Topic 8: Generalized Linear Models

Background

- Generalized Linear Models (GLMs) are an extension of linear models that allow for:
  - a nonlinear link function for nonlinear y's
  - response probability distributions can be any member of the exponential family of distributions (e.g., normal, inverse normal, gamma, binomial, negative binomial, Poisson and multinomial).
  - Unequal variances of the *y*'s (variance of the *y*'s is a function of the mean of *y*, given *x*)
  - Errors are uncorrelated for ease of calculation of the likelihood
  - NOTE: If errors are correlated, Generalized Estimating Equations (GEEs) can be used

Background (after Kery 2010*)

Formally, a GLM is described by the following three components:

1. Statistical distribution used to describe the random variation in the response; e.g. $y_i \sim Normal\ (\mu_i, \sigma^2)$

2. A so-called link function g, that is applied to the expectation of the response E(y); e.g. $\mu = E(y_i)$

3. A linear predictor, which is a linear combination of covariate effects that are thought to make up g(E(y)); $\alpha + \beta * x_i$

*Kery (2010) Introduction to WinBUGS for Ecologists. Academic Press.

Commonly Used GLM's

1. OLS model: y is continuous, LINK=identity, DIST=normal

2. Logistic Regression: y is a proportion (or a 0,1 Bernoulli variable), LINK=logit, DIST=binomial

3. Poisson Regression, log linear model: y is a count (no natural denominator, else use y as a proportion), LINK=log, DIST=Poisson

4. Count using Negative Binomial: y is a count (no natural denominator, else use y as a proportion), LINK=log, DIST=negbin

5. Gamma Model with log linear model: y is a positive continuous variable, LINK=log, DIST=gamma.

Under/Overdispersion

If default variance for specified distribution does not match the data:

- data are *over*- or *underdispersed.*
- Can happen with Poisson, binomial, negbin
- Overdispersion factor can be added to the variance function and an estimate of this found by MLE along with the other parameters.
- Alternatively: another distribution may be more appropriate (e.g., switch to negative binomial for count data).

Model Goodness of Fit

- Asymptotic z-test for individual coefficients.

- Likelihood ratio test for nested models. Better models have higher likelihood (or log likelihood), which is the same as saying the -2 lnL is smaller.

- Pseudo R squared value, based on lnL of the model versus lnL for a "null model" with only the intercept (no explanatory variables), to obtain a similar interpretation to R^2 for linear models.

Model Goodness of Fit

- Use Deviance or Pearson's Chi Squared Statistic, for grouped or ungrouped data to compare unrestricted to restricted models, called, "Deviance partitioning", using a likelihood ratio test.

- Aikaike's Information Criterion (AIC). Smaller is better; gives a "penalty" for number of variables.

$$AIC = -2 * \ln(L) + 2 * k$$

- Schwarz Criterion (SC). Similar to AIC, but includes the number of response levels, the no. of explanatory variables, and the sample size.

$$SC = -2 * \ln(L) + k * \ln(n)$$

Logistic Regression + Predictive Habitat Map Example

- For many problems, the y variable is a Bernouilli random variable (0/1):
  - Detection versus non detection of an animal
  - Healthy versus diseased animal
  - Dead versus live trees

- Can be summarized into a Binomial Distribution which gives the proportion of successes (p) versus failure (q=1-p).

- We then wish to use a model to predict this y variable, given a set of explanatory variables, x, which can be continuous variables, class variables represented by a set of dummy variables, and interactions between continuous and class variables.

- The predicted values from this model will be the probability that y=1.

Logistic Regression (cont.)

$$y_i = \frac{p_i}{1-p_i}, \quad \text{called the "odds ratio"}$$

$$p_i = \text{proportion of successes out of a group of observations}$$

$$\text{as } p_i \text{ goes to 1, the odds ratio goes to } \infty$$

$$\text{as } p_i \text{ goes to 0, the odds ratio goes to 0}$$

Using the logarithm of the odds ratio as a linear function of the *x* variables:

$$\ln\left(\frac{p_i}{1-p_i}\right) = \mathbf{x_i}'\boldsymbol{\beta}, \text{ then } \left(\frac{p_i}{1-p_i}\right) = \exp(\mathbf{x_i}'\boldsymbol{\beta})$$

$$\text{Pr}\hat{\text{o}}\text{b}(Y=1) = p_i = \frac{\exp(\mathbf{x_i}'\boldsymbol{\beta})}{(1+\exp(\mathbf{x_i}'\boldsymbol{\beta}))}$$

