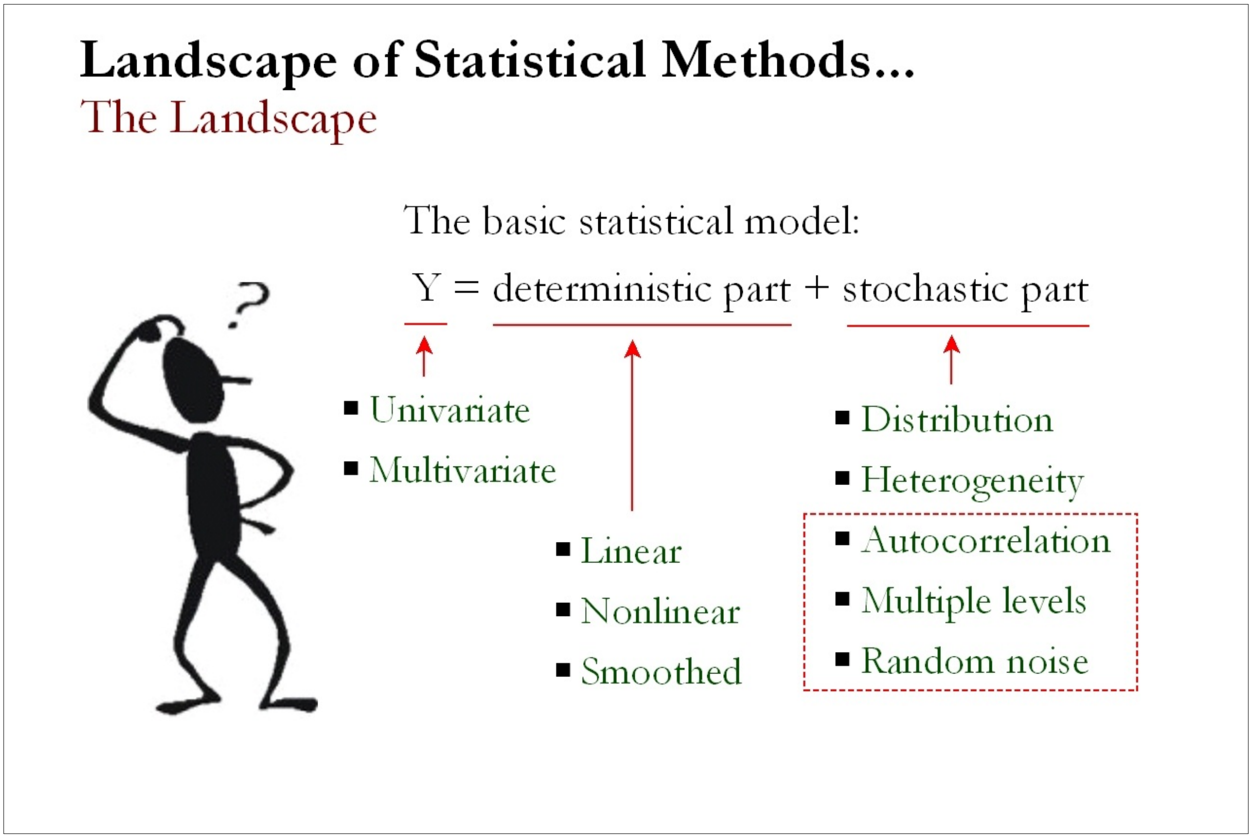
# Design and Analysis of Ecological Data
## Landscape of Statistical Methods: Part 2

*Much of the material in this section is taken from Bolker (2008) and Zur et al. (2009)



## Landscape of Statistical Methods...

The Landscape

The basic statistical model:

