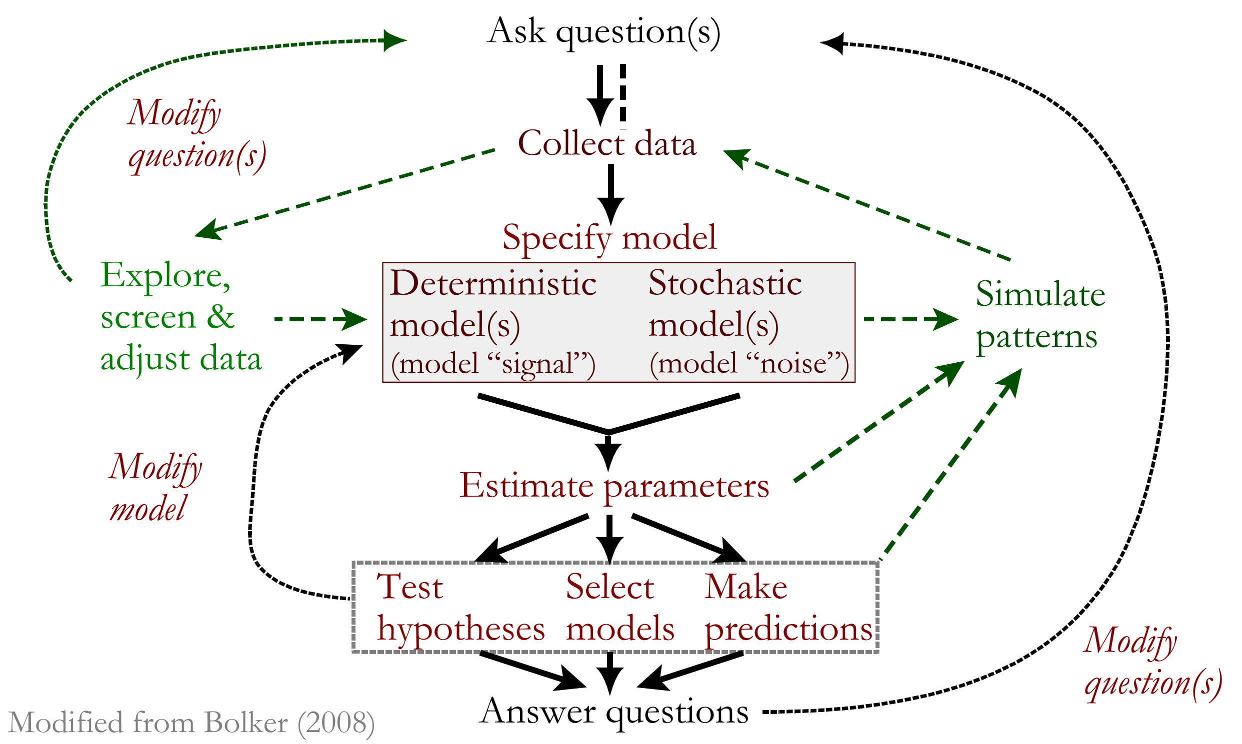
Role of Statistics... *the modeling process*



Landscape of Statistical Methods...

The basic statistical model:

Y = deterministic part + stochastic part

- Univariate
- Multivariate

- Linear
- Nonlinear
- Smoothed

- Distribution
- Heterogeneity
- Autocorrelation
- Multiple levels
- Random noise

Landscape of Statistical Methods...

The basic statistical model:

Y = deterministic part + stochastic part

- Univariate
- Multivariate

- Linear
- Nonlinear
- Smoothed

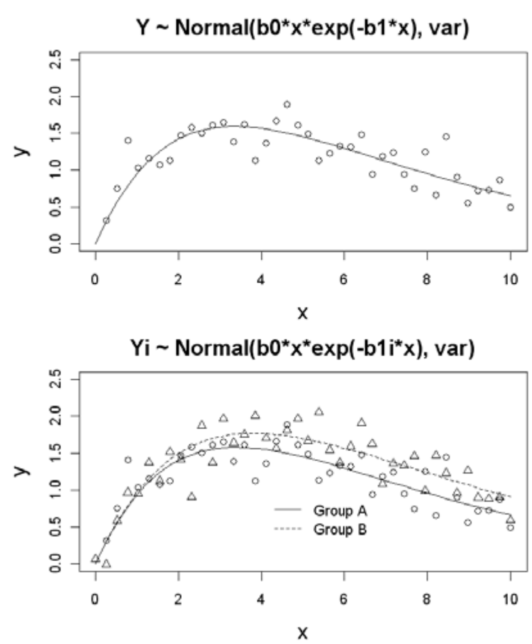
- Distribution
- Heterogeneity
- Autocorrelation
- Multiple levels
- Random noise

Landscape of Statistical Methods...

Dealing with nonlinearity

Nonlinear least squares models (NLS)

- Relax the requirement of linearity (in the parameters) but keep the requirements of independent (and normal and constant) errors (to compute Likelihoods)
- Method: *numerical least squares*



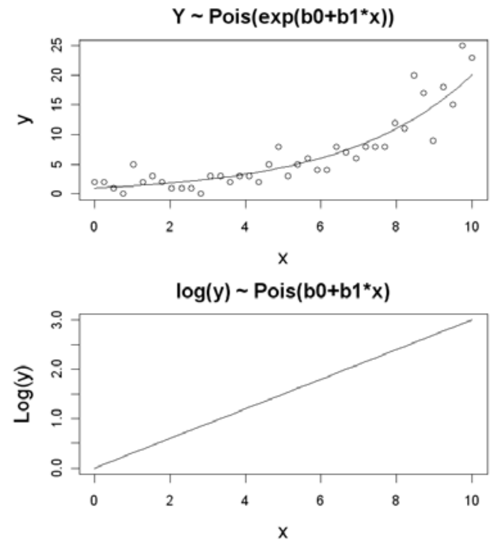
Landscape of Statistical Methods...

Dealing with nonlinearity

Generalized linear models (GLMs)

Link
$$y = e^x$$
$$x = \log(y)$$

- Models that have a particular kind of *nonlinearity* and particular kinds of *nonnormally* distributed (but still independent and constant) errors
- Can fit any nonlinear relationship that has a *linearizing transformation* (link function)
- Method: *iteratively reweighted least squares* (avoids distortions in expected variance that linearizing transformation would induce)



Landscape of Statistical Methods...