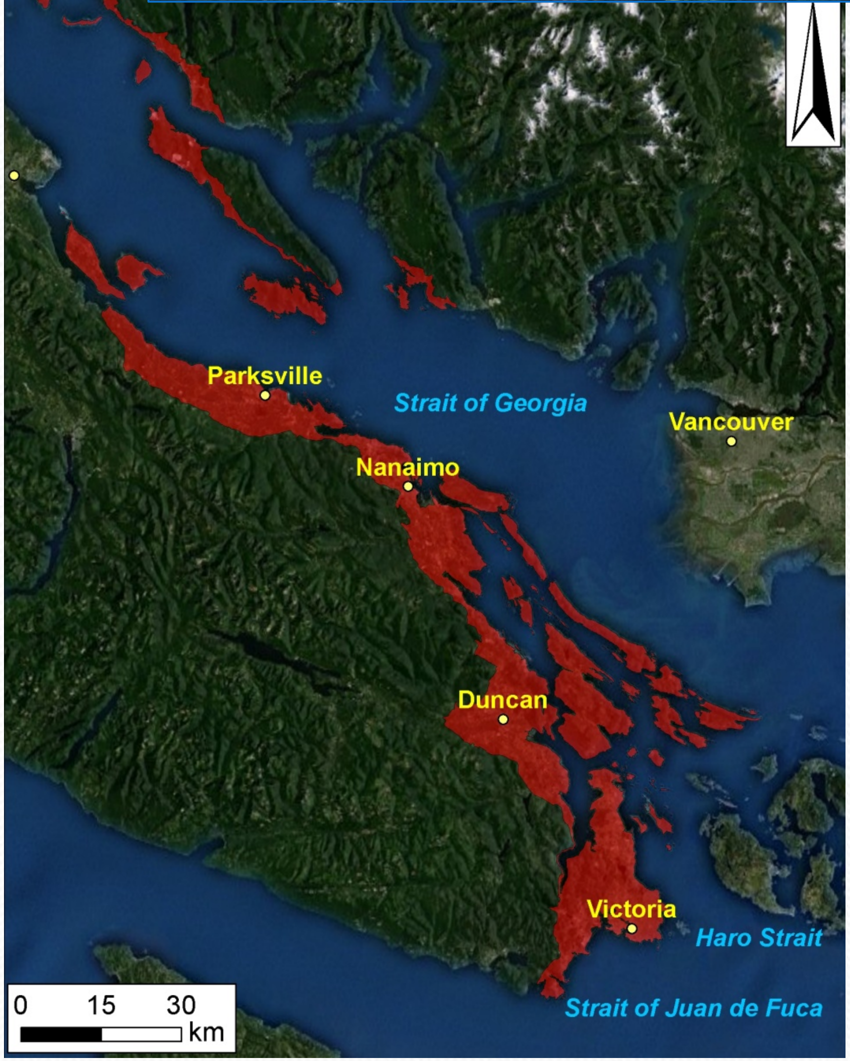
This is called a “logistic function”. It is constrained to give predictions between 0 and 1, which are probabilities that *y*=1.

Goodness of Fit (additions)

- Number of Concordant/Discordant/Ties.
- Classification Table: shows the results for different probability “cutoff” values.
- Receiver Operating Characteristic Curves (ROC) ; Area Under the Curve (AUC)

