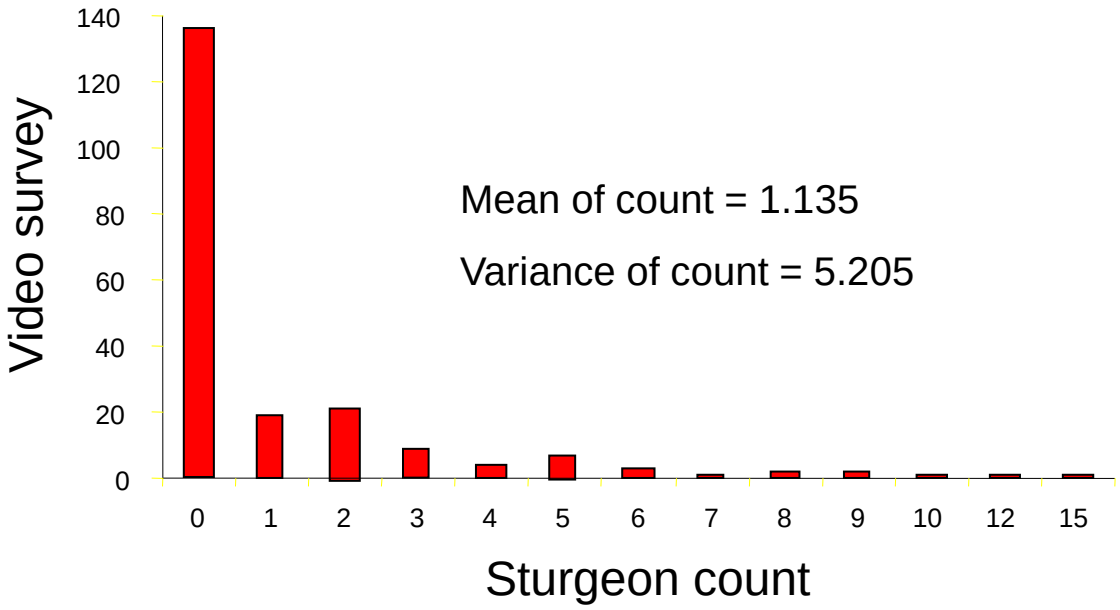


Mean of count = 2.17

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)



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)

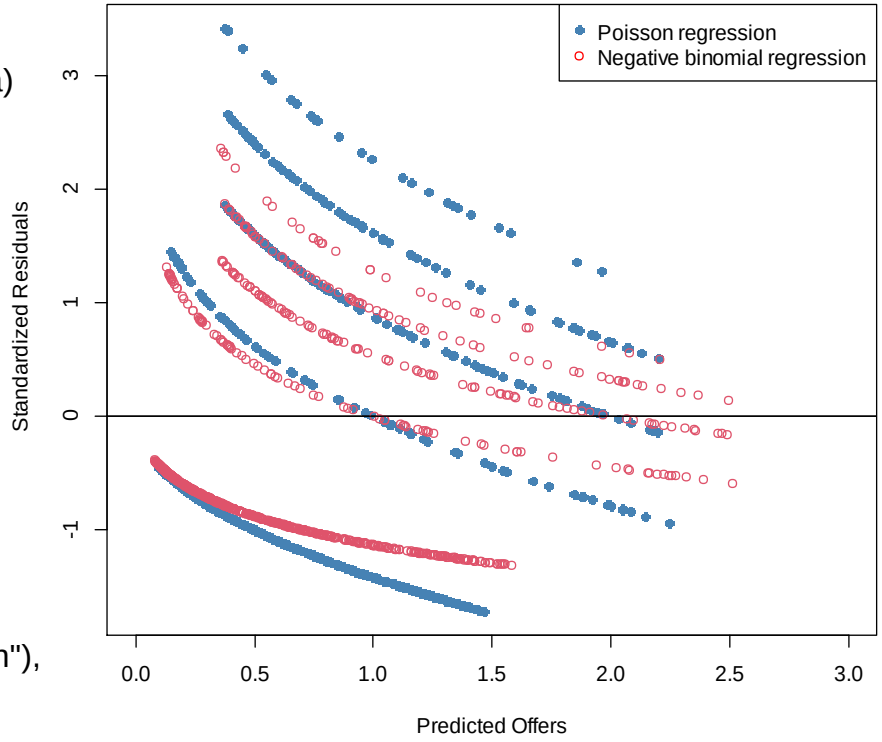
glm.nb (count.video ~ depth + substrate, data = data.count.video)

Fish tracking data (sonar telemetry)

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

Negative binomial regression vs. Poisson regression

```
# make this example reproducible (modified based on Zach)
set.seed(1)

# create dataset
data <- data.frame(offers = c(rep(0, 700), rep(1, 100), rep(2, 100), rep(3, 70), rep(4, 30)),
                   division = sample(c('A', 'B', 'C'), 100, replace = TRUE),
                   exam = c(runif(700, 60, 90), runif(100, 65, 95), runif(200, 75, 95)))
head(data)

# fit Poisson regression model
p_model <- glm(offers ~ division + exam, family = 'poisson', data = data)

# fit negative binomial regression model
library(MASS)
nb_model <- glm.nb(offers ~ division + exam, data = data)

# Residual plot for Poisson regression
p_res <- resid(p_model)
plot(fitted(p_model), p_res, col='steelblue', pch=16, xlim=c(0,3),
    xlab='Predicted Offers', ylab='Standardized Residuals')
abline(0,0)

# Residual plot for negative binomial regression
nb_res <- resid(nb_model)
points(fitted(nb_model), nb_res, col=2)
legend("topright", c("Poisson regression", "Negative binomial regression"),
          pch=c(16,1), col=c("steelblue", "red"))
abline(0,0)

# Perform a Likelihood Ratio Test
pchisq(2 * (logLik(nb_model) - logLik(p_model)), df = 1, lower.tail = FALSE)
```



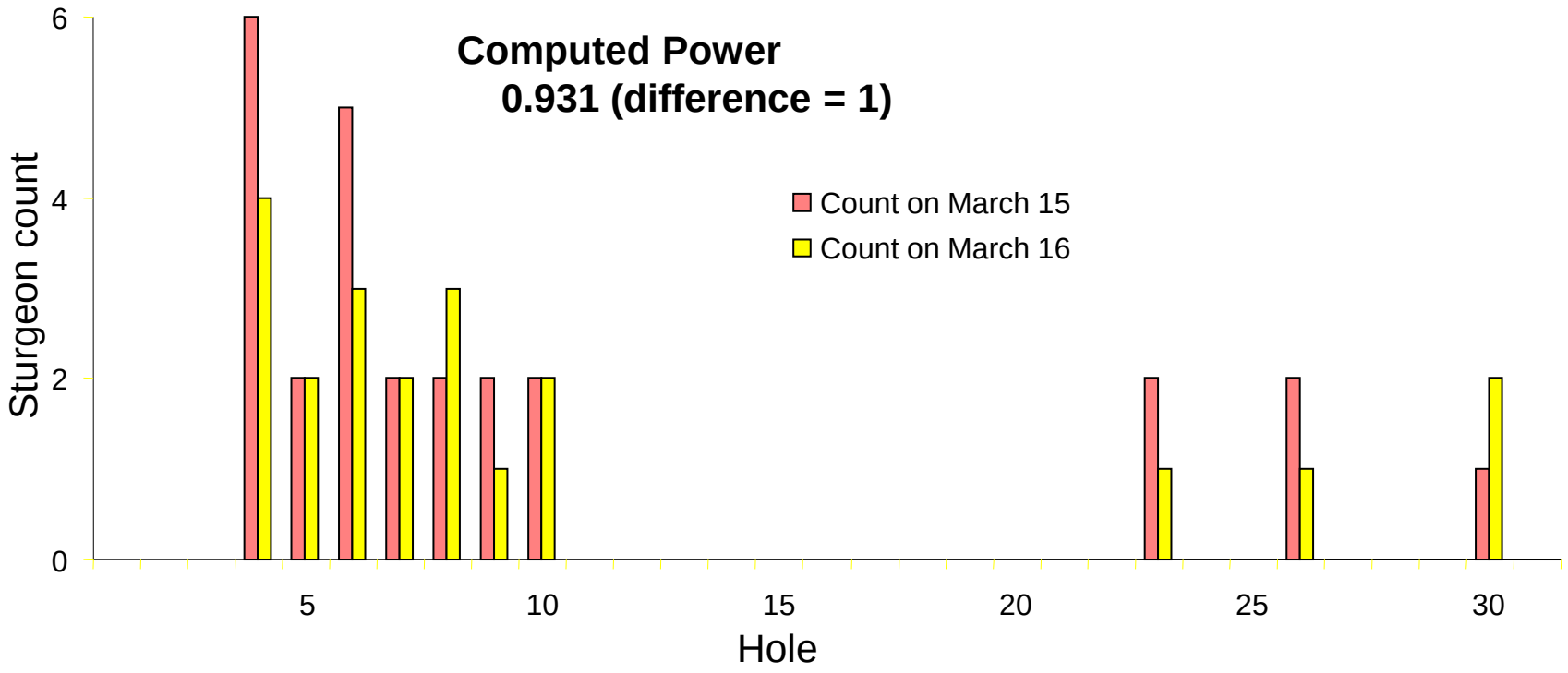
Overwintering: temporal variance

The TTEST Procedure

Difference	N	Lower CL Mean	Mean	Upper CL Mean	Lower CL Std Dev	Std Dev	Upper CL Std Dev	Std Err
Mar15 - Mar16	31	-0.073	0.1613	0.3951	0.5095	0.6375	0.8522	0.1145

T-Tests

Difference	DF	t Value	Pr > |t|
Mar15 - Mar16	30	1.41	0.1692



Results

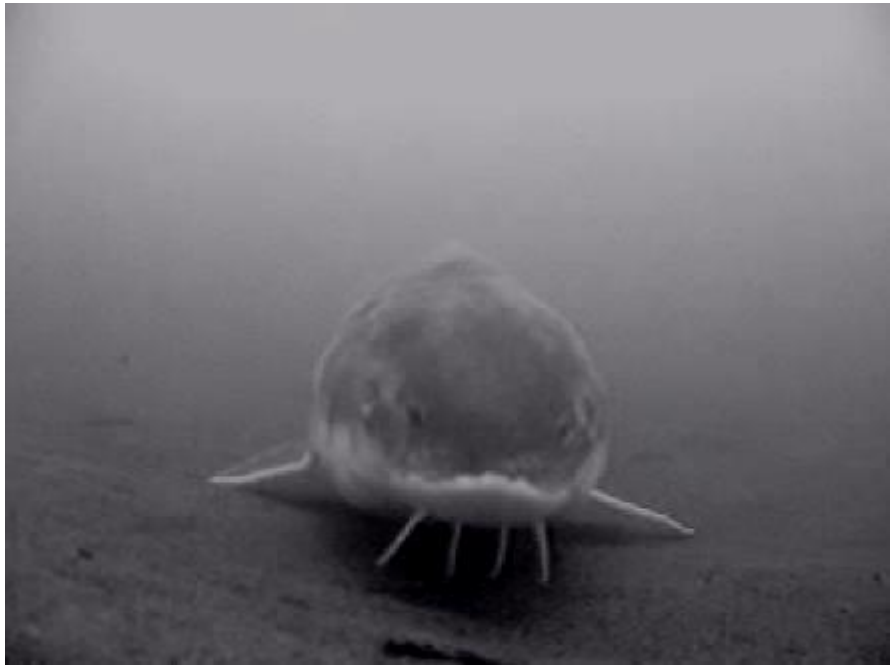
$$\log(\mu) = -2.91 + 0.75\text{Depth} - 0.8\text{Substrate}$$

where $\mu = E(\text{density})$

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: overwintering habitat use

- 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



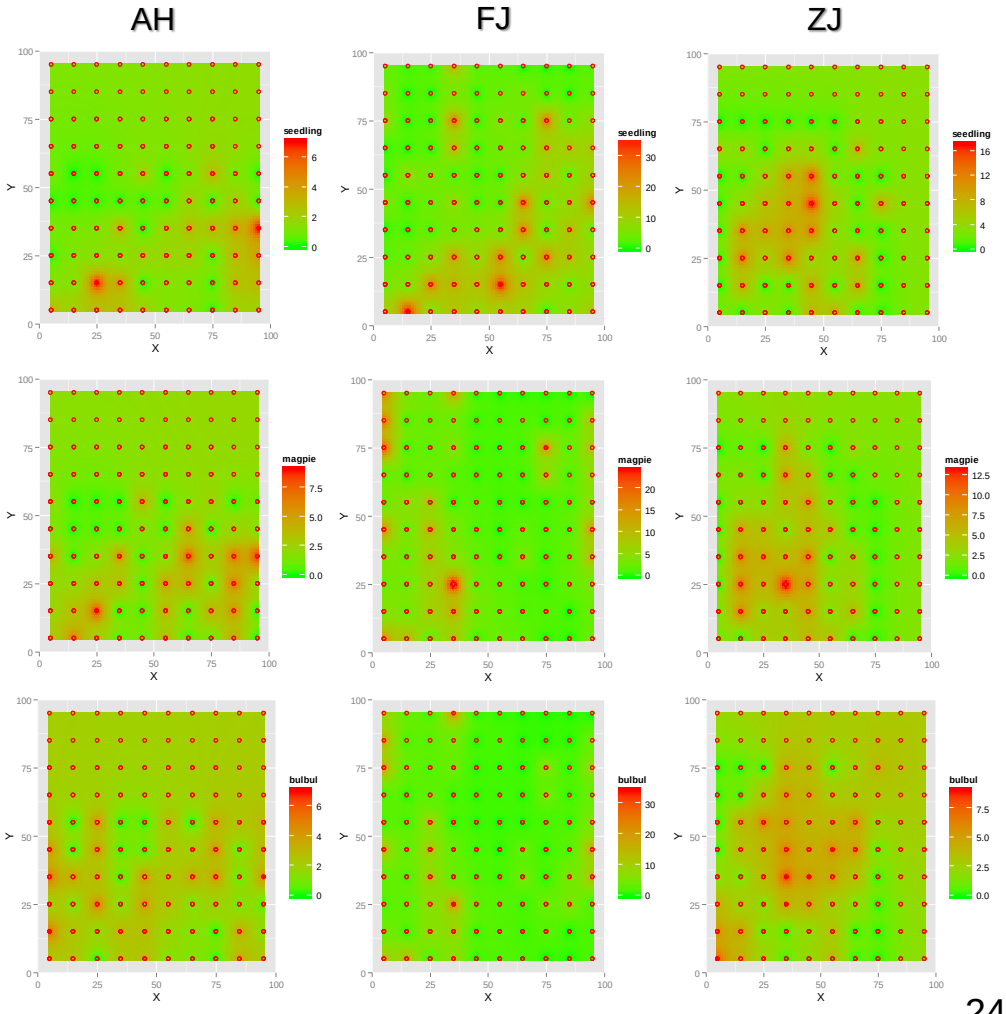
Li, X. H., Matthew K. Litvak, and John E. Hughes Clarke. 2007. Overwintering habitat use of shortnose sturgeon: defining critical habitat using a novel underwater video survey and modeling approach. Canadian Journal of Fisheries and Aquatic Sciences 64: 1248-1257

Generalized linear mixed model (GLMM): spatial autocorrelation

An extension to the generalized linear model in which the linear predictor contains random effects in addition to the usual fixed effects.