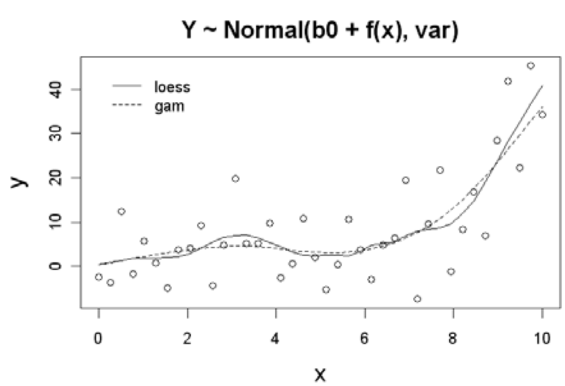
Dealing with nonlinearity

Generalized additive models (GAMs)

- Fit a *smoothing curve* through the data but keep the requirements of independent and constant *parametric* errors; hybrid parametric-nonparametric model
- Note, purely data-driven; phenomenological
- Many different types of smoothers

$$Y \sim Normal\left(b_0 + \boxed{b_1 x}, \sigma^2\right)$$

$$Y \sim Normal\left(b_0 + \boxed{f(x)}, \sigma^2\right)$$



Landscape of Statistical Methods...

The basic statistical model:

Y = deterministic part + stochastic part

- Univariate
- Multivariate

- Linear
- Nonlinear
- Smoothed

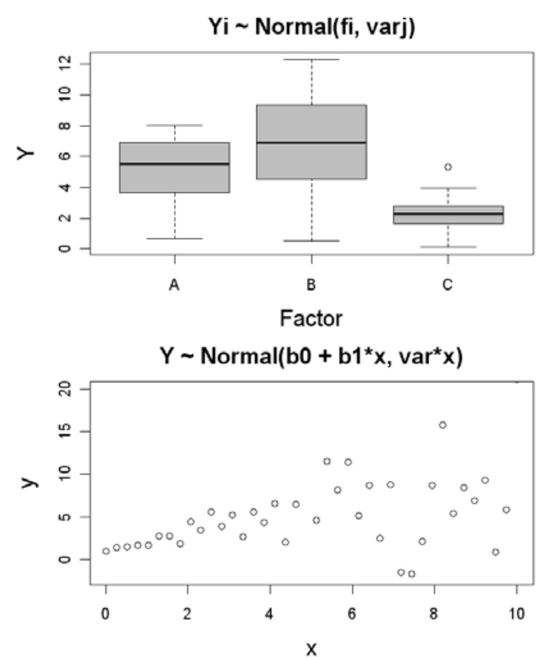
- Distribution
- Heterogeneity
- Autocorrelation
- Multiple levels
- Random noise

Landscape of Statistical Methods...

Dealing with heterogeneity

Generalized (non)linear least squares (GLS/GNLS)

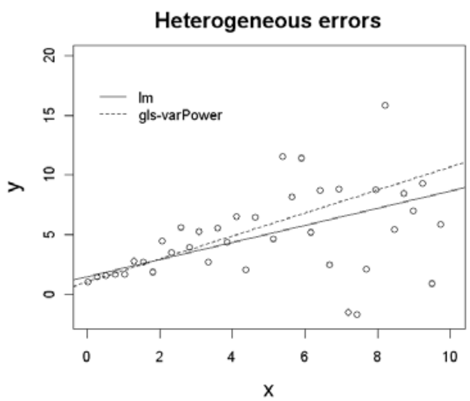
- *Y* is *continuous*
- All observed values are *independent* and *normally* distributed, but with a *nonconstant variance* (heterogeneity)
- Method: *(restricted) maximum likelihood (REML/ML)*

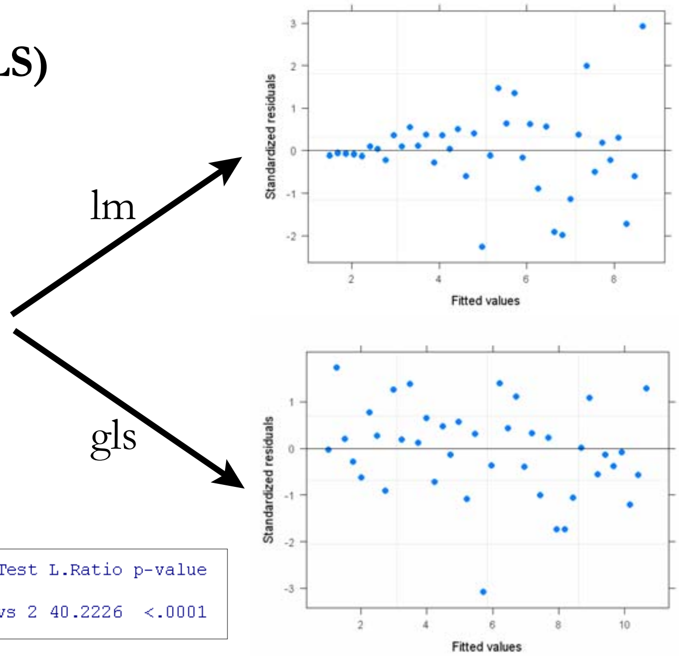


Landscape of Statistical Methods...

Dealing with heterogeneity

Generalized (non)linear least squares (GLS/GNLS)






```
        Model df      AIC      BIC     logLik   Test L.Ratio p-value
fit.lm      1  3 233.5692 238.4819 -113.78458
fit.gls2    2  4 195.3466 201.8969  -93.67328 1 vs 2 40.2226  <.0001
```

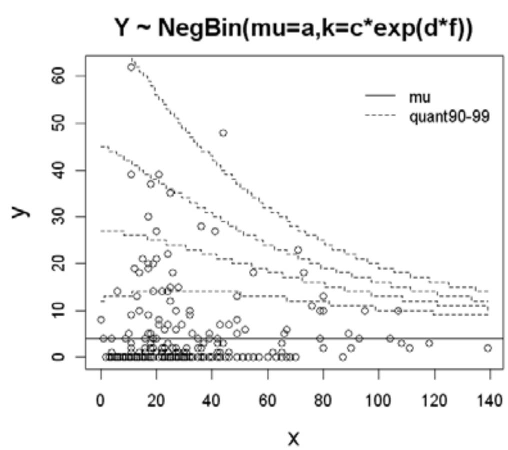

Landscape of Statistical Methods...

Dealing with heterogeneity

Customized linear/nonlinear models with *nonnormal* errors:

- In this example, the observed values of Y are *independent* (counts), distributed *negative binomial* with mean (*mu*) equal to a constant and the overdispersion parameter k (affecting the variance) varying exponentially as a function of X

from Bolker (2008)




```
       AIC df
1 1139.333  3
2 1113.235  3
```

```
Likelihood Ratio Tests
Model 1: fit.mu, [negbinNLL]: a+b+k
Model 2: fit.size, [negbinNLL]: a+c+d
  Tot Df Deviance  Chisq Df Pr(>Chisq)
1      3   1133.3
2      3   1107.2 26.098  0  < 2.2e-16
```

Landscape of Statistical Methods...