Y = deterministic part + stochastic part

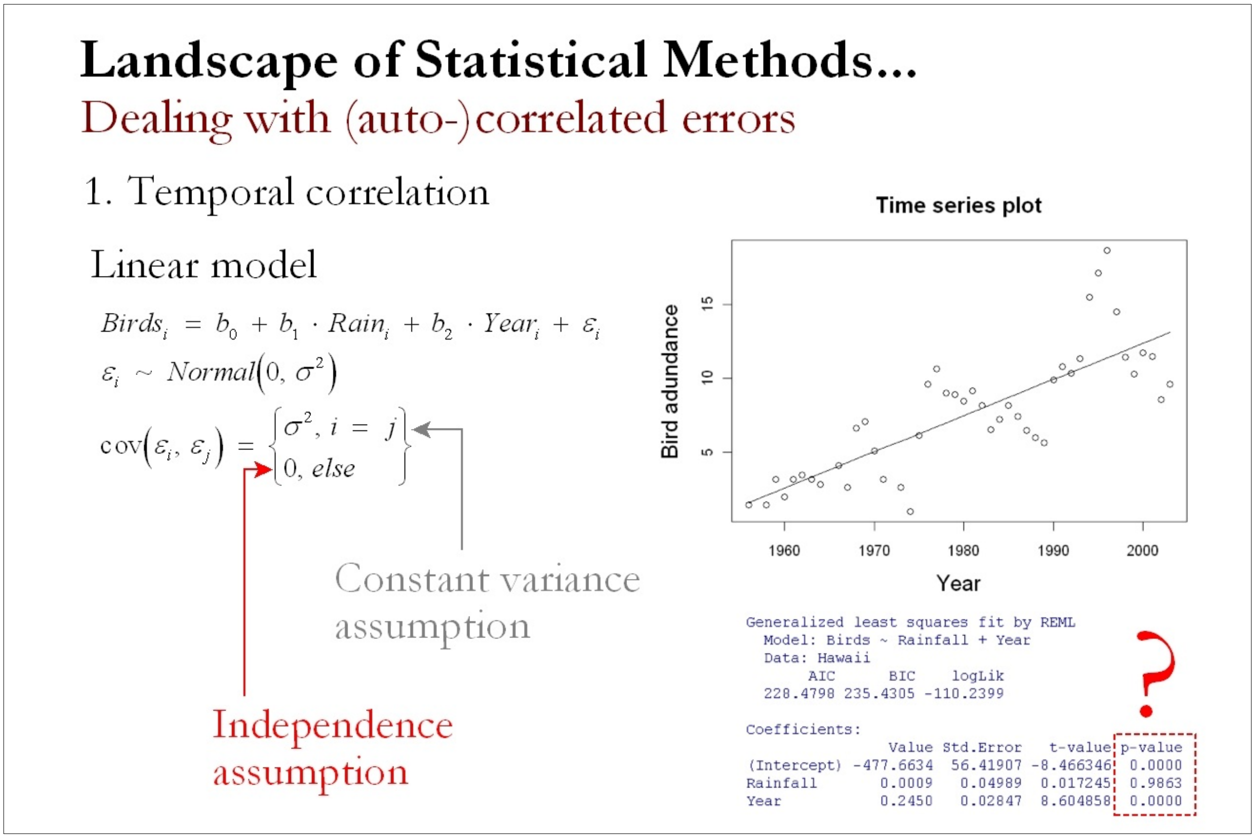
- Univariate
- Multivariate

- Linear
- Nonlinear
- Smoothed

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

## 1. Correlated errors

Up to now we have assumed that the observations in a data set are all independent. When this is true, the likelihood of the entire data set is equal to the product of the likelihoods of each data point, and the (negative) log-likelihood of the data set is equal to the sum of the (negative) log-likelihoods of each point. With considerably more effort, however, we can write and numerically optimize likelihood functions that allow for correlations among observations. However, it is best to avoid correlation entirely by designing the observations or experiments appropriately. Correlations among data points can be a serious headache to model, and always reduces the total amount of information in the data: correlation means that data points are more similar to each other than expected by chance (i.e., they are partially redundant bits of information), so the total amount of information in the data is smaller than if the data were independent. However, sometimes it is difficult, impractical, or simply impossible to avoid correlation, for example when the data come from a spatial array or time series. Moreover, sometimes the correlation in the data is ecologically meaningful; for example, the range of spatial correlation might indicate the spatial scale at which some underlying ecological process operates.