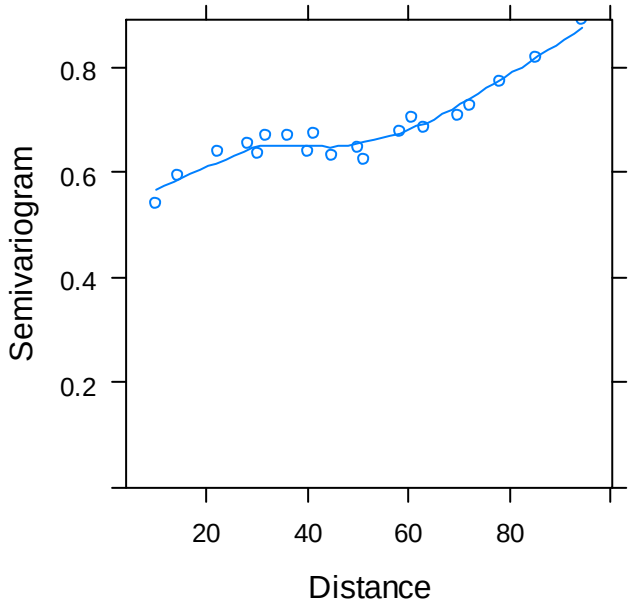
yew[1:12,]

loc	x	y	seedling	year	magpie	bulbul	site
1	5	5	10	2011	12	14	FJ
2	5	15	3	2011	3	5	FJ
3	5	25	1	2011	0	2	FJ
4	5	35	0	2011	1	5	FJ
5	5	45	5	2011	4	23	FJ
6	5	55	0	2011	0	2	FJ
7	5	65	0	2011	2	0	FJ
8	5	75	8	2011	15	21	FJ
9	5	85	5	2011	12	23	FJ
10	5	95	3	2011	10	14	FJ
11	15	5	35	2011	5	23	FJ
12	15	15	4	2011	2	1	FJ



Quantify spatial autocorrelation

```
# spatial aotocorrelation
library(nlme)

f1 <- formula(seedling ~ magpie + bulbul + site)

B1.gls <- gls(f1, data = yew)
Vario.gls <- Variogram(B1.gls, form =~ x + y,
                       robust = TRUE, maxDist = 200,
                       resType = "pearson")
plot(Vario.gls, smooth = T)
yew$x = yew$x + rnorm(nrow(yew),0,0.01) # avoid zero distance
yew$y = yew$y + rnorm(nrow(yew),0,0.01)
B1A <- gls(f1, correlation = corSpher(form =~ x + y, nugget = TRUE), data = yew)
B1B <- gls(f1, correlation = corLin    (form =~ x + y, nugget = TRUE), data = yew)
B1C <- gls(f1, correlation = corRatio (form =~ x + y, nugget = TRUE), data = yew)
B1D <- gls(f1, correlation = corGaus  (form =~ x + y, nugget = TRUE), data = yew)
B1E <- gls(f1, correlation = corExp    (form =~ x + y, nugget = TRUE), data = yew)
AIC(B1A, B1B, B1C, B1D, B1E)
summary(B1D) # best model
```



Generalized linear mixed model (GLMM)

```
library(lme4)
model <- glmer(seedling ~ magpie + bulbul + (1 | year) + (1 | site) + (1 | site:year),
                    data = yew, family = poisson)

summary(model)
anova(model)
```

Random effects:

Groups	Name	Variance	Std.Dev.
site:year	(Intercept)	0.0000	0.0000
site	(Intercept)	0.2909	0.5393
year	(Intercept)	0.0000	0.0000

Number of obs: 448, groups: site:year, 6; site, 3; year, 3

Fixed effects:

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	0.989870	0.313183	3.161	0.00157 **
magpie	0.039440	0.002596	15.193	< 2e-16 ***
bulbul	0.035061	0.003163	11.084	< 2e-16 ***

Analysis of Variance Table

	Df	Sum Sq	Mean Sq	F value
magpie	1	276.00	276.00	276.00
bulbul	1	122.84	122.84	122.84

Li, N., S.-b. Fang, X.-h. Li, S.-q. An, and C.-h. Lu. 2015a. Differential contribution of frugivorous birds to dispersal patterns of the endangered Chinese yew (Taxus chinensis). Scientific Reports **5**.
Li, N., X.-h. Li, S.-q. An, and C.-h. Lu. 2015b. Impact of multiple bird partners on the seed dispersal effectiveness of China's relic trees. Scientific Reports **5**.

Zero-Truncated and Zero-Inflated Models for Count Data (Zuur et al. 2009)

- Zero truncated means the response variable cannot have a value of 0.
  - ✓ A typical example from the medical literature is the duration patients are in hospital.
  - ✓ For ecological data, think of response variables like the time a whale is at the surface before re-submerging, counts of fin rays on fish (e.g. used for stock identification), dolphin group size, age of an animal in years or months, or the number of days that carcasses of road-killed animals remain on the road.

- Zero inflated data are more common in ecological research. In such cases the response variable contains more zeros than expected, based on the Poisson or negative binomial distribution.

Zero-Truncated Poisson distribution

PDF for Poisson distribution:

$$\mathrm{f}(y_i;\mu_i|y_i \geq 0) = \frac{\mu^{y_i} \times e^{-\mu_i}}{y_i!}$$

Probability of being 0:

$$\mathrm{f}(0;\mu_i) = \frac{\mu^0 \times e^{-\mu_i}}{0!} = e^{-\mu_i}$$

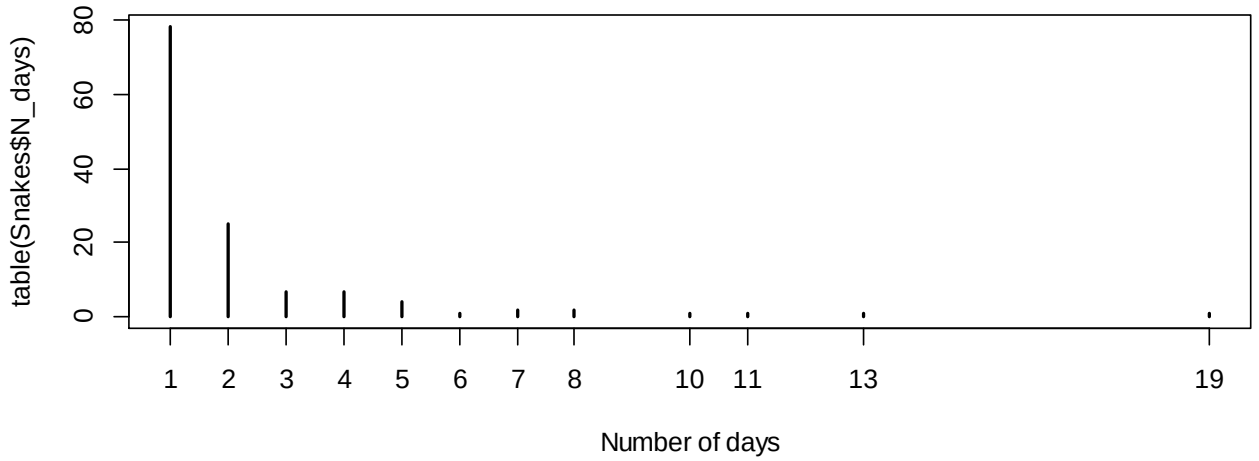
PDF for zero-truncated Poisson distribution:

$$\mathrm{f}(y_i;\mu_i|y_i > 0) = \frac{\mu^{y_i} \times e^{-\mu_i}}{(1 - e^{-\mu_i}) \times y_i!}$$

Zero-truncated GLM: data

- Zuur et al. Mixed Effect Models page 265
- The response variable is the number of days that carcasses of road-killed animals (*Coronella girondica*, *Coluber hippocrepis*, *Elaphe scalaris*, and *Macroprotodon cucullatus*) remain on the road. The data set contains 130 observations.

ID	Road	Month	Season	N_days	Species	Road_Loc	Size_cm	PDayRain	Tot_Rain	Temp_avg
2176	EN114	Jul	Summer	4	Coluberhippocrepis	L	115	0.75	15	24.6
2448	EN114	Aug	Summer	1	Elaphescalaris	L	150	0	0	27.4
2917	EN114	Oct	Autumn	4	Elaphescalaris	L	150	1	40.2	19.1
2927	EN114	Oct	Autumn	2	Elaphescalaris	L	150	1	35.6	17.8
2845	EN114	Oct	Autumn	1	Elaphescalaris	L	150	0	0	22.3
2849	EN114	Oct	Autumn	1	Elaphescalaris	L	150	0	0	22.3



R code

```
# Zuur et al.  2009. Mixed Effect Models page 268
Snakes = read.table('D:/softwares/R/library/AED/data/Snakes.txt', header=T)
library(MASS)
M2A <- glm.nb(N_days ~ PDayRain + Tot_Rain + Road_Loc +
           PDayRain:Tot_Rain, data = Snakes)
library(VGAM)
M2B <- vglm(N_days    ~ PDayRain + Tot_Rain + Road_Loc +
          PDayRain:Tot_Rain, family = negbinomial, data = Snakes)

# Zero-truncated GLM
M3A <- vglm(N_days    ~ PDayRain + Tot_Rain + Road_Loc +
          PDayRain:Tot_Rain, family = posnegbinomial,
          control = vglm.control(maxit = 100), data = Snakes)
```

The `family = posnegbinomial` argument ensures that a zero-truncated NB model is applied

Compare zero-truncated and untruncated GLMs

```
# Zuur et al.  2009. Mixed Effect Models page 269
Z <- cbind(coef(M2A), coef(M3A)[-2])
ZSE <- cbind(sqrt(diag(vcov(M2A))), sqrt(diag(vcov(M3A))[-1]))
Comp <- cbind(Z[,1], Z[,2], ZSE[,1], ZSE[,2])
Comb <- round(Comp, digits = 3)
colnames(Comb) <- c("NB", "Trunc.NB", "SE NB", "SE Trunc.NB")
Comb
```

The coef command extracts the estimated parameters and the vcov the covariance matrix of the estimated parameters. The diagonal elements of this matrix are the estimated variances; hence, the square root of these gives the standard errors. [-2] ensures that only regression parameters are extracted and not the parameter *k*.

	NB	Trun.NB	SE NB	SE Trunc.NB
(Intercept)	0.365	-2.035	0.112	0.267
PDayRain	-0.001	0.114	0.193	0.449
Tot_Rain	0.12	0.254	0.02	0.065
Road_LocV	0.449	1.077	0.148	0.368
PDayRain:Tot_Rain	-0.109	-0.234	0.022	0.07

Zero inflated GLM: why so many zeros

- The habitat is not suitable
- Poor experimental design or sampling practices
  - e.g. counting the number of puffins on the cliffs in the winter. It is highly likely that all samples will be 0 as it is the wrong season and they are all at sea. Another design error is sampling for too short a time period or sampling too small an area.
- Observer error
  - Some bird species look similar, or are difficult to detect. The less experienced the observer, the more likely he/she will end up with zero counts for bird species that are difficult to identify. Alternatively, the observer may be highly experienced, but it is extremely difficult to detect a tiny dark bird in a dark field on a dark day.
- The ‘animal’ error
  - This means that the habitat is suitable, but the site is not used.

Zuur et al. 2009. Mixed Effect Models page 270

ZIP (Poisson) and ZINB (negative binomial) Models

Let Pr(*Yi*) be the probability that at site *i* where an animal is recorded

Pr(*Yi* = 0) = Pr(False zeros) + (1 − Pr(False zeros))×Pr(Count process gives a zero)

- We divide the data in two imaginary groups; the first group contains only zeros (the false zeros). This group is also called the observations with zero mass.

- The second group is the count data, which may produce zeros (true zeros) and as well as values larger than zero.

- Note that we are not actively splitting the data in two groups; it is just an *assumption* that we have these two groups. We do not know which of the observations with zeros belong to a specific group. All that we know is that the non-zeros (the counts) are in group 2.