The basic statistical model:

Y = deterministic part + stochastic part

- Univariate
- Multivariate

- Linear
- Nonlinear
- Smoothed

- Distribution
- Heterogeneity
- Autocorrelation
- Multiple levels
- Random noise

Landscape of Statistical Methods...
Dealing with (auto-)correlated errors

1. Spatial auto-correlation

Linear model

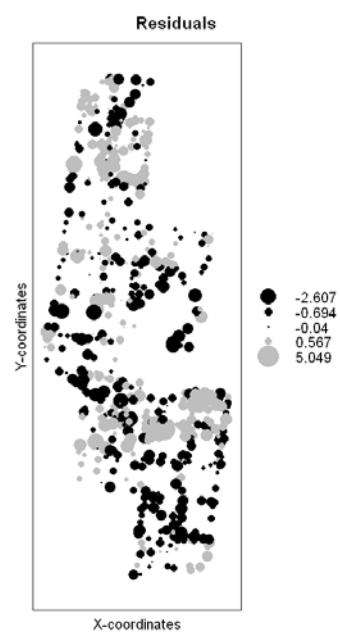
$$BFI_i = b_0 + b_1 \cdot Wetness_i + \varepsilon_i$$

$$\varepsilon_i \sim Normal\left(0, \sigma^2\right)$$

$$\operatorname{cov}\left(\varepsilon_i, \varepsilon_j\right) = \begin{Bmatrix} \sigma^2, i = j \\ 0, else \end{Bmatrix}$$

Constant variance assumption

Independence assumption



from Zuur et al (2009)

Landscape of Statistical Methods...

Dealing with (auto-)correlated errors

1. Spatial auto-correlation

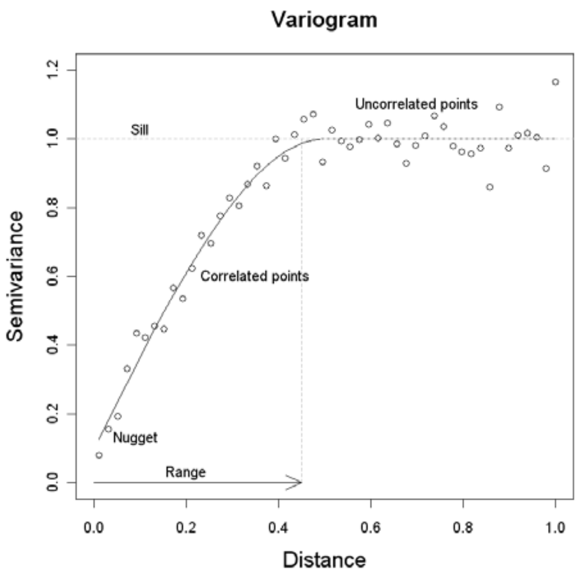
(Semi-)Variogram method:

Semivariance:

$$\gamma(x_i, x_j) = \frac{1}{2} E\left[\left(Z(x_i) - Z(x_j)\right)^2\right]$$

Experimental variogram:

$$\hat{\gamma}(h) = \frac{1}{2N(h)} \sum_{i=1}^{N(h)} \left[z(x_i + h) - z(x_i)\right]^2$$



from Zuur et al (2009)

Landscape of Statistical Methods...

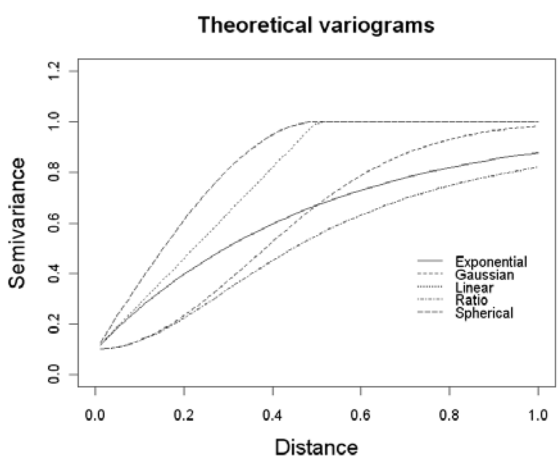
Dealing with (auto-)correlated errors

1. Spatial auto-correlation

Theoretical variograms:

- *Exponential* correlation (corExp)
- *Gaussian* corrrelation (corGaus)
- *Linear* correlation (corLin)
- *Rational* quadratic correlation (corRatio)
- *Spherical* correlation (corSpher)

$$cor(\varepsilon_i \varepsilon_j) = \begin{Bmatrix} 1, i = j \\ \boxed{h(\varepsilon_i, \varepsilon_j, \rho),} \; else \end{Bmatrix}$$



Landscape of Statistical Methods...

Dealing with (auto-)correlated errors

1. Spatial auto-correlation

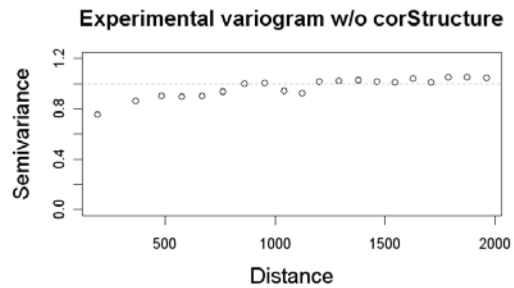
Experimental variogram:

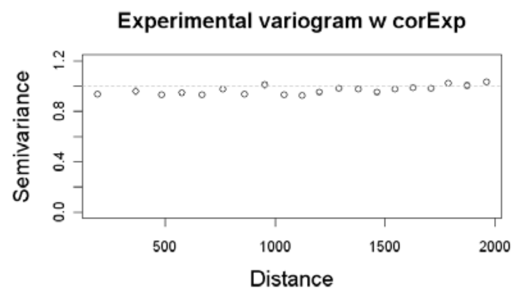
```
          df      AIC
B.none     3 2844.541
B.spher    5 2737.012
B.lin      5 2848.510
B.ratio    5 2732.930
B.gaus     5 2736.292
B.exp      5 2732.224
```

$$cor(\varepsilon_i \varepsilon_j) = \begin{cases} 1, \; i = j \\ 1 - e^{\frac{d}{r}}, \; else \end{cases}$$

d = distance
r = estimated range

```
       Model df      AIC      BIC    logLik   Test  L.Ratio p-value
B.none     1  3 2844.541 2857.365 -1419.271
B.exp      2  5 2732.224 2753.597 -1361.112 1 vs 2 116.3175  <.0001
```





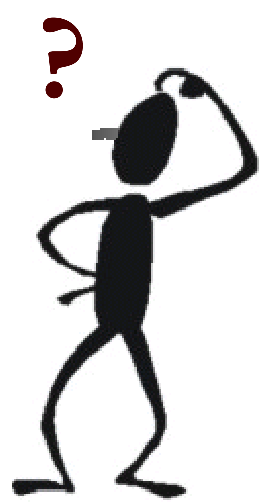
from Zuur et al (2009)

Landscape of Statistical Methods...