# Landscape of Statistical Methods…

## Dealing with (auto-)correlated errors

1. Temporal correlation

Linear model

$$Birds_i = b_0 + b_1 \cdot Rain_i + b_2 \cdot Year_i + \varepsilon_i$$

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

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

Constant variance assumption

Independence assumption

```
Generalized least squares fit by REML
  Model: Birds ~ Rainfall + Year
  Data: Hawaii
       AIC      BIC    logLik
  228.4798 235.4305 -110.2399

Coefficients:
                Value Std.Error   t-value p-value
(Intercept) -477.6634  56.41907 -8.466346  0.0000
Rainfall       0.0009   0.04989  0.017245  0.9863
Year           0.2450   0.02847  8.604858  0.0000
```

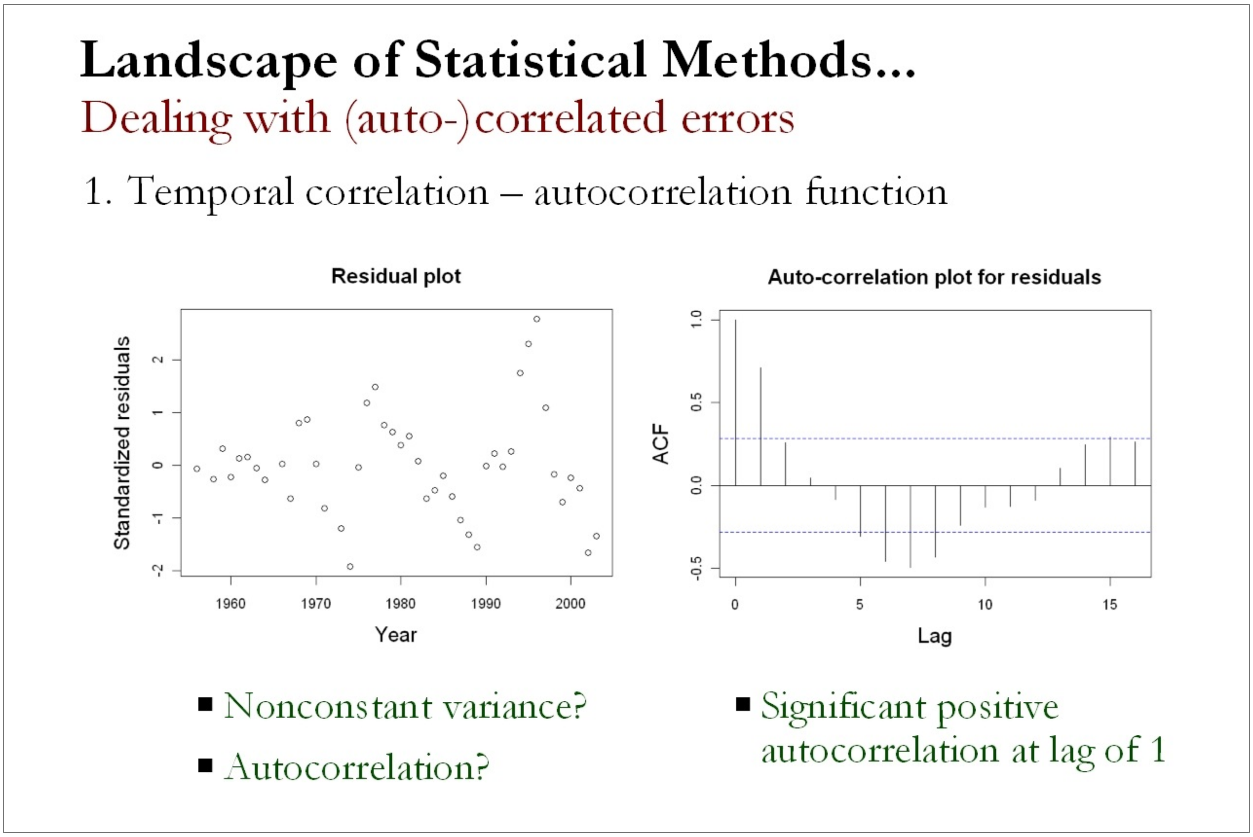
### 1.1 Temporal correlation

Let's begin with the issue of temporally correlated observations that arise in repeated measures or time series data, where the value of any observation is likely to be correlated with the value of the observations nearby in time. In other words, if the observations are spaced close together in time relative to the temporal scale at which the phenomenon under study changes, then we will probably be able to predict the value of an observation at least partially by the previous value or values. In this situation we clearly violate the assumption of independent observations (or rather, independent errors) and somehow must account for this in the model. Fortunately, if we can specify a reasonable form for this temporal autocorrelation, then we may be able to address it in the model.