ZIP Model

Assuming the count *Yi* follow a Poisson distribution with expectation μi

PDF for Poisson distribution:

$$\mathrm{f}(y_i;\mu_i|y_i \geq 0) = \frac{\mu^{y_i} \times e^{-\mu_i}}{y_i!}$$

Probability of being 0:

$$\mathrm{f}(0;\mu_i) = \frac{\mu^0 \times e^{-\mu_i}}{0!} = e^{-\mu_i}$$

Assuming the probability that *Yi* is a false zero is binomially distributed with probability *πi*, the following is the probability distribution for a ZIP model.

$$\mathrm{f}(y_i = 0) = \pi_i + (1-\pi_i) \times e^{-\mu_i}$$

$$\mathrm{f}(Y_i = y_i\,|y_i > 0) = (1-\pi_i) \times \frac{\mu^{y_i} \times e^{-\mu_i}}{y_i!}$$

ZIP Model

In Poisson GLM, we model the mean *μi* of the positive count data as

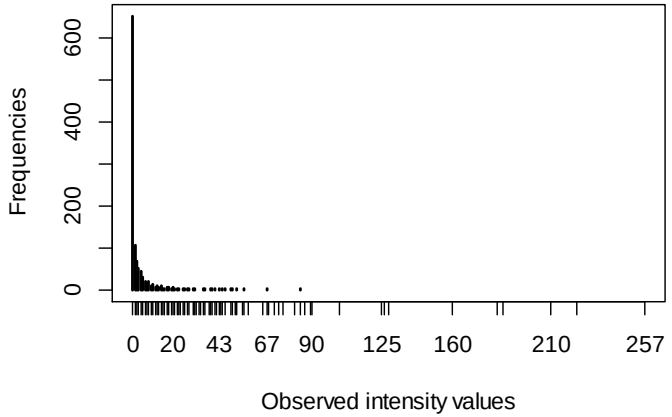
$$\mu_i = e^{\alpha+\beta_1\times X_{i1}+\cdots+\beta_q\times X_{q1}}$$

Hence, covariates are used to model the positive counts. What about the probability of having a false zero, *πi*? The easiest approach is to use a logistic regression:

$$p_i = \frac{e^{\alpha+\beta_1\times X_{i1}+\cdots+\beta_q\times X_{q1}}}{1+e^{\alpha+\beta_1\times X_{i1}+\cdots+\beta_q\times X_{q1}}}$$

R code

```
# Zuur et al.  2009. Mixed Effect Models page 270
ParasiteCod = read.table('D:/softwares/R/library/AED/data/ParasiteCod.txt', header=T)
ParasiteCod$fArea <- factor(ParasiteCod$Area)
ParasiteCod$fYear <- factor(ParasiteCod$Year)
I1 <- is.na(ParasiteCod$Intensity) |  is.na(ParasiteCod$fArea) |
          is.na(ParasiteCod$fYear) |  is.na(ParasiteCod$Length)
ParasiteCod2 <- ParasiteCod[!I1, ]
plot(table(ParasiteCod2$Intensity),
      ylab = "Frequencies",
      xlab = “Observed intensity values”) #Figure to the right
```




```
library(pscl)
f1 <- formula(Intensity ~ fArea*fYear + Length | fArea * fYear + Length)
Zip1 <- zeroinfl(f1, dist = "poisson",
                 link = "logit", data = ParasiteCod2)
summary(Zip1)
```

Results

Count model coefficients (poisson with log link):

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	3.9554	0.0609	64.9210	0
fArea2	0.3132	0.0713	4.3930	1.12E-05
fArea3	-0.2819	0.0731	-3.8540	0.000116
fArea4	0.8222	0.0458	17.9440	0
fYear2000	0.1358	0.0736	1.8450	0.06508
fYear2001	-0.9147	0.1937	-4.7230	2.33E-06
Length	-0.0368	0.0010	-37.5220	0
fArea2:fYear2000	-0.6000	0.1343	-4.4680	7.91E-06
fArea3:fYear2000	0.8350	0.1042	8.0120	1.13E-15
fArea4:fYear2000	0.2673	0.0819	3.2630	0.001102
fArea2:fYear2001	0.9711	0.2103	4.6180	3.87E-06
fArea3:fYear2001	1.0107	0.2115	4.7770	1.78E-06
fArea4:fYear2001	0.8913	0.1974	4.5150	6.34E-06

Zero-inflation model coefficients (binomial with logit link):

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	0.0086	0.3003	0.029	0.9771
fArea2	1.3230	0.2861	4.624	3.77E-06
fArea3	1.4402	0.2452	5.874	4.26E-09
fArea4	-0.3064	0.2719	-1.127	0.2597
fYear2000	-0.3924	0.3474	-1.129	0.2587
fYear2001	2.5483	0.4396	5.797	6.75E-09
Length	-0.0089	0.0048	-1.854	0.0637
fArea2:fYear2000	0.0633	0.5124	0.123	0.9017
fArea3:fYear2000	-0.8522	0.4538	-1.878	0.0604
fArea4:fYear2000	-0.8594	0.5947	-1.445	0.1484
fArea2:fYear2001	-2.6229	0.5387	-4.869	1.12E-06
fArea3:fYear2001	-2.6347	0.5061	-5.206	1.93E-07
fArea4:fYear2001	-2.4314	0.5243	-4.637	3.53E-06

Conditional logistic regression

Used in matched case–control studies, e.g. a case (read diseased) subject is matched with a number of controls (read non-diseased) based on some matching or confounding factors (Breslow et al. 1978; Breslow and Day 1980).

Conditional logistic regression: case study

The family of the black snub-nosed monkey consist of one male and several females. Single males (usually juveniles) occasionally challenge the adult male in a family in order to take control.

In one population of the black snub-nosed monkey, there are about 6-7 families with 42-60 individuals. In the past 10 years, the family members have kept changing. In total, 48 challenge behaviors were observed in this period.

The probability of a male in a family being challenged, compared with other 5-6 unchallenged males, is associated with the number of females in the family (F_tot), number of available females (not in pregnant or lactation period) (F_ava), and the rank of the male.

Data from Dr. Pingfen Zhu

R code and results

```
library(survival)
fit = clogit(Attempt ~ F_tot * F_ava * Rank + strata(Stratum), data=D)
summary(fit)
```

	coef	exp(coef)	se(coef)	z	Pr(>|z|)
F_tot	-2.08E+00	1.25E-01	6.94E-01	-3.003	0.00267
F_ava	-6.51E+00	1.49E-03	2.29E+00	-2.846	0.00443
Rank	-1.20E+01	6.18E-06	4.33E+00	-2.77	0.0056
F_tot:F_ava	1.78E+00	5.91E+00	6.04E-01	2.941	0.00327
F_tot:Rank	3.32E+00	2.77E+01	1.19E+00	2.788	0.0053
F_ava:Rank	1.05E+01	3.60E+04	3.90E+00	2.693	0.00708
F_tot:F_ava:Rank	-2.89E+00	5.55E-02	1.06E+00	-2.733	0.00628

Stratum	F_tot	F_ava	Rank	Attempt
1	3	2	0.709	1
1	4	0	1	0
1	6	2	0.132	0
1	4	3	0.709	0
1	5	1	0	0
1	2	1	0.046	0
1	3	2	0.046	0
2	4	3	0	1
2	4	0	1	0
2	6	2	0.841	0
2	3	2	0.709	0
2	5	1	0	0
2	2	1	0.046	0
2	3	2	0.046	0
3	3	2	0.181	1
3	4	0	0.947	0
3	6	1	1	0
3	3	1	0.713	0
3	4	3	0.144	0
3	5	1	0.144	0
3	2	1	0	0

Data from Dr. Pingfen Zhu

Multinomial logistic regression

It is a classification method that generalizes logistic regression to multiclass problems, i.e. with more than two possible discrete outcomes.

It assumes a linear combination of the observed features and some problem-specific parameters can be used to determine the probability of each particular outcome of the dependent variable.

It has some other names:

- polytomous logistic regression
- multiclass logistic regression
- softmax regression
- multinomial logit

Rationale

For *K* possible outcomes, running *K*-1 independent binary logistic regression models, in which one outcome is chosen as a “pivot” and then the other *K*-1 outcomes are separately regressed against the pivot outcome (e.g. the K’s outcome).

$$\ln \frac{\Pr(Y_i = 1)}{\Pr(Y_i = K)} = \beta_1 X_i \qquad \ln \frac{\Pr(Y_i = 2)}{\Pr(Y_i = K)} = \beta_2 X_i \qquad \ln \frac{\Pr(Y_i = K-1)}{\Pr(Y_i = K)} = \beta_{K-1} X_i$$

All K of the probabilities must sum to one, we have:

$$\Pr(Y_i = K) = \frac{1}{1 + \sum_{k=1}^{K-1} e^{\beta_k X_i}} \qquad \Pr(Y_i = 1) = \frac{e^{\beta_1 X_i}}{1 + \sum_{k=1}^{K-1} e^{\beta_k X_i}}$$

Multinomial logistic regression

```
head(mtcars)
mtcars$cyl = as.factor(mtcars$cyl)
mtcars$gear = as.factor(mtcars$gear)

# Re-leveling data
mtcars$cyl <- relevel(mtcars$cyl, ref = "4")
table(mtcars$cyl); table(mtcars$gear)
```

	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	card
Mazda RX4	21.0	6	160.0	110	3.90	2.620	16.46	0	1	4	4
Mazda RX4 Wag	21.0	6	160.0	110	3.90	2.875	17.02	0	1	4	4
Datsun 710	22.8	4	108.0	93	3.85	2.320	18.61	1	1	4	1
Hornet 4 Drive	21.4	6	258.0	110	3.08	3.215	19.44	1	0	3	1
Hornet Sportabout	18.7	8	360.0	175	3.15	3.440	17.02	0	0	3	2
Valiant	18.1	6	225.0	105	2.76	3.460	20.22	1	0	3	1

```
options(digits=4)
library("nnet")
test <- multinom(gear ~ wt + cyl, data = mtcars)
summary(test)  # Coefficients and SE
```

```
Coefficients:
  (Intercept)    wt    cyl6    cyl8
4       5.794 -1.456 -0.4469 -10.757
5      12.231 -5.374  3.2431   5.352

Std. Errors:
  (Intercept)    wt   cyl6   cyl8
4       4.725 1.709  1.673 55.808
5       5.680 2.557  2.979  4.106

Residual Deviance: 33.46
AIC: 49.46
```

```
# significance
z <- summary(test)$coefficients / summary(test)$standard.errors
z
p <- (1 - pnorm(abs(z), 0, 1))*2; p
exp(coef(test))
```

```
  (Intercept)    wt    cyl6    cyl8
4       1.226 -0.8518 -0.2672 -0.1927
5       2.153 -2.1021  1.0885  1.3035

  (Intercept)    wt   cyl6   cyl8
4     0.22008 0.39431 0.7893 0.8472
5     0.03131 0.03555 0.2764 0.1924
```

Multinomial logistic regression

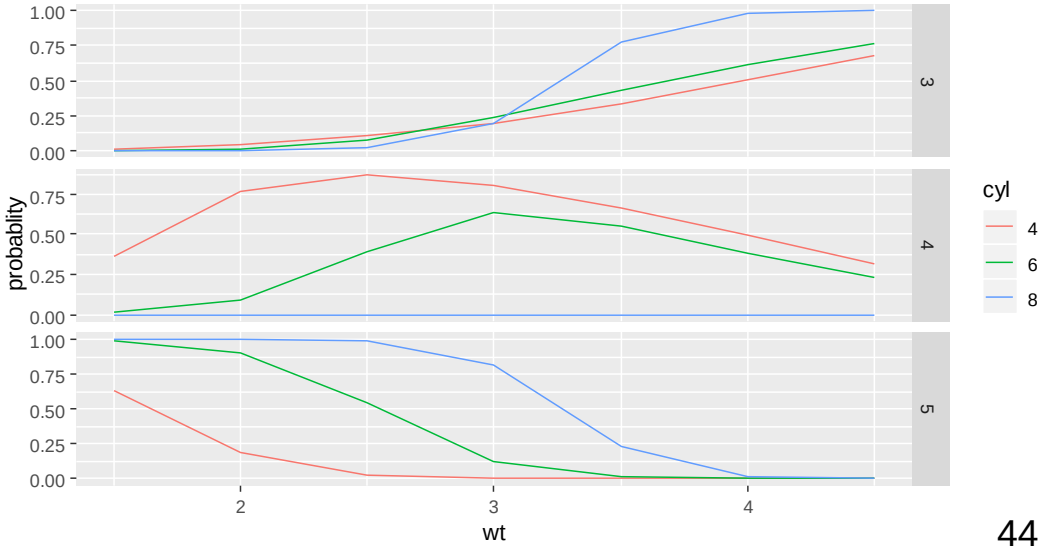
```
# prediction
head(fitted(test)) # fitted values
expanded = expand.grid(cyl = c("4", "6", "8"),
                       wt = c(1.5, 2, 2.5, 3, 3.5, 4, 4.5))
predicted = predict(test,expanded, type = "probs")
bpp = cbind(expanded, predicted)
```

	3	4	5
Mazda RX4	0.10352	0.47965	0.41683
Mazda RX4 Wag	0.19159	0.61246	0.19594
Datsun 710	0.07692	0.86243	0.06064
Hornet 4 Drive	0.32121	0.62595	0.05284
Hornet Sportabout	0.71193	0.00003	0.28804
Valiant	0.41524	0.56645	0.01831

```
# “melts” data with the purpose of each row
# being a unique id-variable combination
library("reshape2")
bpp2 = melt(bpp, id.vars = c("cyl", "wt"), value.name = "probablity")
head(bpp2)
```

	cyl	wt	variable	probablity
1	4	1.5	3	9.742e-03
2	6	1.5	3	5.949e-04
3	8	1.5	3	7.326e-05
4	4	2.0	3	4.298e-02
5	6	2.0	3	7.990e-03
6	8	2.0	3	1.075e-03

```
library("ggplot2")
ggplot(bpp2, aes(x = wt, y = probablity, colour = cyl))
 + geom_line() + facet_grid(variable ~ ., scales="free")
```



Ordered logistic regression

```
require(foreign); require(ggplot2); require(MASS)
require(Hmisc); require(reshape2)

mtcars$gear = ordered(mtcars$gear, levels = c(3, 4, 5)); mtcars$gear;
str(mtcars)

m <- polr(gear ~ wt + cyl, data = mtcars, Hess=TRUE) # require(MASS)

summary(m)
```

Coefficients:

	Value	Std. Error	t value
wt	-1.723	0.906	-1.902
cyl6	0.735	1.147	0.641
cyl8	-0.535	1.526	-0.351

Intercepts:

	Value	Std. Error	t value
3|4	-5.538	2.325	-2.382
4|5	-2.906	2.076	-1.400

Residual Deviance: 47.97
AIC: 57.97

```
ctable <- coef(summary(m))
p <- pnorm(abs(ctable[, "t value"]), lower.tail = FALSE) * 2; p
ctable <- cbind(ctable, "p value" = p); ctable
```