Dealing with (auto-)correlated errors

Other methods for dealing with spatial and temporal auto-correlation:

- Autocovariate models
- Spatial eigenvector mapping (SEVM)
- Generalized least squares (GLS)
  - Conditional autoregressive models (CAR)
  - Simultaneous autoregressive models (SAR)
  - Generalized linear mixed models (GLMM)
- Generalized estimating equations GEE)
- Wavelet-revised model (WRM)



Landscape of Statistical Methods...

The basic statistical model:

Y = deterministic part + stochastic part

- Univariate
- Multivariate

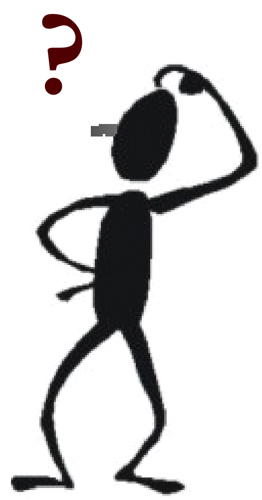
- Linear
- Nonlinear
- Smoothed

- Distribution
- Heterogeneity
- Autocorrelation
- Multiple levels
- Random noise

Landscape of Statistical Methods...

Dealing with multiple levels of error

- *Zero-inflated data* – models for data with too many zeros involving a mixture of two distributions
- *Nested data* – models for nested or blocked data, which divide observations into discrete groups according to their spatial or temporal locations, or other characteristics
- *Observation-process models* – models with process error and measurement (observation) error in the same model



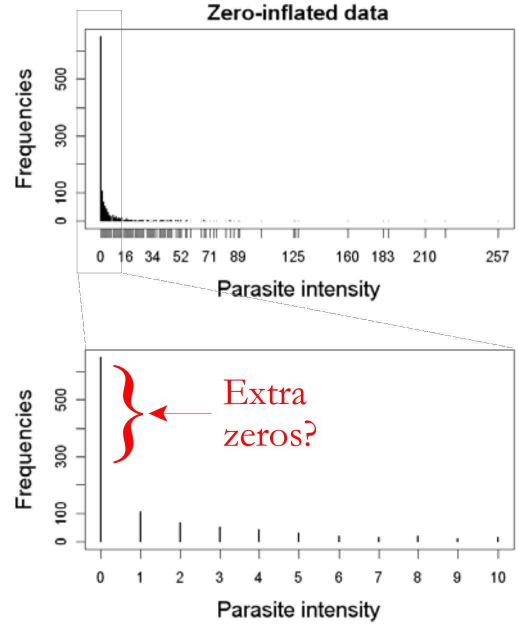
Landscape of Statistical Methods...

Dealing with multiple levels of error

1. Zero-inflated data

Why so many zeros?

- *True zeros* (positive zeros or true negatives) – structural errors, absent because the habitat is not suitable
- *False zeros* (false negatives) – due either to study design (in the wrong place or at the wrong time), survey method (low detectability), or observer error



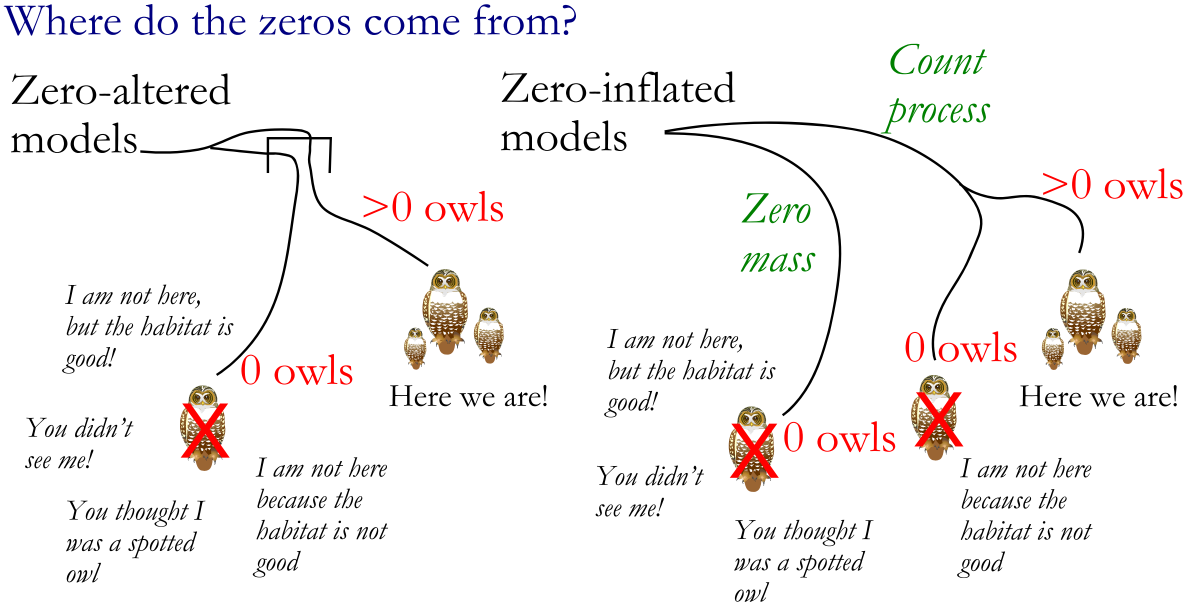
from Zuur et al (2009)

Landscape of Statistical Methods...

Dealing with multiple levels of error

1. Zero-inflated data

Where do the zeros come from?



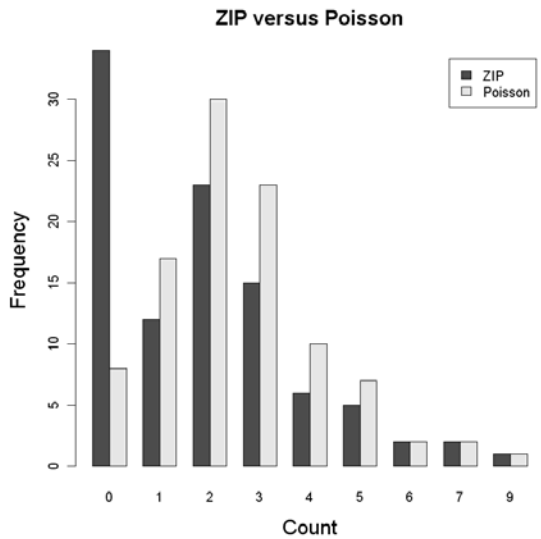
Landscape of Statistical Methods...