To illustrate this, let's take an example of a multiple linear regression model involving the abundance of a bird species (Moorhen) measured at one site (Kauai, Hawaii) annually from 1956 to 2003 (example taken from Zur et al. 2009, chapter 6). The square root transformed abundance of birds is modeled as a function of annual rainfall and the variable Year (representing a long-term trend) using the following linear regression model:

$$Birds_i = b_0 + b_1 \cdot Rain_i + b_2 \cdot Year_i + \varepsilon_i$$
$$\varepsilon_i \sim Normal\left(0, \sigma^2\right)$$
$$\mathrm{cov}\left(\varepsilon_i, \varepsilon_j\right) = \begin{Bmatrix} \sigma^2, i = j \\ 0, else \end{Bmatrix}$$

Note, we have written the model in a slightly different form than usual in order to highlight the independent error assumption. The model reads: bird abundance at time *i* is a linear function of rainfall at time *i* plus year at time *i* plus a random error, where the errors are *normally distributed* with mean zero (centered on the predicted value of birds at time *i*) and a *constant variance*, and the covariance between the errors at time *i* and time *j* is equal to the variance if $i = j$ and zero if $i \neq j$. Note that the *covariance* between a data point and itself ($i = j$) is simply its variance, and if the covariance between two points is zero they are statistically independent (i.e., they don't covary). Thus, the covariance part of this model simply says that the expected variance for each point is the same for all points (i.e., constant variance) and that they are completely independent. When we assume independence among observations, we usually leave off the covariance description of the model for convenience – but it is nonetheless still there.



# Landscape of Statistical Methods…

## Dealing with (auto-)correlated errors

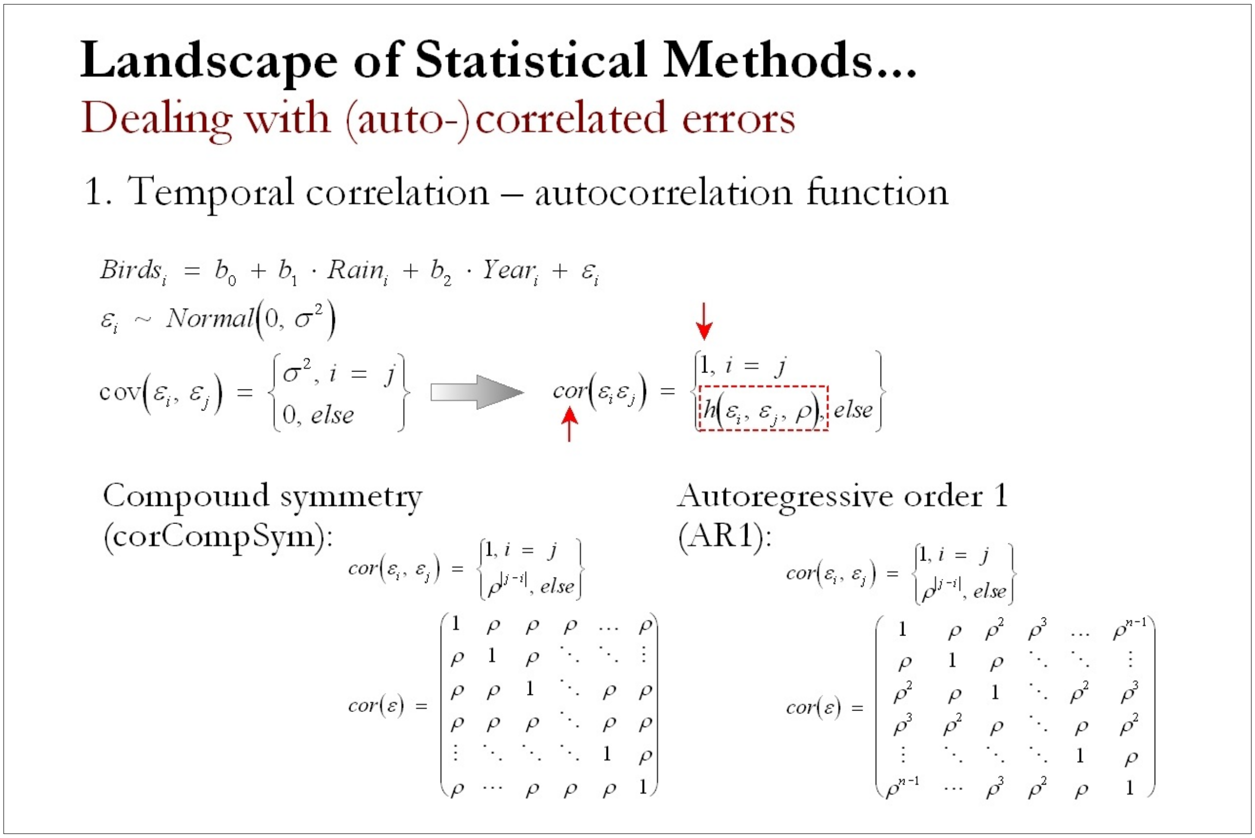
1. Temporal correlation – autocorrelation function