	Value	Std. Error	t value	p value
wt	-1.7234	0.9063	-1.9017	0.05721
cyl6	0.7353	1.1470	0.6411	0.52148
cyl8	-0.5355	1.5264	-0.3508	0.72574
3|4	-5.5382	2.3254	-2.3816	0.01724
4|5	-2.9065	2.0764	-1.3998	0.16158

```
ci <- confint(m);ci # confidence intervals
exp(coef(m))
exp(cbind(OR = coef(m), ci)) ## OR and CI
# enhance this model to obtain better prediction estimates
summary(update(m, method = "probit",    Hess = TRUE), digits = 3)
summary(update(m, method = "logistic",  Hess = TRUE), digits = 3)
summary(update(m, method = "cloglog",  Hess = TRUE), digits = 3)
```

	OR	2.5 %	97.5 %
wt	0.1785	0.02459	0.8866
cyl6	2.0861	0.22591	22.0358
cyl8	0.5854	0.02649	12.3334

Ordered logistic regression

```
# mixed effect model
library(ordinal)
fmm1 <- clmm(cyl ~ wt + (1|gear), data = mtcars)

summary(fmm1)
fmm2 <- clmm(cyl ~ wt + (1|gear), data = mtcars,
          link = "probit", threshold = "equidistant")

summary(fmm2)
```

```
> summary(fmm1)
Cumulative Link Mixed Model fitted with the Laplace approximation

formula: cyl ~ wt + (1 | gear)
data:    mtcars

 link  threshold nobs logLik  AIC    niter       max.grad   cond.H
 logit flexible  32   -14.64  37.29 117(237) 1.90e-06   1.0e+03

Random effects:
 Groups Name        Variance Std.Dev.
 gear   (Intercept)  1.38     1.17
Number of groups:  gear 3

Coefficients:
   Estimate Std. Error z value Pr(>|z|)
wt     5.61       1.83    3.06   0.0022 **

Threshold coefficients:
    Estimate Std. Error z value
4|6    15.35       5.28    2.91
6|8    18.83       6.20    3.03
```

Complete or quasi-complete separation in logistic regression

```
Y  <-  c(0, 0, 0, 0, 0,  1, 1, 1, 1, 1)
X1 <- c(1, 1, 2, 3, 3,  5, 6, 7, 9, 9)
X2 <- c(1, 2, 1, 1, 5,  4, 1, 0, 3, 6)
fit  <- glm(Y ~ X1 + X2, family = binomial)
```

```
Warning message:
glm.fit: fitted probabilities numerically 0 or 1 occurred
```

We can see that observations with Y = 0 all have values of X1<=3 and observations with Y = 1 all have values of X1>3. In other words, X1 separates Y perfectly.

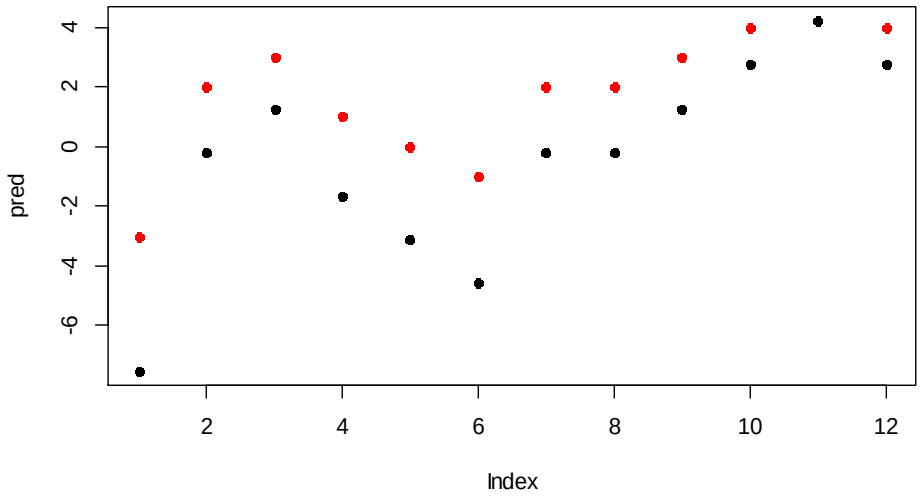
Now the maximum likelihood estimate for X1 does not exist. For this example, the larger the coefficient for X1, the larger the likelihood. The coefficient for X1 should be as large as it can be, which is infinity.

Residuals

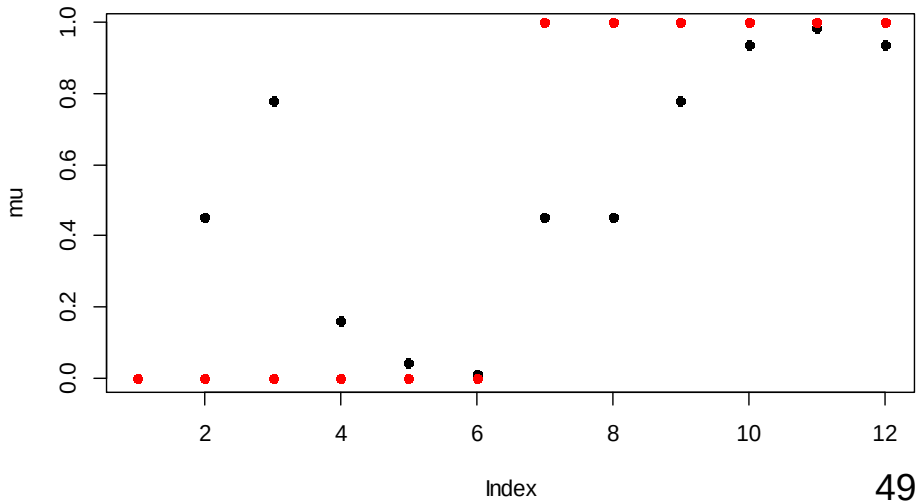
Predicted values in logistic regression

```
# sample data
y = c(0,0,0,0,0,0,1,1,1,1,1,1)
x = c(-3,2,3,1,0,-1,2,2,3,4,5,4)
fit = glm(y ~ x, family = 'binomial')
```

```
pred = predict(fit)
plot(pred, pch=16)
points(x, col="red", pch=16)
```

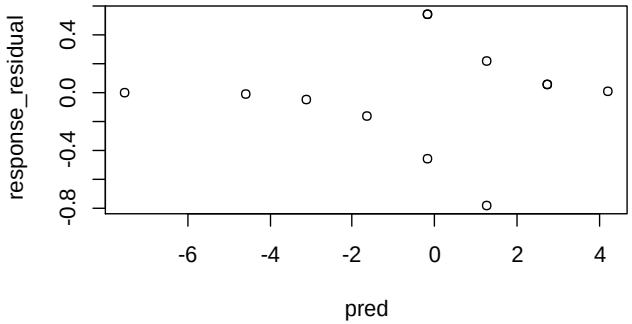



```
mu = exp(pred)/(1+exp(pred))
plot(mu, pch=16)
points(y, col="red", pch=16)
```

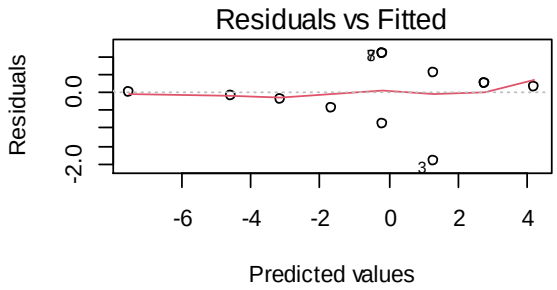


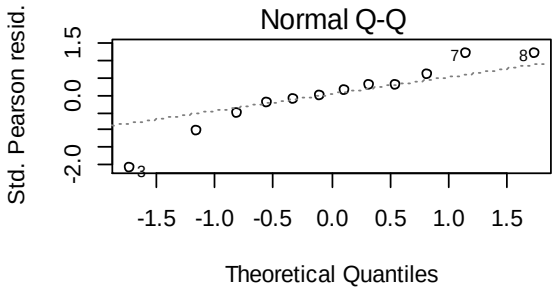
Response residuals

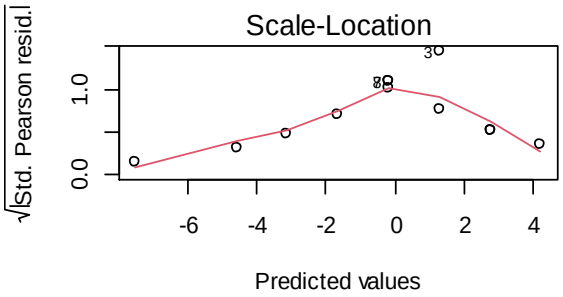
```
resid(fit, type="response")
(response_residual = y - mu) # same
plot(pred, response_residual)
```

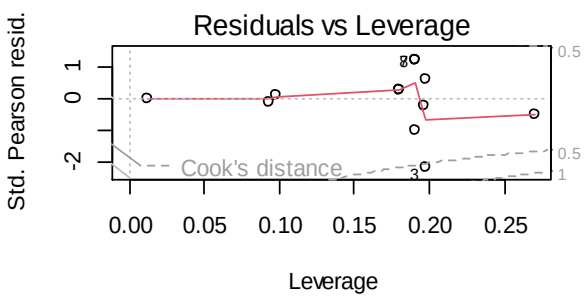



```
par(mfrow=c(2,2)); plot(fit)
```



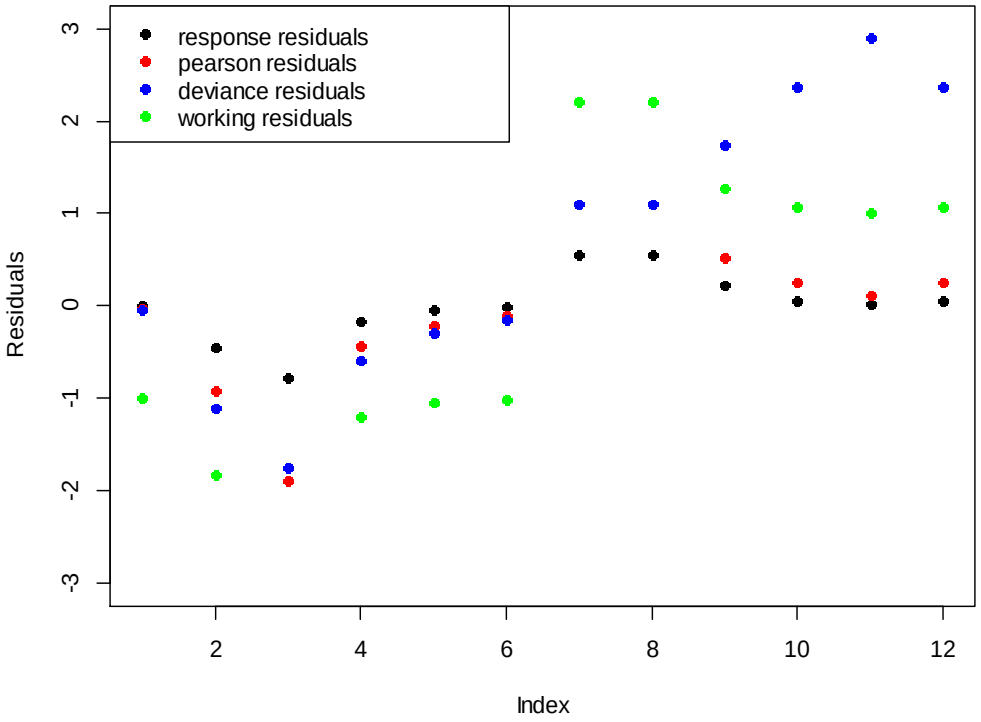






Pearson residuals, deviance residuals, and working residuals

```
plot(response_residual, col="black", pch=16, ylim=c(-3,3), ylab="Residuals")

# manually calculating the pearson residuals
resid(fit, type="pearson")
pearson_residual = (y-mu) / sqrt(mu*(1-mu)) # same
points(pearson_residual, col="red", pch=16)

# manually calculating the deviance residuals
resid(fit, type="deviance")
deviance_residual = sqrt(-2*log(1-mu))*sign(y-mu) # same
points(deviance_residual, col="blue", pch=16)

# manually calculating the working residuals
resid(fit, type="working")
working_residual = (y-mu) / (mu*(1-mu)) # same
points(working_residual, col="green", pch=16)

legend("topleft", legend=c("response residuals",
       "pearson residuals","deviance residuals","working residuals"),
       col=c("black","red","blue","green"), pch=16)
```



Model comparison

Popular species distribution models, history, complexity levels, popularity in climate change studies, types of species data, and reference papers (Li & Wang 2013) (Integrative Zoology 2013; 8: 124 - 135)

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 or abundance	(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 or abundance	(Breiman *et al.* 1984)
Generalized Boosting Models (GBM)	1999	medium	0/14	p/a or abundance	(Friedman *et al.* 2000)
Random Forest	1995	high	26	p/a or abundance	(Breiman 2001)
Artificial neural networks (ANN)	1943	high	96/75	p/a or abundance	(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)

biomod2

```
library(biomod2)
# species occurrences
DataSpecies <- read.csv(system.file
                  ("external/species/mammals_table.csv",
                    package="biomod2"))

# the name of studied species
myRespName <- 'GuloGulo'

# the presence/absences data for our species
myResp <- as.numeric(DataSpecies[,myRespName])

# the XY coordinates of species data
myRespXY <- DataSpecies[,c("X_WGS84","Y_WGS84")]

# Environmental variables extracted from BIOCLIM (bio_3, bio_4, bio_7, bio_11 & bio_12)
myExpl = stack( system.file( "external/bioclim/current/bio3.grd",
                      package="biomod2"),
                system.file( "external/bioclim/current/bio4.grd",
                      package="biomod2"),
                system.file( "external/bioclim/current/bio7.grd",
                      package="biomod2"),
                system.file( "external/bioclim/current/bio11.grd",
                      package="biomod2"),
                system.file( "external/bioclim/current/bio12.grd",
                      package="biomod2"))

# 1. Formatting Data
myBiomodData <- BIOMOD_FormatingData(resp.var = myResp,
                                     expl.var = myExpl,
                                     resp.xy = myRespXY,
                                     resp.name = myRespName)

# 2. Defining Models Options using default options.
myBiomodOption <- BIOMOD_ModelingOptions()
```