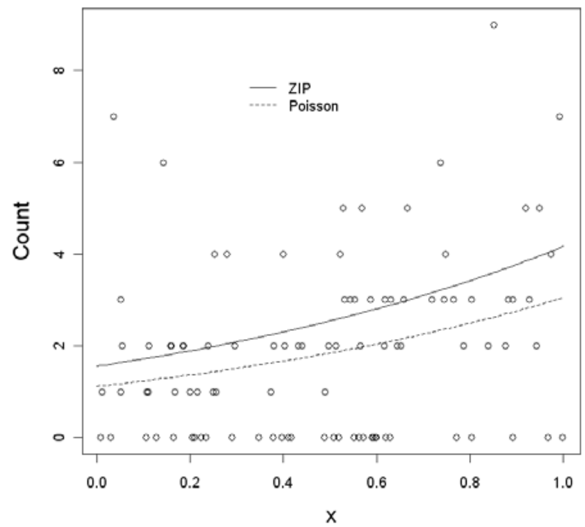
Dealing with multiple levels of error

1. Zero-inflated data

Example:

```
          df      AIC
fit.pois   2 386.9589
fit.zip1   3 354.6750
```



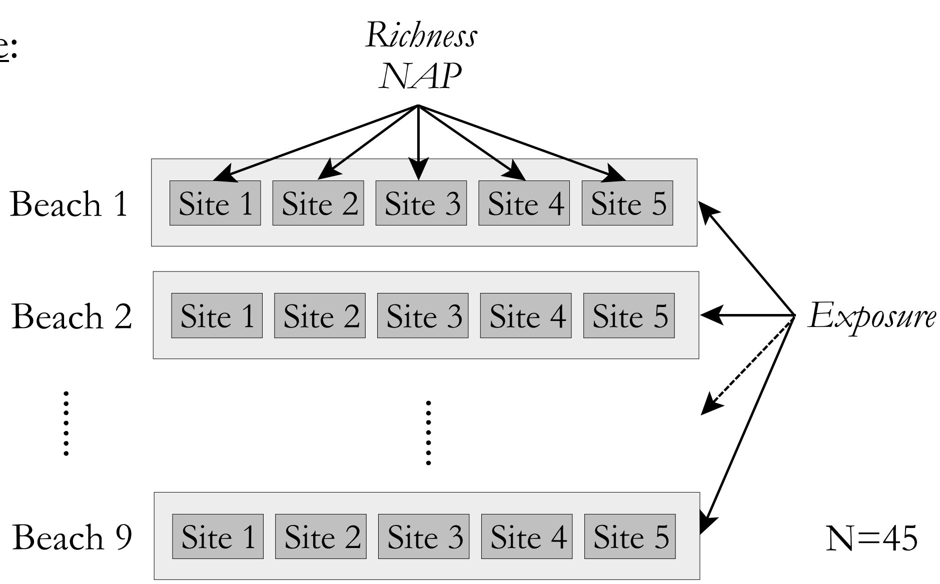


Landscape of Statistical Methods...

Dealing with multiple levels of error

2. Nested data

Example:



from Zuur et al (2009)

Landscape of Statistical Methods...

Dealing with multiple levels of error

2. Nested data

(a) Single-level model:

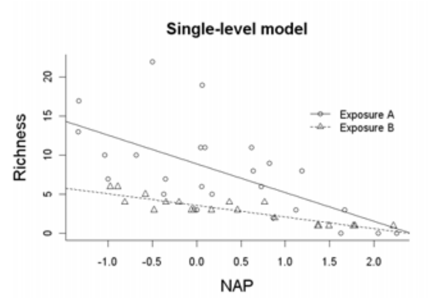
- Ignore nested structure

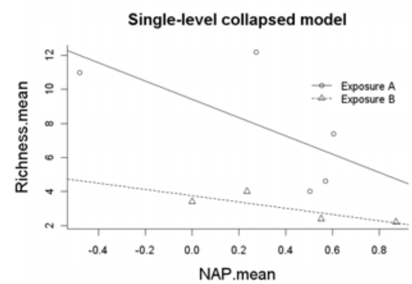
$$R_{ij} = \alpha + \beta_1 \cdot NAP_{ij} + \beta_2 Exposure_i + \varepsilon_{ij}$$
$$\varepsilon_{ij} \sim Normal\left(0, \sigma^2\right)$$
$$i = 1, \ldots, 9$$
$$j = 1, \ldots, 5$$

- Collapse to group level

$$\overline{R}_i = \alpha + \beta_1 \cdot \overline{NAP}_i + \beta_2 Exposure_i + \varepsilon_i$$

Neither approach is ideal





Landscape of Statistical Methods...

Dealing with multiple levels of error

2. Nested data

(b) Two-stage method:

- Stage 1 – separate regression for each beach

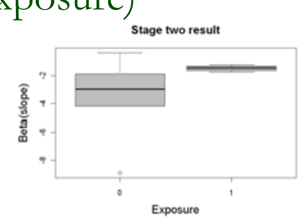
$$R_{ij} = \alpha + \beta_i \cdot NAP_{ij} + \varepsilon_{ij}$$
$$j = 1, \ldots, 5$$

- Stage 2 – model estimated regression coefficients as a function of group-level covariates (exposure)

$$\boxed{\hat{\beta}_i} = \eta + \tau \cdot Exposure_i + b_i$$
$$i = 1, \ldots, 9$$

Stage 1 result:

```
beach      betas
    1 -0.3718279
    2 -4.1752712
    3 -1.7553529
    4 -1.2485766
    5 -8.9001779
    6 -1.3885120
    7 -1.5176126
    8 -1.8930665
    9 -2.9675304
```



Stage 2 result:

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    -3.662      1.099  -3.332   0.0126 *
fExposure91     2.184      1.649   1.325   0.2268
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' '