Residual plot

Auto-correlation plot for residuals

- Nonconstant variance?
- Autocorrelation?
- Significant positive autocorrelation at lag of 1

*Autocorrelation Function*
If the independence assumption is not met we cannot trust the *p*-values associated with the significance tests for each of the parameters. To evaluate this assumption, the first thing we can do is plot the standardized residuals against time. Note that there is a clear pattern in the residuals; in fact, two patterns emerge. The most obvious pattern is the increasing spread of residuals with time, which suggests that the variance is not constant over time. We have already discussed how to incorporate heterogeneity using variance covariates. Here, we are going to focus instead on the independence assumption. The second pattern (and the one that concerns us here) is the apparent autocorrelation in the residuals: residuals closer together in time are more alike than those farther apart in time. A more formal tool to detect autocorrelation patterns is the *autocorrelation function* (ACF). The value of the ACF at different time lags gives an indication of whether this is any autocorrelation in the data. Note in the example here that the ACF plot shows a clear violation of the independence assumption; various time lags have a significant correlation. The ACF plot has a general pattern of decreasing values for the first 5 years.



# Landscape of Statistical Methods…

## Dealing with (auto-)correlated errors

### 1. Temporal correlation – autocorrelation function

$$Birds_i = b_0 + b_1 \cdot Rain_i + b_2 \cdot Year_i + \varepsilon_i$$

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

$$\operatorname{cov}(\varepsilon_i, \varepsilon_j) = \begin{Bmatrix} \sigma^2, i = j \\ 0, else \end{Bmatrix} \Longrightarrow cor(\varepsilon_i \varepsilon_j) = \begin{Bmatrix} 1, i = j \\ \boxed{h(\varepsilon_i, \varepsilon_j, \rho),} else \end{Bmatrix}$$

Compound symmetry (corCompSym):

$$cor(\varepsilon_i, \varepsilon_j) = \begin{Bmatrix} 1, i = j \\ \rho^{|j-i|}, else \end{Bmatrix}$$

$$cor(\varepsilon) = \begin{pmatrix} 1 & \rho & \rho & \rho & \cdots & \rho \\ \rho & 1 & \rho & \ddots & \ddots & \vdots \\ \rho & \rho & 1 & \ddots & \rho & \rho \\ \rho & \rho & \rho & \ddots & \rho & \rho \\ \vdots & \ddots & \ddots & \ddots & 1 & \rho \\ \rho & \cdots & \rho & \rho & \rho & 1 \end{pmatrix}$$

Autoregressive order 1 (AR1):

$$cor(\varepsilon_i, \varepsilon_j) = \begin{Bmatrix} 1, i = j \\ \rho^{|j-i|}, else \end{Bmatrix}$$

$$cor(\varepsilon) = \begin{pmatrix} 1 & \rho & \rho^2 & \rho^3 & \cdots & \rho^{n-1} \\ \rho & 1 & \rho & \ddots & \ddots & \vdots \\ \rho^2 & \rho & 1 & \ddots & \rho^2 & \rho^3 \\ \rho^3 & \rho^2 & \rho & \ddots & \rho & \rho^2 \\ \vdots & \ddots & \ddots & \ddots & 1 & \rho \\ \rho^{n-1} & \cdots & \rho^3 & \rho^2 & \rho & 1 \end{pmatrix}$$