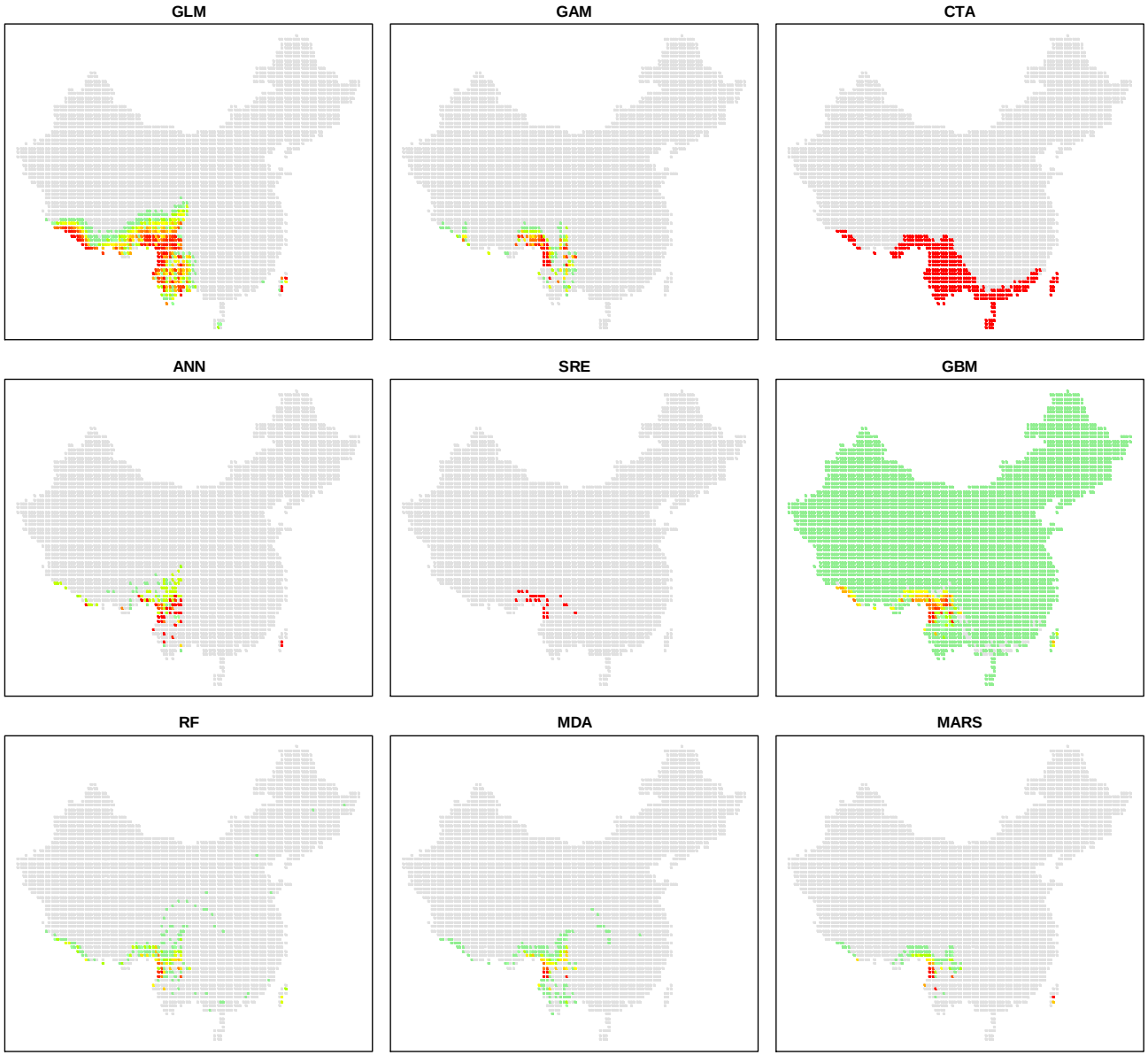
'GLM'
'GBM'
'GAM'
'CTA'
'ANN'
'SRE'
'FDA'
'MARS'
'RF'
'MAXENT.Phillips'
'MAXENT.Tsuruoka'

```
# 3. Doing Modelisation

myBiomodModelOut <- BIOMOD_Modeling( myBiomodData,
                                     models = c('SRE','RF'),
                                     models.options = myBiomodOption,
                                     NbRunEval=2,
                                     DataSplit=80,
                                     VarImport=0,
                                     models.eval.meth = c('TSS','ROC'),
                                     do.full.models=TRUE,
                                     modeling.id="test")

## print a summary of modeling stuff
myBiomodModelOut
```

Model comparison

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

Black snub-nose monkey



Li, X. H., H. D. Tian, R. Q. Li, Z. M. Song, F. C. Zhang, M. Xu, and D. M. Li. 2013. Vulnerability of 208 endemic or endangered species in China to the effects of climate change. Regional Environmental Change 13:843-852.

Model evaluation

- AUC (Area Under the Curve) in ROC (receiver operating characteristic) curve (Hanley & McNeil 1982)
- Cohen's Kappa (Landis & Koch 1977)

Hanley, J. A., and B. J. McNeil. 1982. The meaning and use of the area under a receiver operating characteristic (roc) curve. Radiology 143:29-36.

Landis, J. R., and G. G. Koch. 1977. Measurement of observer agreement for categorical data. Biometrics 33:159-174.

Model performance: Cohen's kappa

Cohen's kappa coefficient is a statistical measure of inter-rater agreement or *inter-annotator agreement* for qualitative (categorical) items. It is generally thought to be a more robust measure than simple percent agreement calculation since κ takes into account the agreement occurring by chance. Some researchers (e.g. Strijbos, Martens, Prins, & Jochems, 2006) have expressed concern over κ's tendency to take the observed categories' frequencies as givens, which can have the effect of underestimating agreement for a category that is also commonly used; for this reason, κ is considered an overly conservative measure of agreement.

$$\kappa = \frac{\Pr(a) - \Pr(e)}{1 - \Pr(e)}$$

where Pr(*a*) is the relative observed agreement among raters, and Pr(*e*) is the hypothetical probability of chance agreement, using the observed data to calculate the probabilities of each observer randomly saying each category.

If the raters are in complete agreement then κ = 1. If there is no agreement among the raters (other than what would be expected by chance) then κ ≤ 0.

```
library(psych)  # for for Cohen's Kappa
# "obs" has observed 0/1 values, "pred" has predicted 0/1 values
Kappa = cohen.kappa(data[, c("obs", "pred")])$kappa
```

Key measures of the performance of a binary classification

	Presence (predicted)	Absence (predicted)
Presence	A. 5 true positives (actual cats that were correctly classified as cats)	C. 3 false negatives (cats that were incorrectly marked as dogs)
Absence	B. 2 false positives (dogs that were incorrectly labeled as cats)	D. 17 true negatives (all the remaining animals, correctly classified as non-cats

Sensitivity (or Recall, R) = A/(A+C)
Specificity = D/(B+D)
Precision (P) = A/(A+B)
Overall prediction success (OPS) = (A+D)/n
Odds ratio = (AD)/(CB)
True skill statistics (TSS) = sensitivity + specificity - 1

The **prevalence (**$(a+c)/n$**)** of species occurrence is defined as the proportion of species occurrences among all the sites.

$$\text{Kappa} = \frac{(a+d) - [(a+c)(a+b)+(b+d)(c+d)]/n}{n - [(a+c)(a+b)+(b+d)(c+d)]/n}$$

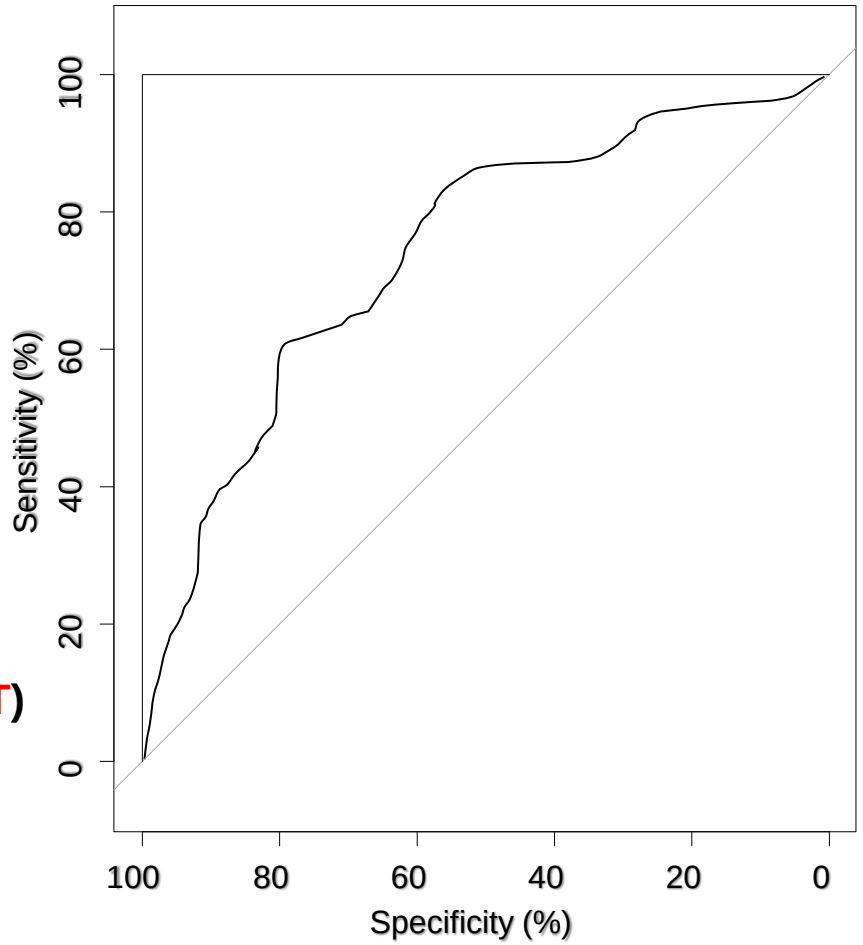
The sum of sensitivity and specificity can be maximized to give the threshold (Manel et al. 2001), which is equivalent to finding a point on the ROC (receiver operating characteristics) curve (i.e. sensitivity against 1-specificity) whose tangent slope is equal to 1 (Cantor et al. 1999).

ROC curve

The receiver operating characteristic (ROC) is a graphical plot which illustrates the performance of a binary classifier system as its discrimination threshold is varied.

It is created by plotting the fraction of true positives out of the total actual positives (TPR = true positive rate) vs. the fraction of false positives out of the total actual negatives (FPR = false positive rate), at various threshold settings.

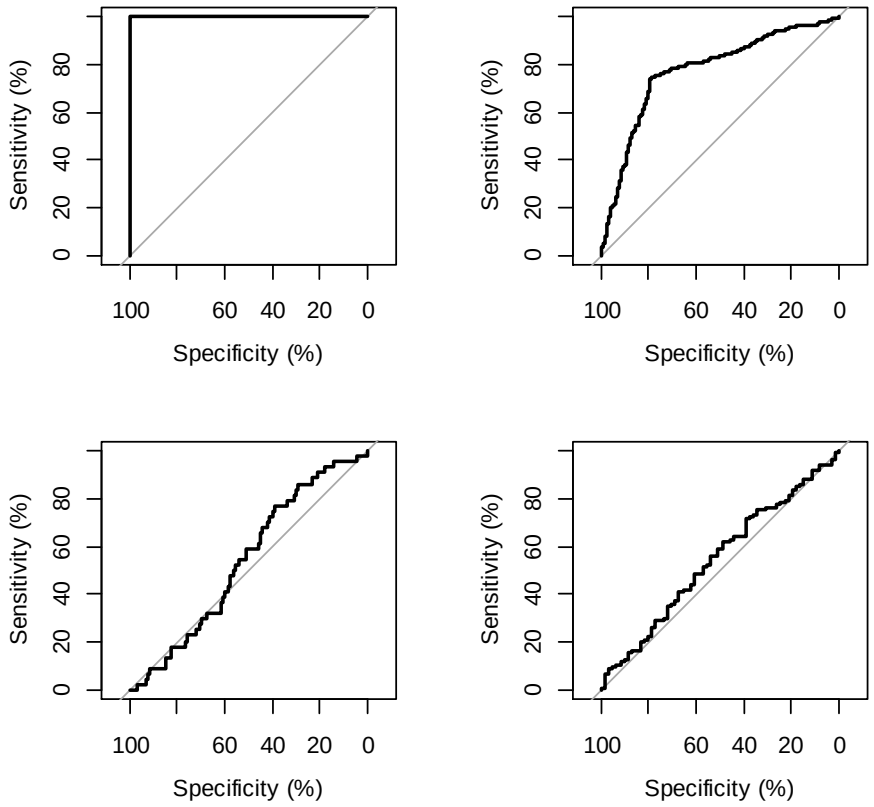
Sensitivity = True positive rate
1- Specificity = False positive rate




```
library(pROC)  # for ROC
# use has the 0/1 values, prb has the predicted p values
roc1 = roc(use, prb , percent = T, auc = T, plot = T)
roc1$auc # the AUC value

# specificity, sensitivity, and threshold
SST  = coords(roc1, 'best',
         ret = c('spec', 'sens', 'threshold'))
```

ROC curves

```
library(pROC)  # for ROC
par(mfrow=c(2,2))

# perfect prediction
prb = runif(1000)
use = round(prb)
roc1 = roc(use, prb , percent = T, auc = T, plot = T)

# half perfect, half random
use1 = round(prb)[1:500]
use2 = sample(c(0,1), 500, rep=T)
use = c(use1, use2)
roc1 = roc(use, prb , percent = T, auc = T, plot = T)

# random prediction
use = sample(c(0,1), 1000, rep=T)
roc1 = roc(use, prb , percent = T, auc = T, plot = T)

# high presence
use = sample(c(0,1), 1000, rep=T, prob=c(0.05, 0.95))
roc1 = roc(use, prb , percent = T, auc = T, plot = T)
```



AUC: a misleading measure of the performance of predictive distribution models (Lobo et al. 2008)

(1) It ignores the predicted probability values and the goodness-of-fit of the model;

(2) It summarises the test performance over regions of the ROC space in which one would rarely operate;

(3) It weights omission and commission errors equally;

(4) It does not give information about the spatial distribution of model errors; and, most importantly,

(5) The total extent to which models are carried out highly influences the rate of well-predicted absences and the AUC scores.