Residual standard error: 2.458 on 7 degrees of freedom
Multiple R-squared: 0.2005,     Adjusted R-squared: 0.08625
F-statistic: 1.755 on 1 and 7 DF,  p-value: 0.2268
```

Landscape of Statistical Methods...

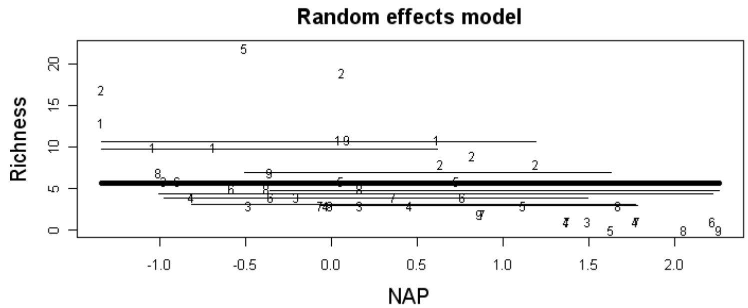
Dealing with multiple levels of error

2. Nested data

(c) Mixed-effects model:

- *Random effects model –* allows intercept to vary among beaches but no slope

$$R_{ij} \sim Normal\left(\mu_i, \sigma^2\right)$$
$$\mu_i = \beta_{0_{i|j|}}$$
$$\beta_{0_i} \approx Norm\left(\mu_{\beta_0}, \sigma^2_{\beta_0}\right)$$
$$\mu_{\beta_0} = \gamma_0 + \gamma_1 \cdot Exposure_i$$



Landscape of Statistical Methods...

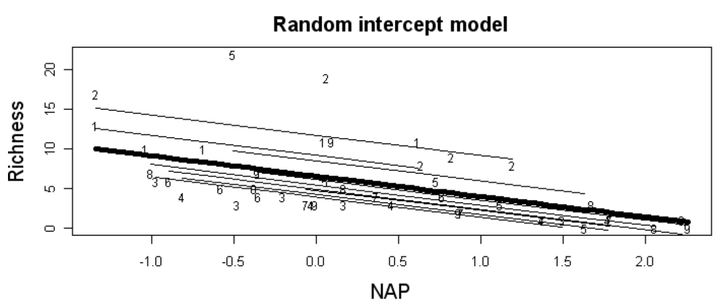
Dealing with multiple levels of error

2. Nested data

(c) Mixed-effects model:

- *Random intercept model –* allows intercept to vary among beaches

$$R_{ij} \sim Normal\left(\mu_i, \sigma^2\right)$$
$$\mu_i = \beta_{0_{i|j|}} + \beta_{1_{|j|}} \cdot NAP_{ij}$$
$$\beta_{0_i} \approx Norm\left(\mu_{\beta_0}, \sigma^2_{\beta_0}\right)$$
$$\mu_{\beta_0} = \gamma_0 + \gamma_1 \cdot Exposure_i$$



Landscape of Statistical Methods…

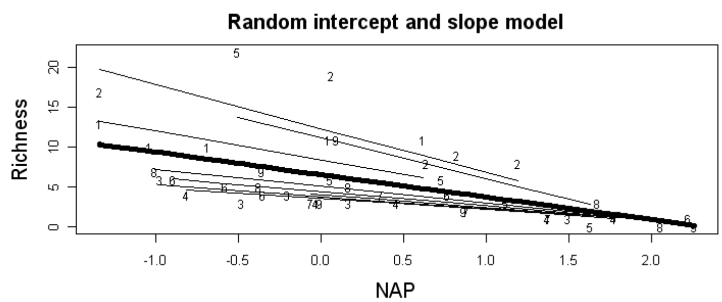
Dealing with multiple levels of error

2. Nested data

(c) Mixed-effects model:

- *Random intercept and slope model* – allows intercept and slope to vary among beaches

$$R_{ij} \sim Normal\left(\mu_i, \sigma^2\right)$$
$$\mu_i = \beta_{0_{i|j|}} + \beta_{1_{i|j|}} \cdot NAP_{ij}$$
$$\begin{pmatrix} \beta_{0_i} \\ \beta_{1_i} \end{pmatrix} \approx Norm\left(\begin{pmatrix} \mu_{\beta_0} \\ \mu_{\beta_1} \end{pmatrix}, \begin{pmatrix} \sigma^2_{\beta_0}, \rho\sigma_{\beta_0}\sigma_{\beta_1} \\ \rho\sigma_{\beta_0}\sigma_{\beta_1}, \sigma^2_{\beta_1} \end{pmatrix}\right)$$
$$\mu_{\beta_0} = \gamma_0 + \gamma_1 \cdot Exposure_i$$
$$\mu_{\beta_1} = \tau_0 + \tau_1 \cdot Exposure_i$$



Landscape of Statistical Methods…

Dealing with multiple levels of error

3. Observation-Process models

Models that account for the ecological process and the observation process separately in a single model

"Few animals are so conspicuous that they are always detected at each survey."
MacKenzie et al. (2002)

- When detection bias is suspected to be significant, it is necessary to account for it in the model to achieve accurate estimates of the parameters associated with the ecological process of interest