Predictive Habitat Map Example

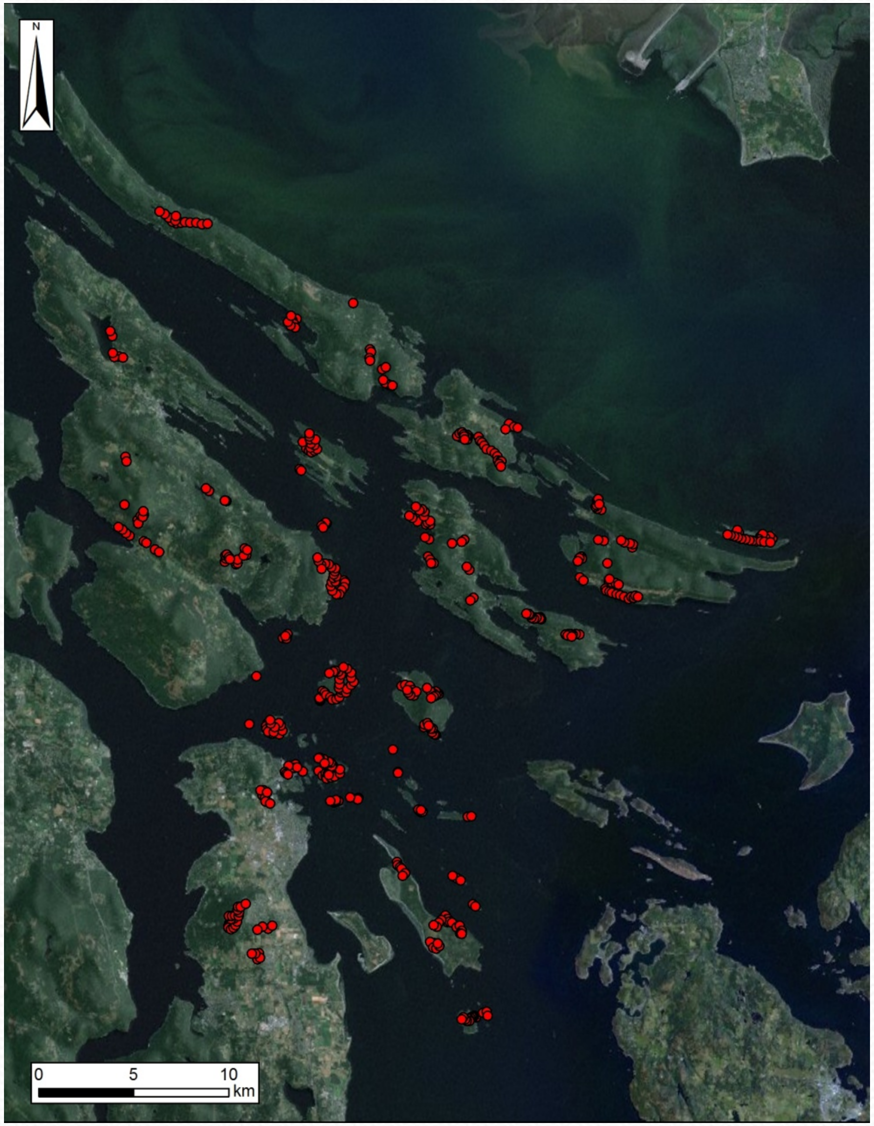
- **Coastal Douglas Fir Zone: Most Imperilled Ecosystem in BC**
- >60% Converted to Human Use
- ~84% Private Land
- <1% Old-Growth Forest
- 115 'Species At Risk'
- Study area: ~2120 km^2



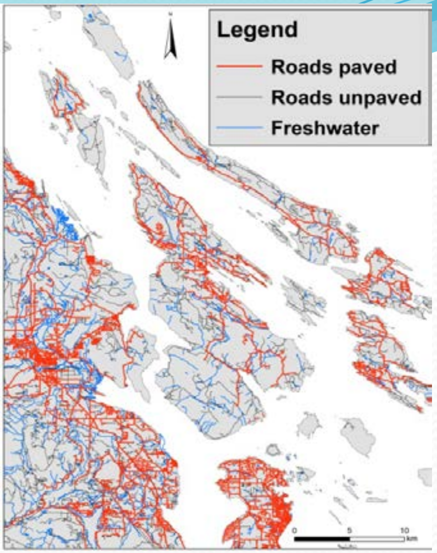
Austin et al. (2008); Madrone Env. Services (2008)

Data

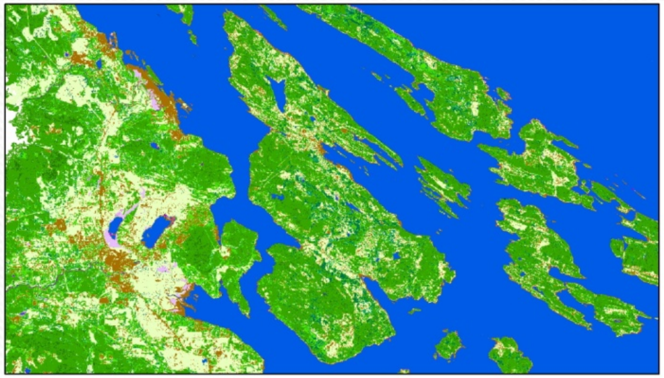
- presence/absence data
- ~460 plots
- Brown Creeper as an example