Instead of using only the AUC, we propose that sensitivity and specificity should be also reported, so that the relative importance of commission and omission errors can be considered to assess the method performance. Unfortunately, we cannot recommend any useful method to compare model performance among species.

Jorge M. Lobo, Alberto Jiménez-Valverde, and Raimundo Real. 2008. AUC: a misleading measure of the performance of predictive distribution models. *Global Ecology and Biogeography,* **17:** 145–151

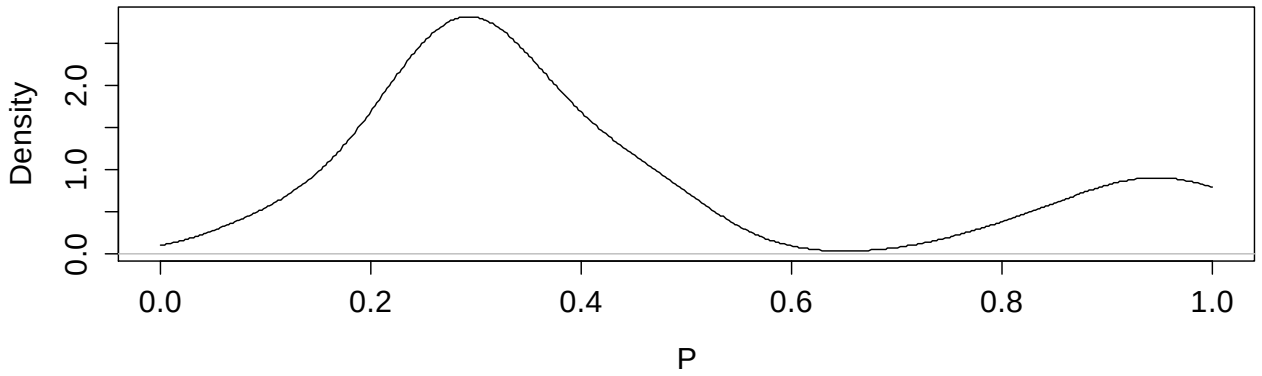
Prediction

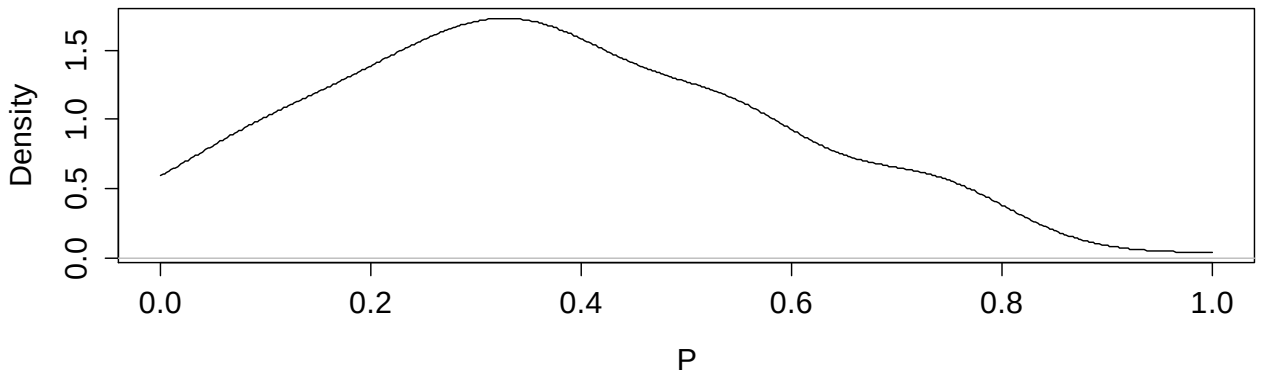
```
# predicted probability of presence
# random forest
p <- predict(model, new.data, type = 'response',
                    predict.all = F)
# GAM
p <- predict.gam(model, type="link", newdata = new.data)


# some random data
ID = 1:200
use = sample(c(0,1), 200, replace=T, prob=c(0.75,0.25))
p = use / 4 + rnorm(200, mean=0.3, sd = 0.2)
# p = use / 1.5 + rnorm(200, mean=0.3, sd = 0.1)
occ = data.frame(ID, use, p)
```

ID	use	p	Pred
1	1	0.710630789	1
2	0	0.606700985	1
3	1	0.693370703	1
4	0	0.168996797	0
5	1	0.581859344	1
6	0	0.243226211	0
7	0	0.688768088	1
8	0	0.192407567	0
9	1	0.640645997	1
10	0	0.122985393	0
11	0	0.145059298	0
12	0	0.128858521	0
13	1	0.736497782	1
14	1	0.370274550	0
15	0	0.278402645	0

Distribution of the probability of presence

```
plot(density(p, from=0, to =1), xlab="P", main="")
```





Model evaluation indices

```
# some random data
ID = 1:200
use = sample(c(0,1), 200, replace=T, prob=c(0.75,0.25)); p = use / 3 + rnorm(200, mean=0.3, sd = 0.2)
occ = data.frame(ID, use, p)

modEva <- function(data) {
  library(pROC)  # for ROC, specificity, sensitivity, threshold
  library(psych) # for Cohen's Kappa
  use = data[,"use"];  prb = data[,"p"]
  roc1 = roc(use, prb , percent = T, auc = T, plot = T) # for AUC
  SST  = coords(roc1, 'best', ret = c('spec', 'sens', 'threshold'), transpose = TRUE) # for specificity, sensitivity, threshold
  ACC = (SST[1]*sum(use) + SST[2]*(nrow(data) - sum(use))) / nrow(data) # for accuracy
  AUC = roc1$auc  # AUC
  Pred <- ifelse(prb >= SST[3], 1, 0) # classification based on best threshold
  data  = cbind(data, Pred)
  kappa = cohen.kappa(data[, c("use", "p")])$weighted.kappa
  L = numeric(6);  L[1] = AUC;  L[2] = as.numeric(ACC);  L[3] = kappa
  L[4] = as.numeric(SST[1]);  L[5] = as.numeric(SST[2]);  L[6] = as.numeric(SST[3]);  L = round(L, 3)
  names(L) = c("AUC", "Accuracy(%)", "Kappa", "Specificity", "Sensitivity", "Threshold")
  return(L)
}
modEva(occ)

TSS  = sensitivity + specificity – 1 # true skill statistic
```

AUC	Accuracy(%)	Kappa	Specificity	Sensitivity	Threshold
84.091	70.794	0.604	93.464	63.83	0.562

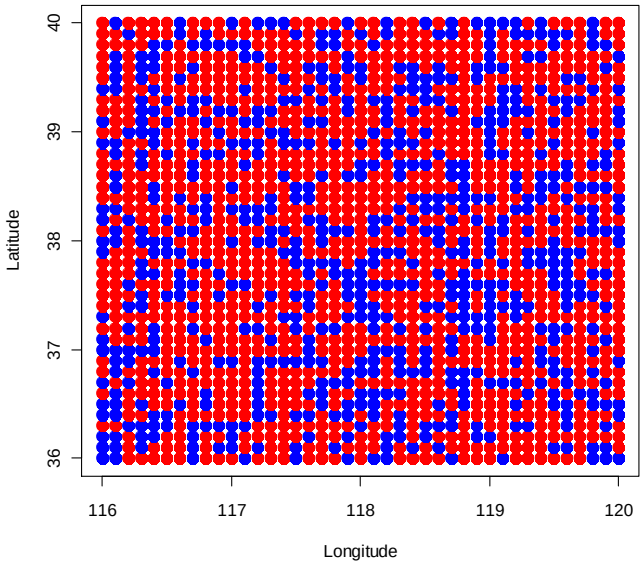
Case study: effect of climate change to species distribution

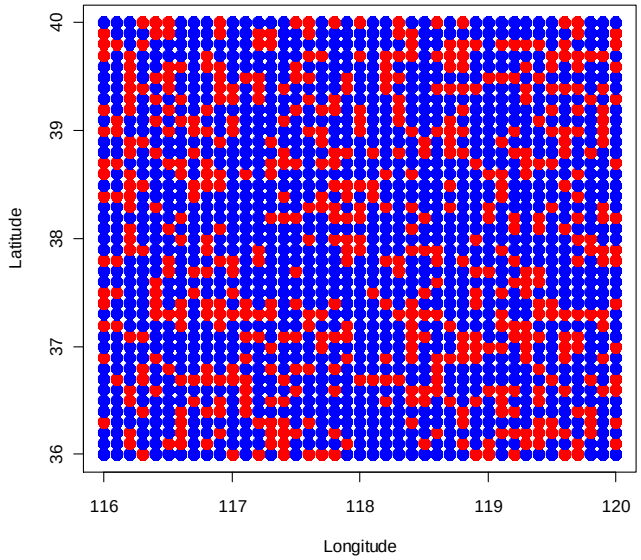
R code for logistic regression
– an example for species habitat prediction in future climate conditions

```
#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 <- as.factor(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 <- glm(use ~ elev + aspect + temperature, data = distr, family = binomial()); summary(fit)
# fit <- step(glm(use ~ elev + aspect + temperature, data = distr, family = binomial())); summary(fit)
p.current <- predict(fit, newdata = distr, type='response') # predicted p values hist(p.current)

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]>0.5) points(Lon[i],  Lat[i],  col = "red",    cex = 1.5,  pch = 19)
if(p.current[i]<0.5) points(Lon[i],  Lat[i],  col = "blue",  cex = 1.5,  pch = 19)}
```

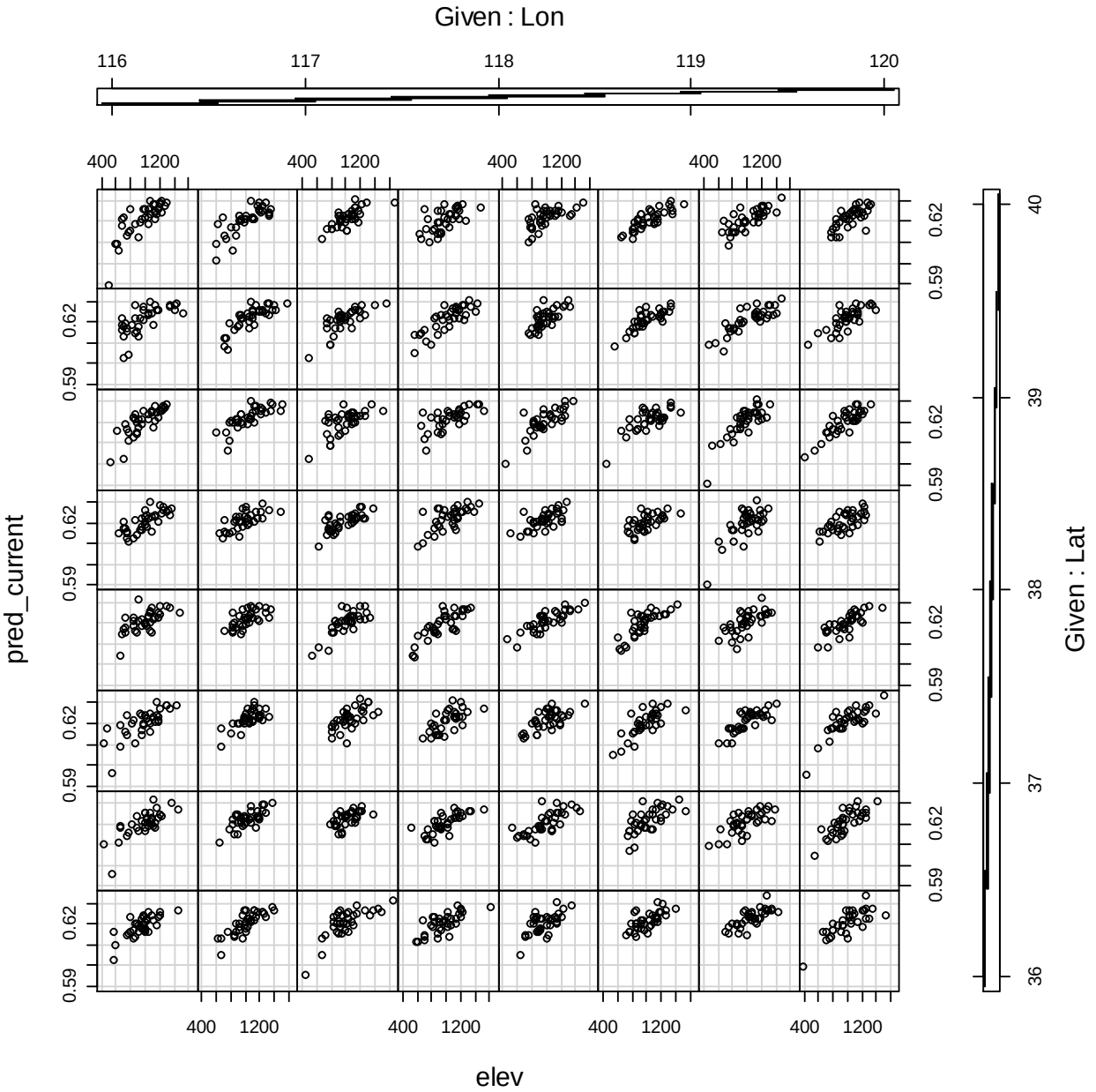
R code for logistic regression – an example for habitat prediction in future

```
#predict future distribution
distr.future <- distr
distr.future$temperature <- distr.future$temperature + 2 # a global warming scenario
p.future <- predict(fit , type = c("response"), newdata = distr.future);

#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)}
```