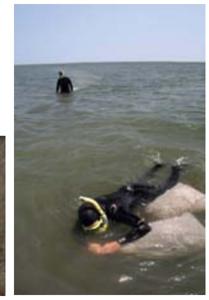
Landscape of Statistical Methods...

Dealing with multiple levels of error

3. Observation-Process models

Example:

Estimate occupancy rate of an invasive species of crab along a coastline in relation to the percent of the substrate covered by cobbles - a potentially important habitat covariate



site	survey.1	survey.2	survey.3	waterClarity.1	waterClarity.2	waterClarity.3	% cover cobbles
1	0	0	0	3.06	1.14	1.92	75.1
2	0	0	0	1.79	0.72	0.54	79.9
3	1	1	1	6.61	9.18	5.43	28.1
4	1	1	1	8.68	8.51	7.92	19.4
5	0	0	0	2.49	1.68	2.91	91.0
6	1	0	1	9.98	6.80	8.44	100.0
7	1	1	0	7.95	7.38	8.74	90.2
.							
.							
.							
100	0	0	0	6.59	8.41	8.31	84.6

from Royle and Dorazio (2008)

Landscape of Statistical Methods...

Dealing with multiple levels of error

3. Observation-Process models

Statistical model:

Process model

$$\boxed{z_i} \sim Bern(\psi_i)$$

$$Logit(\psi_i) = \beta_0 + \beta_1 \cdot \boxed{Cobble_i}$$

Process covariate

Observation model

$$y_{ij} \sim Bern(p_{ij} \cdot \boxed{z_i})$$

$$Logit(p_{ij}) = \alpha_0 + \alpha_1 \cdot \boxed{waterClarity_{ij}}$$

Observation covariate

Z_i = Unobserved state variable (presence/absence at site *i*)

y_{ij} = Observed data (detected/not detected at site *i* on survey *j*)

a_i, β_i = Parameters to estimate

Landscape of Statistical Methods...

Dealing with multiple levels of error

3. Observation-Process models

Model selection:

(1)
$$z_i \sim Bern(\psi_i)$$
$$y_{ij} \sim Bern(p_{ij} \cdot z_i)$$

(2)
$$z_i \sim Bern(\psi_i)$$
$$Logit(\psi_i) = \beta_0 + \beta_1 \cdot Cobble_i$$
$$y_{ij} \sim Bern(p_{ij} \cdot z_i)$$

(3)
$$z_i \sim Bern(\psi_i)$$
$$y_{ij} \sim Bern(p_{ij} \cdot z_i)$$
$$Logit(p_{ij}) = \alpha_0 + \alpha_1 \cdot waterClarity_{ij}$$

(4)
$$z_i \sim Bern(\psi_i)$$
$$Logit(\psi_i) = \beta_0 + \beta_1 \cdot Cobble_i$$
$$y_{ij} \sim Bern(p_{ij} \cdot z_i)$$
$$Logit(p_{ij}) = \alpha_0 + \alpha_1 \cdot waterClarity_{ij}$$

Model	n	K	AIC	ΔAIC	AICwt	R-squared	AICwtCum	Ψ	S.E.(Ψ)
p(clarity), psi(cobbles)	100	4	210.67	0.000	0.679	0.310	0.679	0.418	0.088
p(clarity), psi(.)	100	3	212.16	1.497	0.321	0.281	1.000	0.420	0.070
p(.), psi(cobbles)	100	3	237.55	26.885	0.000	0.042	1.000	0.318	0.067
p(.), psi(.)	100	2	239.31	28.642	0.000	0.000	1.000	0.320	0.050

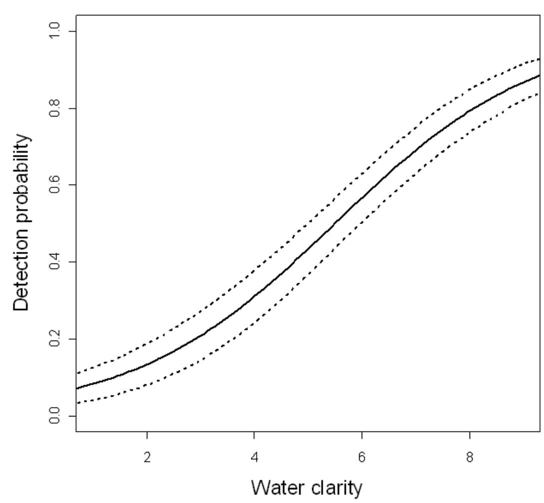
Landscape of Statistical Methods...

Dealing with multiple levels of error