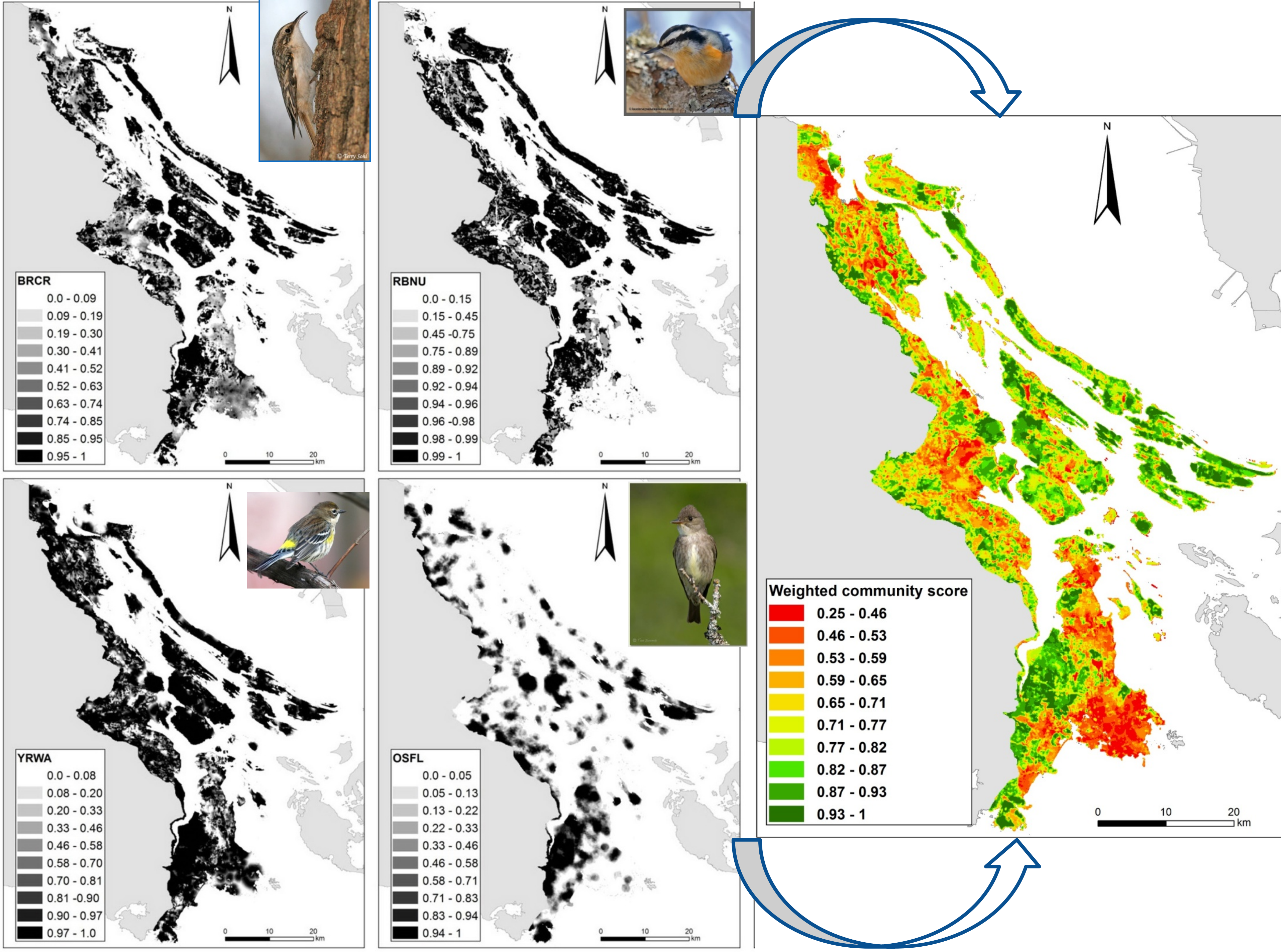
Model covariates



- Coarse (1km) and fine scale (100m)[1]
- 25 landscape covariates
  - e.g. road length; urban; rural; young and older forest; nearest water
- Using: Hawth‘s Tools, ArcGIS, Geospatial Modelling Environment, R



[1] Guisan and Thuiller, 2005; Jewell et al., 2007


N
BRCR
0.0 - 0.09
0.09 - 0.19
0.19 - 0.30
0.30 - 0.41
0.41 - 0.52
0.52 - 0.63
0.63 - 0.74
0.74 - 0.85
0.85 - 0.95
0.95 - 1
0 10 20 km
© Terry Sohl
N
RBNU
0.0 - 0.15
0.15 - 0.45
0.45 -0.75
0.75 - 0.89
0.89 - 0.92
0.92 - 0.94
0.94 - 0.96
0.96 -0.98
0.98 - 0.99
0.99 - 1
0 10 20 km
N
YRWA
0.0 - 0.08
0.08 - 0.20
0.20 - 0.33
0.33 - 0.46
0.46 - 0.58
0.58 - 0.70
0.70 - 0.81
0.81 -0.90
0.90 - 0.97
0.97 - 1.0
0 10 20 km
N
OSFL
0.0 - 0.05
0.05 - 0.13
0.13 - 0.22
0.22 - 0.33
0.33 - 0.46
0.46 - 0.58
0.58 - 0.71
0.71 - 0.83
0.83 - 0.94
0.94 - 1
0 10 20 km
N
Weighted community score
0.25 - 0.46
0.46 - 0.53
0.53 - 0.59
0.59 - 0.65
0.65 - 0.71
0.71 - 0.77
0.77 - 0.82
0.82 - 0.87
0.87 - 0.93
0.93 - 1
0 10 20 km

Example in R

- Package: MASS
- Function: *glm()*

- *brcr_m1 <- glm(BRCR ~ CR_CL_2Z, family=binomial, data=brddata)*

- For help on this (or any other) function use:

  help(glm) *or* ?glm