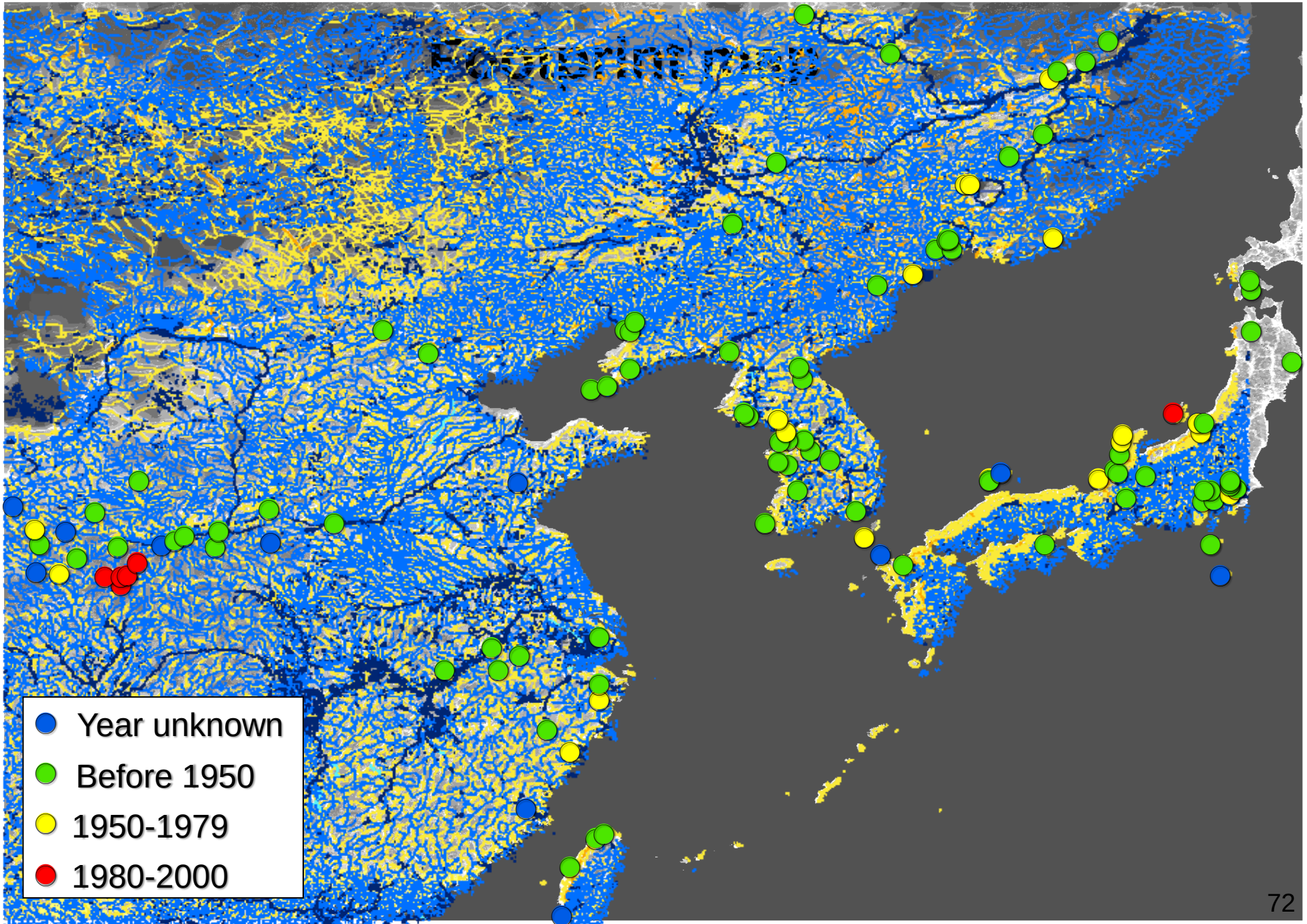


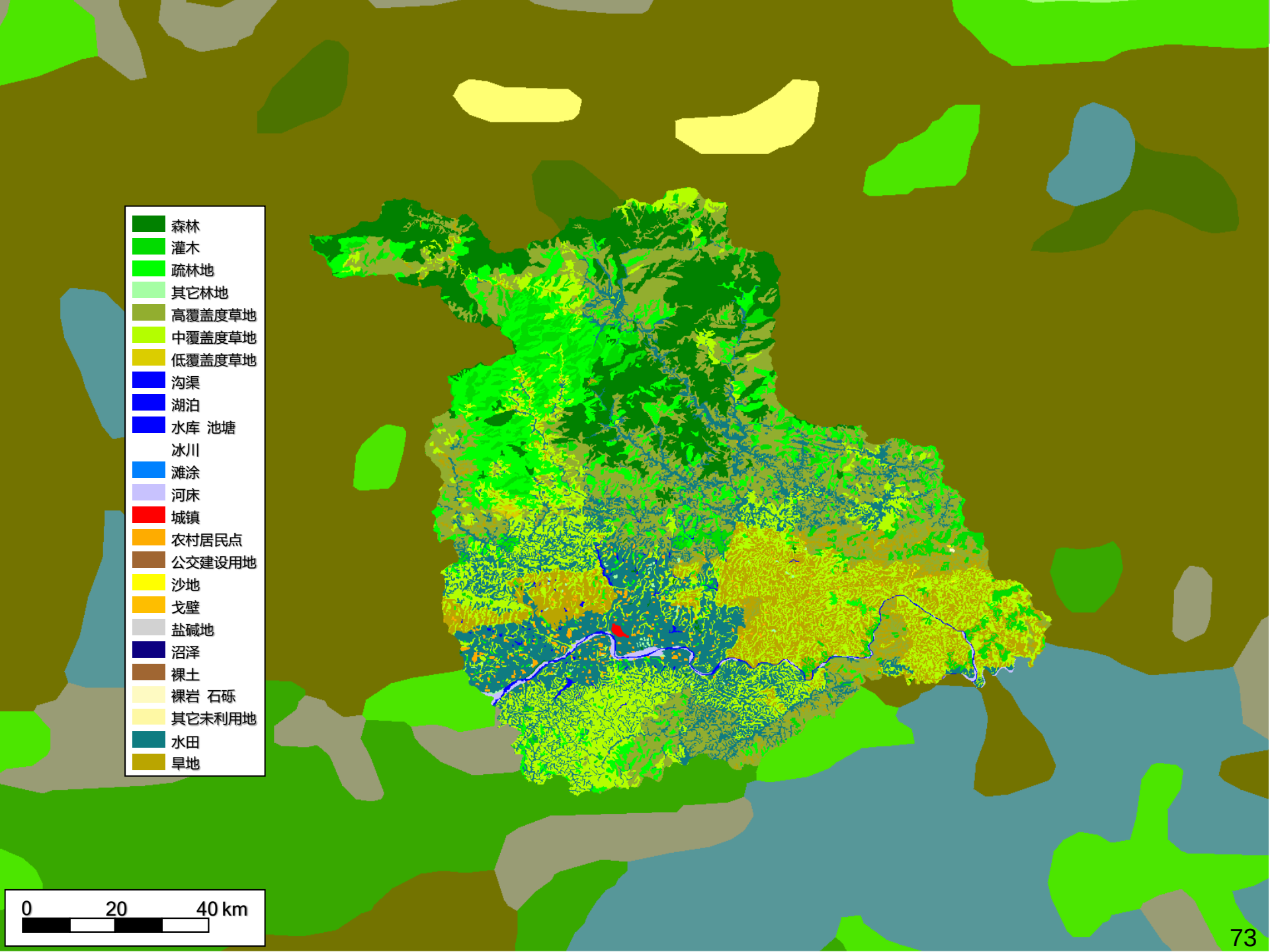
coplot(pred.current ~ elev | Lon*Lat, data=distr, overlap = 0, number = c(8,8))

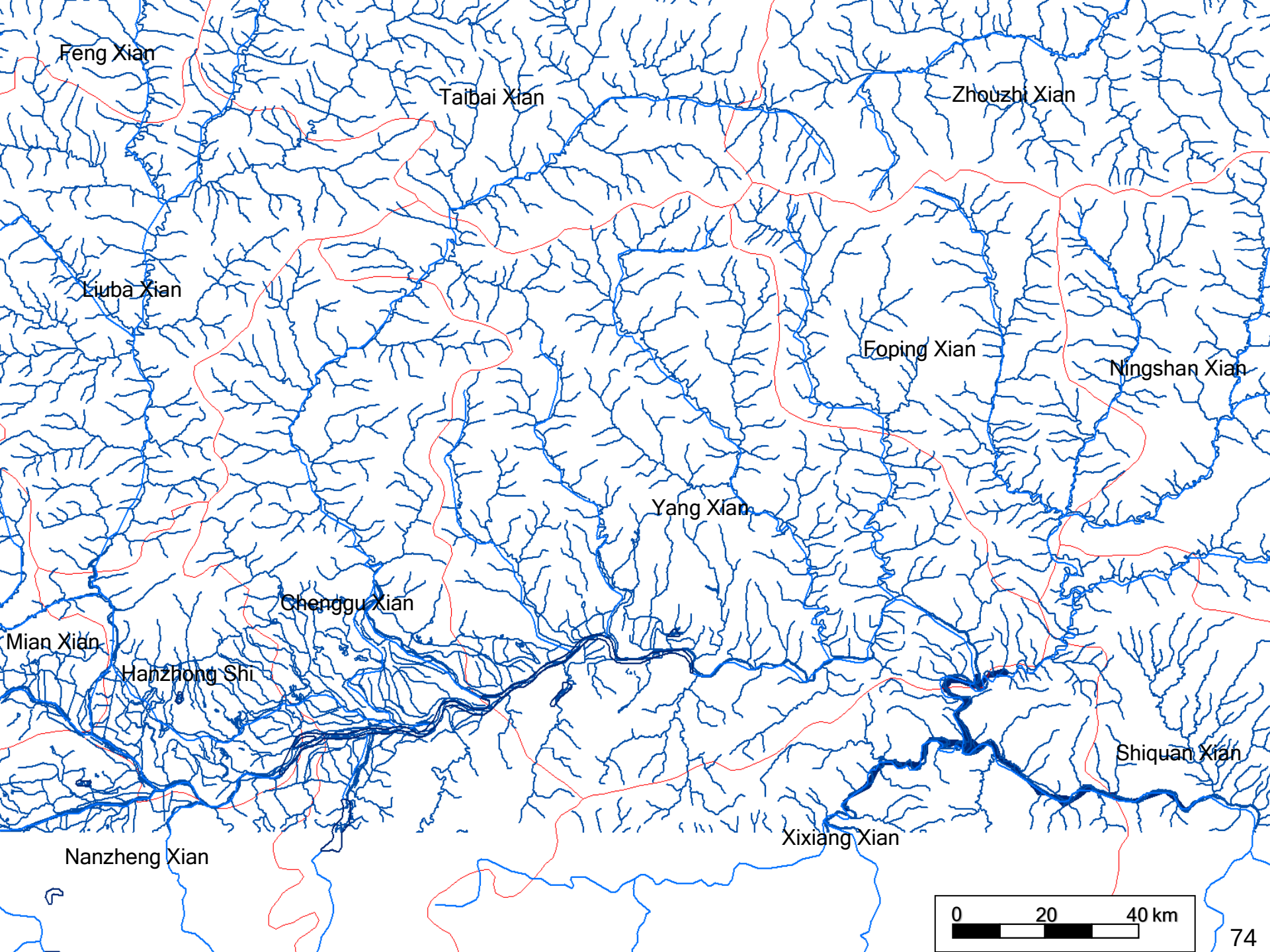


Scale dependent logistic regression: a case study of habitat use by the crested ibis

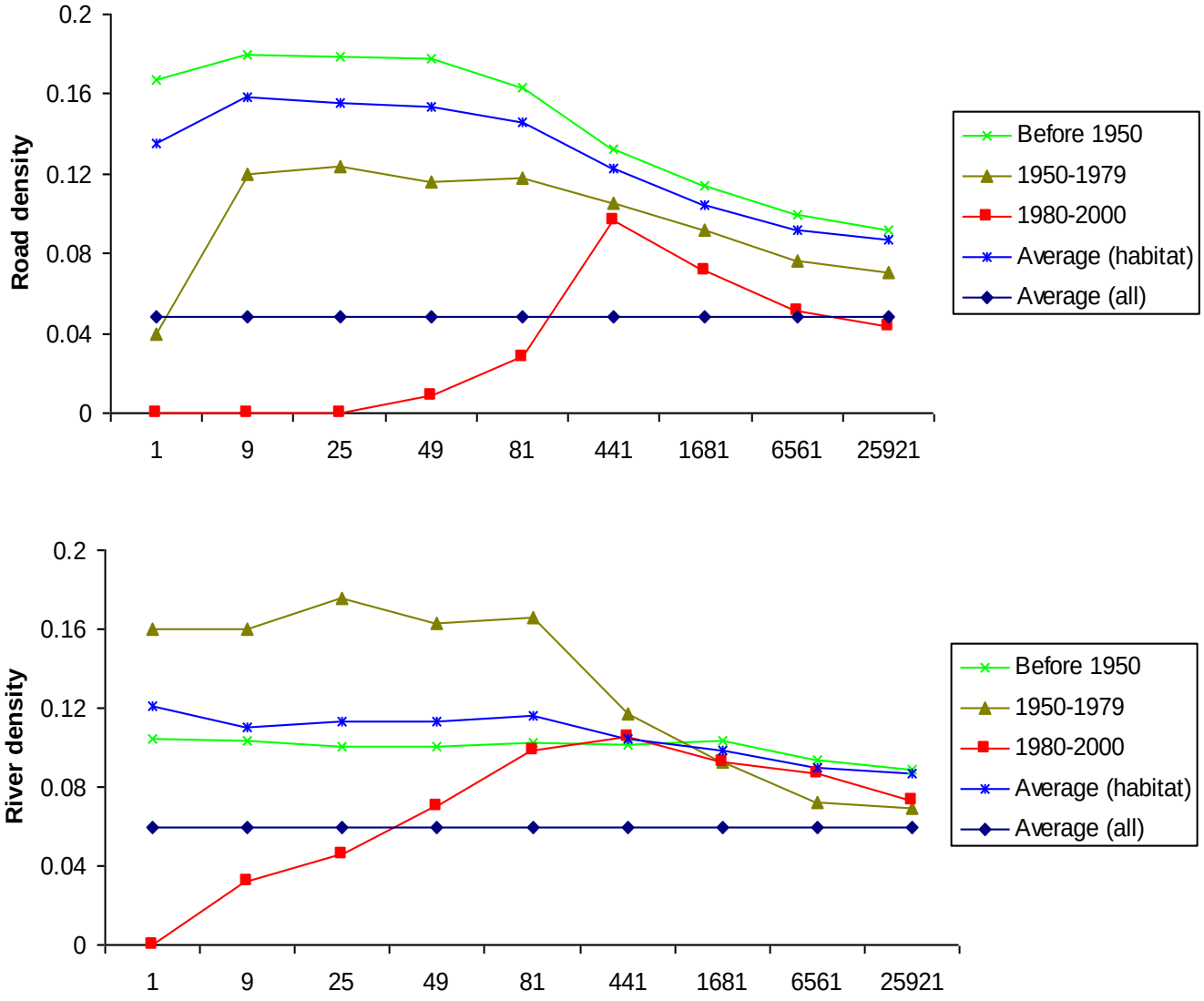

AMUR
KHABAROVSK
RUSSIA
MONGOLIA
HEILONGJIANG
PRIMORYE
JILIN
LIAONING
CHINA
NORTH KOREA
SOUTH KOREA
HEBEI
SHANXI
SHAANXI
GANSU
SHANDONG
HENAN
ANHUI
SICHUAN
SHANGHAI
ZHEJIANG
FUJIAN
TAIWAN
JAPAN
PACIFIC OCEAN