So how do we account for this autocorrelation in the model. The underlying principle is rather simple; instead of using the '0 else' in the covariance model, we model the autocorrelation between residuals of different time points by introducing a function *h*(.):

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

Note here that we have expressed the covariance structure as a *correlation* structure instead, where correlation is simply the standardized covariance, such that the correlation between an observation and itself is 1 and the correlation between any two different observations ranges from -1 to 1. This means we assume that the correlation between the residuals at time *i* and *j* only depends on their time difference *i* - *j*. Hence, the correlation between residuals at time *i* and *j* is assumed to be the same as that between time *i*+1 and *j*+1, between time *i*+2 and *j*+2, etc. Our task is to find the optimal parameterization of *h*(.)

There are a number of common correlation structures that can easily be incorporated into the model using existing functions, or we can always generate our own custom model if none of the existing models are appropriate. Here are some examples:

- *Compound symmetry* (corCompSymm) – in this autocorrelation structure we assume that whatever the distance in time between two observations, their residual correlation is the same. This can be modeled as follows (and the resulting correlation structure):

$$cor\left(\varepsilon_i, \varepsilon_j\right) = \left\{\begin{matrix} 1, i = j \\ \rho^{|j-i|}, else \end{matrix}\right\}$$

$$cor(\varepsilon) = \begin{pmatrix} 1 & \rho & \rho & \rho & \cdots & \rho \\ \rho & 1 & \rho & \ddots & \ddots & \vdots \\ \rho & \rho & 1 & \ddots & \rho & \rho \\ \rho & \rho & \rho & \ddots & \rho & \rho \\ \vdots & \ddots & \ddots & \ddots & 1 & \rho \\ \rho & \cdots & \rho & \rho & \rho & 1 \end{pmatrix}$$

- *Autoregressive model of order 1* (corAR1) – in this autocorrelation structure we model the residual at time *i* as a function of the residual of time *i - j* along with noise, as follows:

$$cor\left(\varepsilon_i, \varepsilon_j\right) = \left\{\begin{matrix} 1, i = j \\ \rho^{|j-i|}, else \end{matrix}\right\}$$

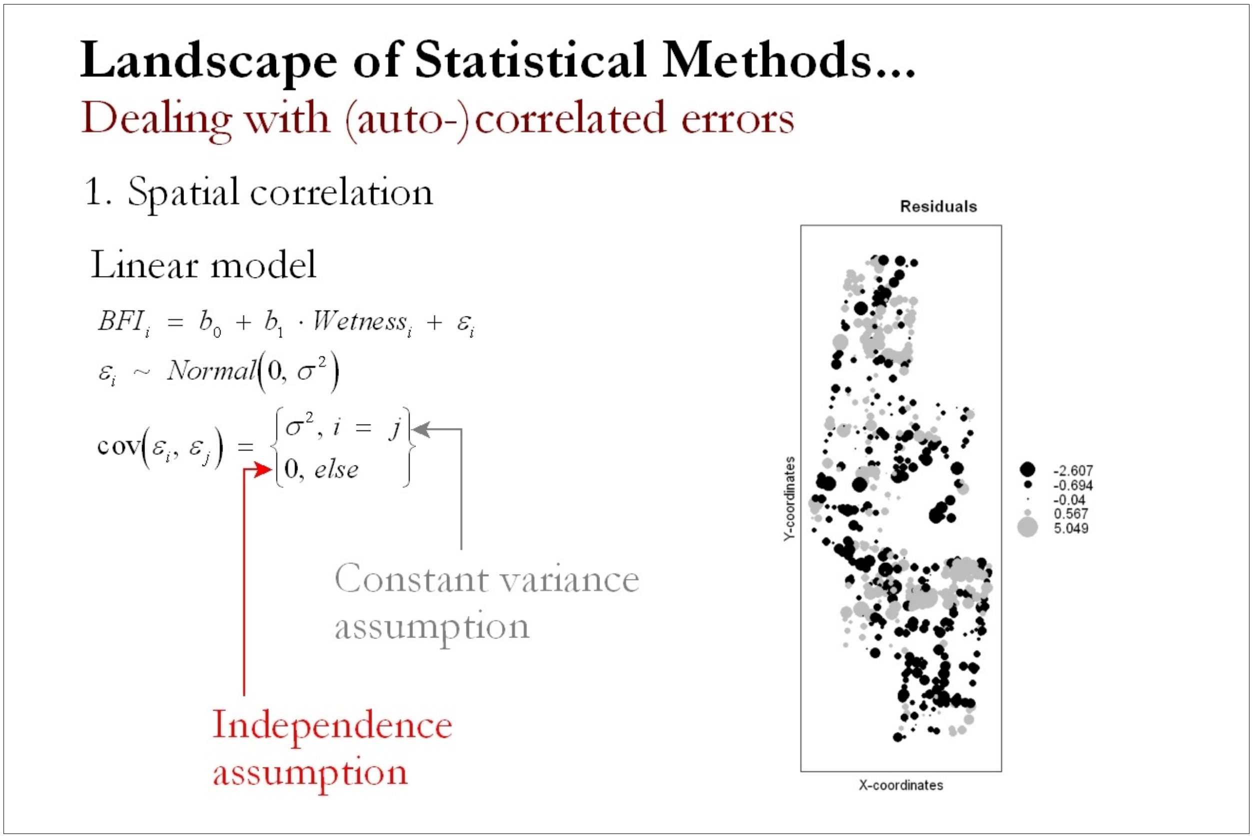
$$cor(\varepsilon) = \begin{pmatrix} 1 & \rho & \rho^2 & \rho^3 & \cdots & \rho^{n-1} \\ \rho & 1 & \rho & \ddots & \ddots & \vdots \\ \rho^2 & \rho & 1 & \ddots & \rho^2 & \rho^3 \\ \rho^3 & \rho^2 & \rho & \ddots & \rho & \rho^2 \\ \vdots & \ddots & \ddots & \ddots & 1 & \rho \\ \rho^{n-1} & \cdots & \rho^3 & \rho^2 & \rho & 1 \end{pmatrix}$$