3. Observation-Process models

Detectability function:

- Estimating the detectability function can be useful in the design of future studies



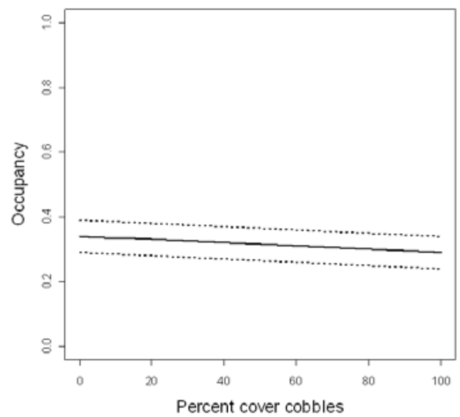
Landscape of Statistical Methods…

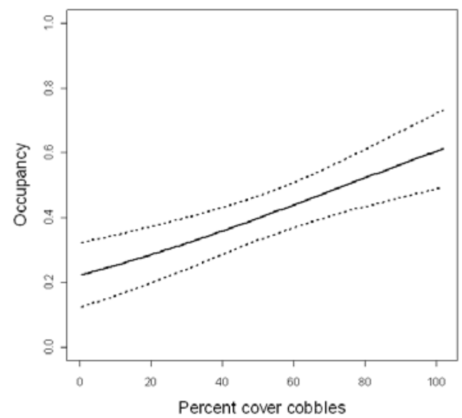
Dealing with multiple levels of error

3. Observation-Process models

Occupancy function:

- Estimates of occupancy can be significantly biased and even misleading if detectability is not taken into account





Landscape of Statistical Methods…

The basic statistical model:

Y = deterministic part + stochastic part

- Univariate
- Multivariate

- Linear
- Nonlinear
- Smoothed

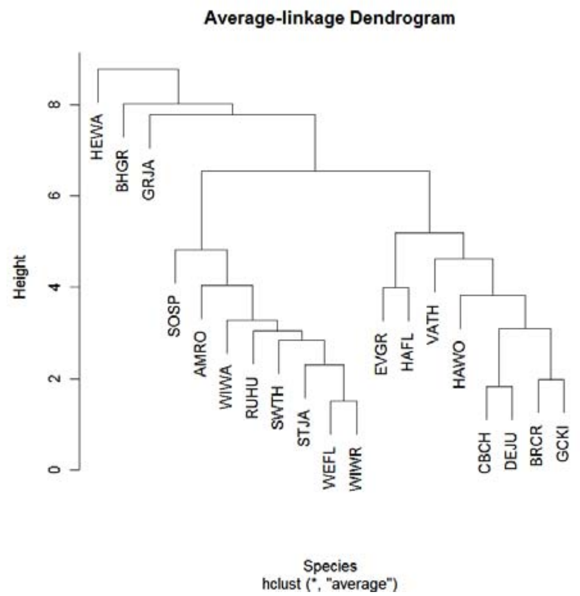
- Distribution
- Heterogeneity
- Autocorrelation
- Multiple levels
- Random noise

Landscape of Statistical Methods...

Mulivariate methods

Finding groups (Cluster analysis)

- Large family of techniques with similar goals; operating on data sets for which *pre-specified, well-defined groups do "not" exist*; characteristics of the data are used to assign entities into artificial groups

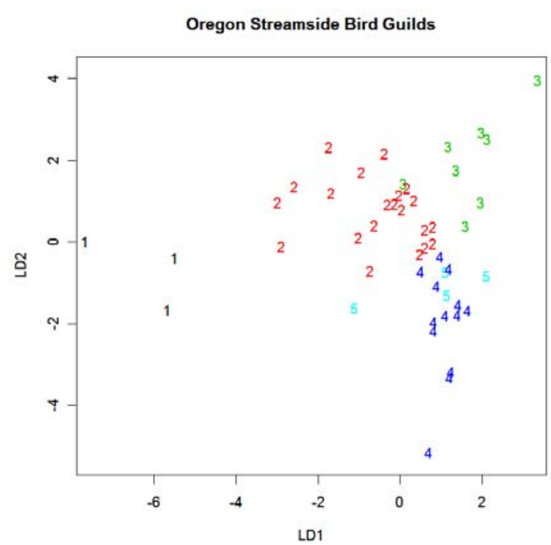


Landscape of Statistical Methods...

Mulivariate methods

Testing/describing differences among groups (e.g., DA, ISA, mCART, MRPP, MANTEL)

- Large family of different methods for testing and/or describing differences among *pre-specified, well-defined groups* based on a set of discriminating variables

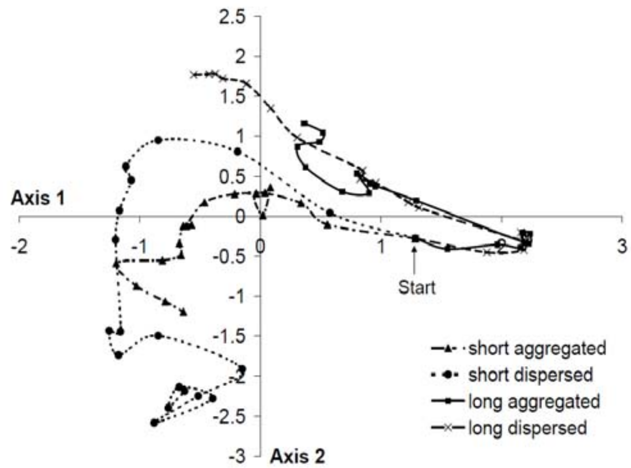


Landscape of Statistical Methods…

Mulivariate methods

Unconstrained ordination (e.g., PCA, CA, NMDS)

- A family of different methods for organizing sampling entities (e.g., species, sites, observations, etc.) along continuous gradients based on a set of interdependent variables



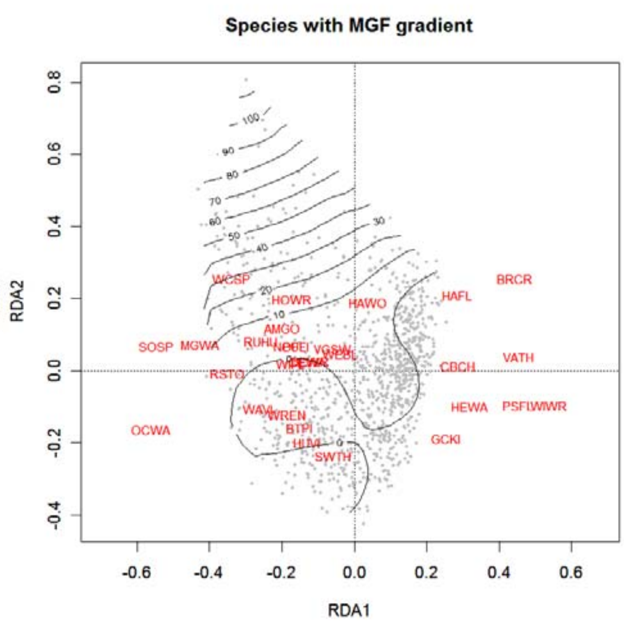
Landscape of Statistical Methods…

Mulivariate methods

Constrained ordination (e.g., RDA, CCA, CAPS)

- A family of different methods for extending unconstrained ordination in which the solution is constrained to be expressed by ancillary variables

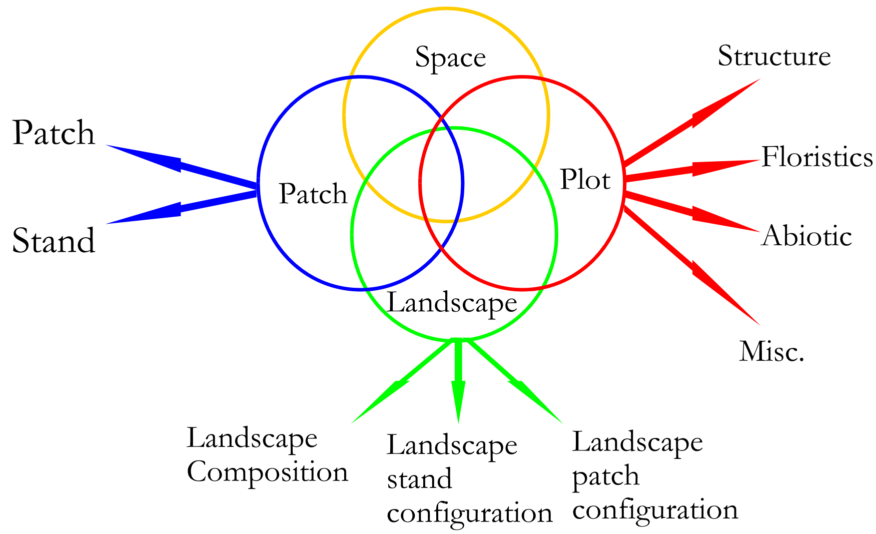
Triplot (samples, species, environment)



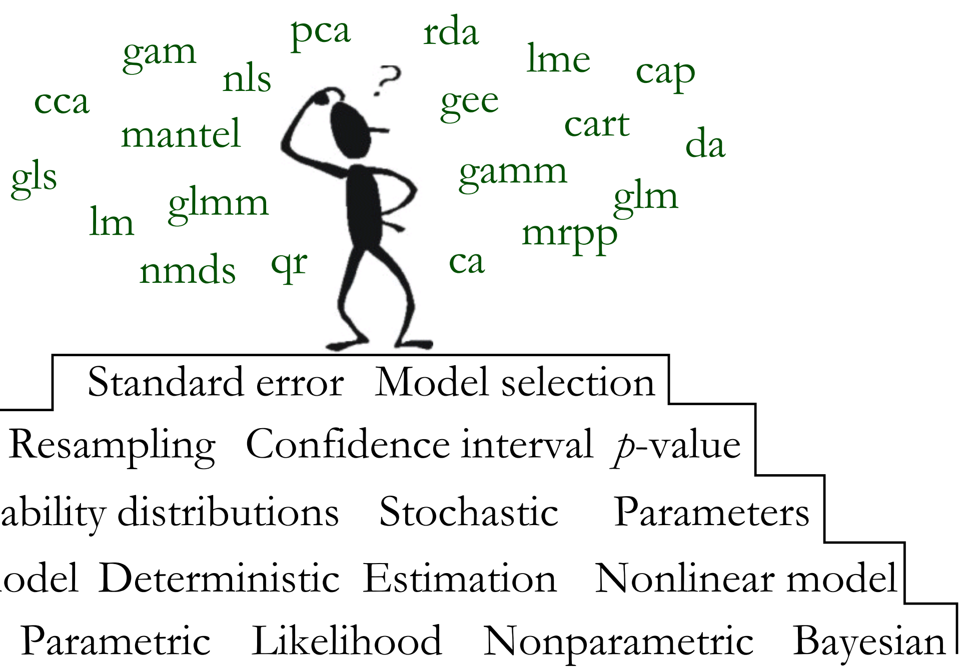
Landscape of Statistical Methods...

Mulivariate methods

Ordination variance partioning



Foundation for Understanding Statistical Methods...



Foundation for Understanding Statistical Methods...

Statistical foundations