Year unknown
Before 1950
1950-1979
1980-2000

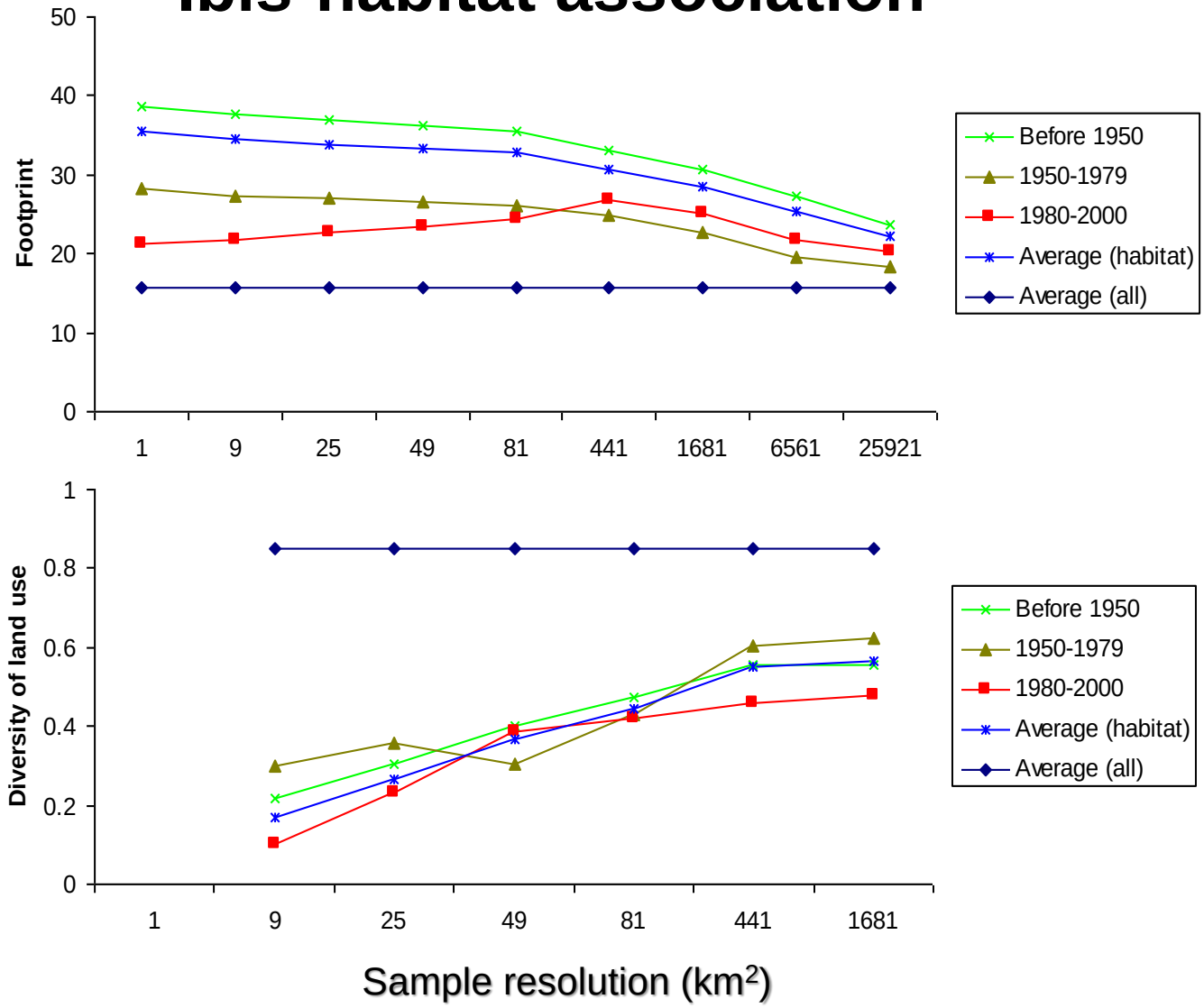

森林
灌木
疏林地
其它林地
高覆盖度草地
中覆盖度草地
低覆盖度草地
沟渠
湖泊
水库 池塘
冰川
滩涂
河床
城镇
农村居民点
公交建设用地
沙地
戈壁
盐碱地
沼泽
裸土
裸岩 石砾
其它未利用地
水田
旱地
0
20
40 km


Feng Xian
Taibai Xian
Zhouzhi Xian
Liuba Xian
Foping Xian
Ningshan Xian
Yang Xian
Chenggu Xian
Mian Xian
Hanzhong Shi
Shiquan Xian
Xixiang Xian
Nanzheng Xian
0
20
40 km

Ibis-habitat association



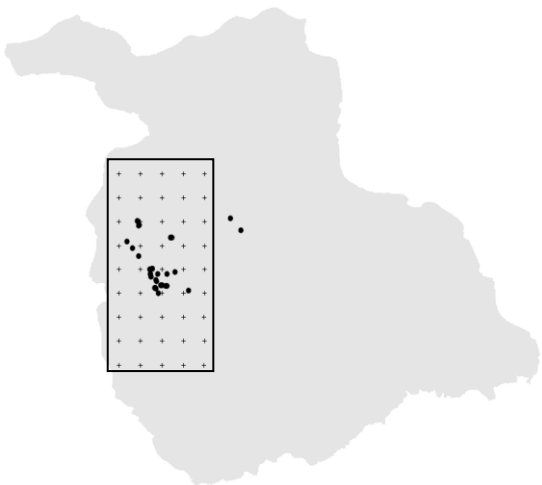
Ibis-habitat association



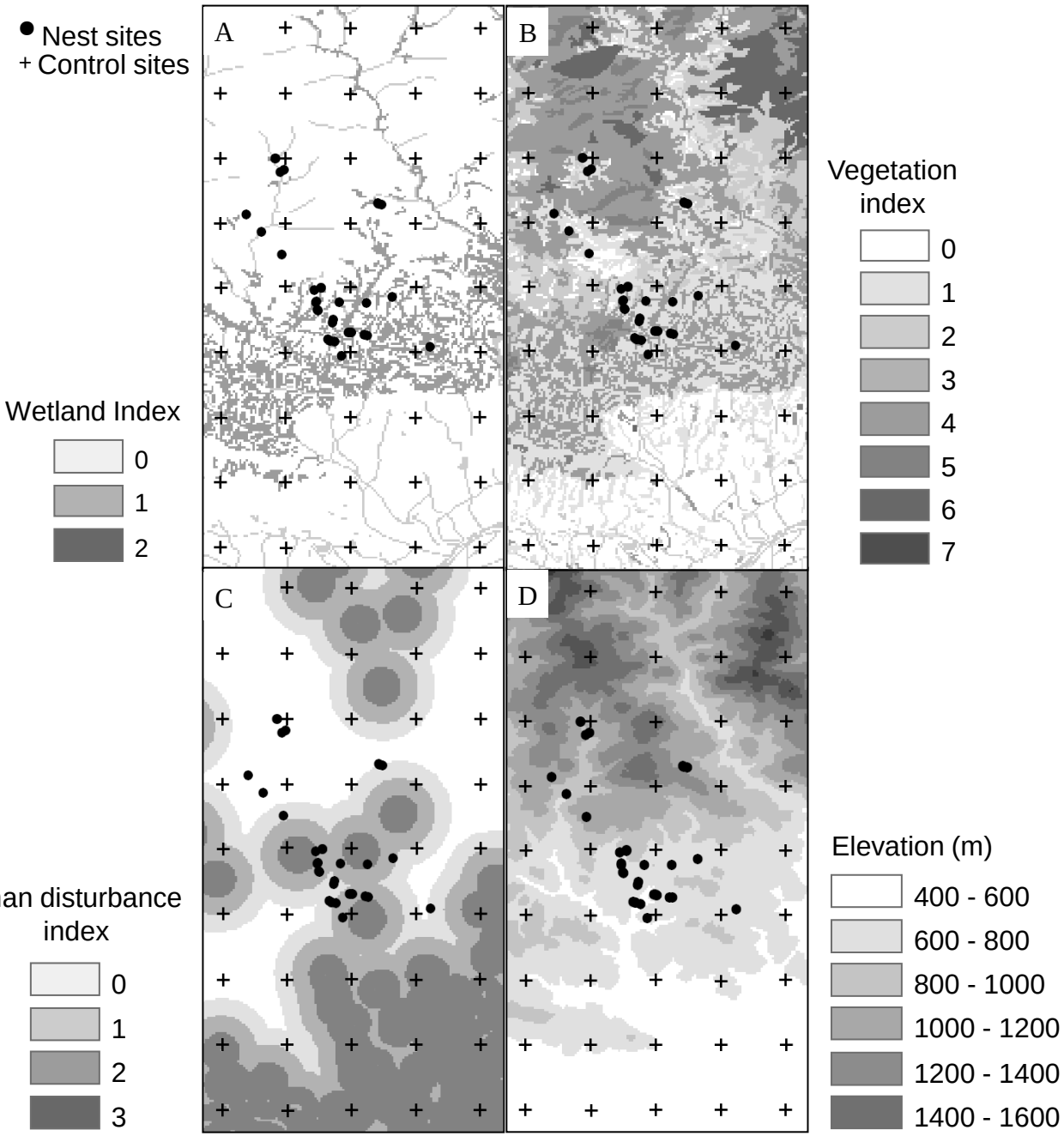
Li, X. H., H. D. Tian, and D. M. Li.2009. Why the crested ibis declined in the middle twentieth century? Biodiversity and Conservation 18: 2165-2172

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

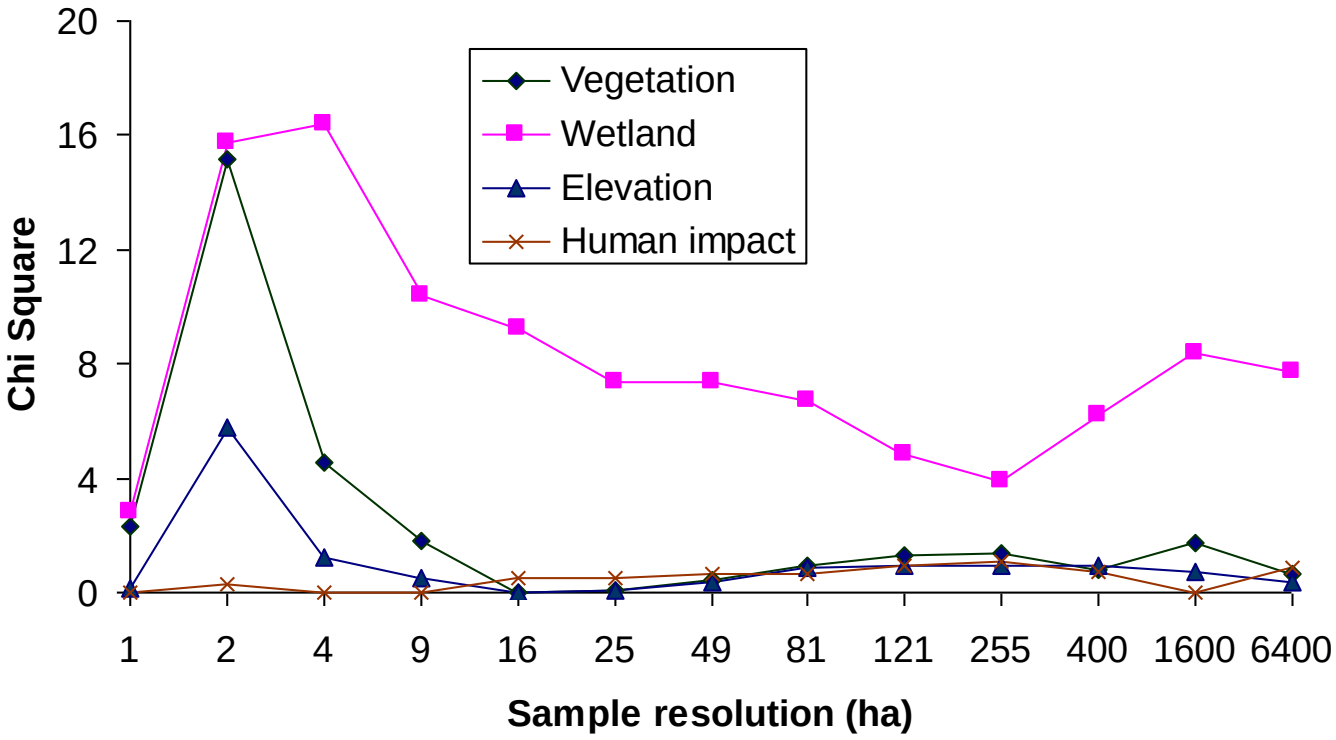
A. Wetland index;
B. Vegetation index;
C. Human impact index;
D. Elevation.



Study area



The contribution (Wald chi square) of each of the four habitat variables to the logistic regression models for nest site selection of crested ibis when habitat variables were sampled at 13 grain sizes.



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.

Assignment

General objectives: learn about Poisson regression.

- Develop a dataset to perform:
  - Poisson regression Y - X1, X2, X3, etc.
- Describe the dataset, check the correlation between each two independent variables, check the over dispersion, describe the significance of each independent variables and overall model fit.