  In this model the parameter $\rho$ is unknown and needs to be estimated from the data. With this correlation structure, if $\rho = 0.5$, then the correlation between residuals separated by one unit in time is 0.5; if the separation is two units in time the correlation is $0.5^2 = 0.25$; if it is three units in time the correlation is $0.5^3 = 0.125$; and so on. Hence, the further away two residuals are separated in time, the lower their correlation. For many ecological examples, this makes sense.

- *Autoregressive moving average* (ARMA) – in this autocorrelation structure we extend the AR1 model to handle more complex structures involving a combination of an autoregressive model of specified order and a moving average model of specified order, where order refers to the number of previous time steps considered. Unlike the previous two structures, the function *h*(.) for an ARMA structure does not have any easy formulation. A description of these ARMA structures is beyond the scope of our survey, but suffice it to say that it can be seen as somewhat of a black box to fix correlation problems.

- *Other structures* – not surprisingly, there are several other established correlation structures.

The good news is that may not be that important to find the perfect correlation structure; finding one that is adequate may be sufficient. Selecting an adequate structure can be accomplished by trying a number of structures and using AIC to select the best.



**Landscape of Statistical Methods…**

Dealing with (auto-)correlated errors

1. Spatial correlation

Linear model

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

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

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

Constant variance assumption

Independence assumption

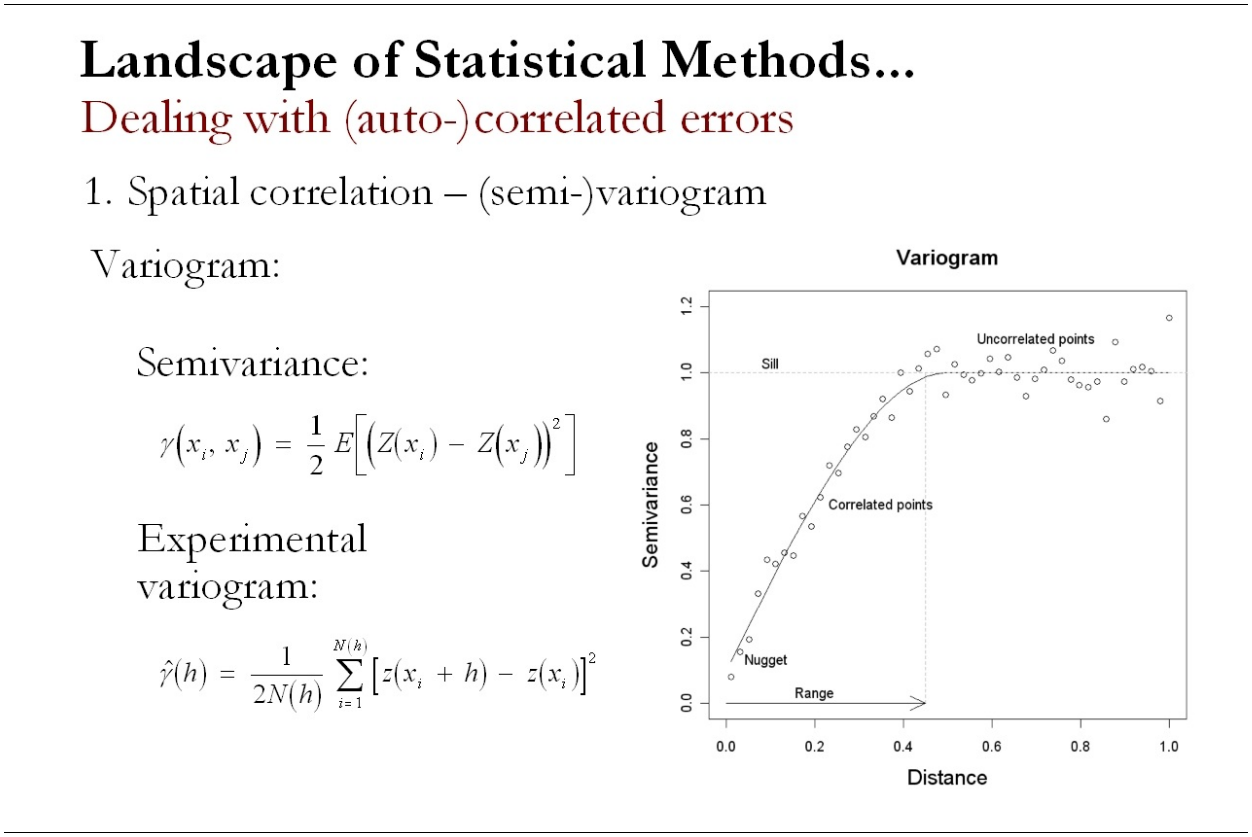
## 1.2 Spatial correlation

The issue of temporally correlated errors is effectively the same as for spatially correlated errors. The problem of spatially correlated errors is particularly relevant in ecological studies. The general principle with spatial data is that things that are close together are likely to be more alike than things that are farther apart. Thus, if we know the value of a variable at a particular sample location, we will probably be able to predict, albeit imperfectly, the value for nearby locations. In this situation, if our observations are too close together in space, we clearly violate the assumption of independent observations (or rather, independent errors) and must account for this in the model. Fortunately, as before, if we can specify a reasonable form for this spatial autocorrelation, then we may be able to address it in the model.

To illustrate this, let's take an example of a simple linear regression model involving the relative abundance of boreal tree species measured at 533 sites in Russia (example taken from Zur et al. 2009, chapter 7). The boreal forest index (BFI) is modeled as a function of a wetness index derived from LANDSAT imagery using the following simple linear regression model:

$$
\begin{aligned}
BFI_i &= b_0 + b_1 \cdot Wetness_i + \varepsilon_i \\
\varepsilon_i &\sim Normal\left(0, \sigma^2\right) \\
\operatorname{cov}\left(\varepsilon_i, \varepsilon_j\right) &= \begin{Bmatrix} \sigma^2, i = j \\ 0, else \end{Bmatrix}
\end{aligned}
$$

Note that in this model we assume that the errors are independent; i.e., that the covariance between any two different points is zero. Based on the results of fitting this linear model, it appears that wetness is highly significant and based on a plot of the residuals against the fitted values, it appears that homogeneity is a reasonable assumption. As a first step to verify independence, we can plot the residuals versus their spatial location using a "bubble plot". The size of the dots is proportional to the value of the residuals. This plot should not show any spatial pattern; e.g., groups of negative or positive residuals close together). If it does, then there may be a missing covariate or spatial correlation. In this case, there seems to be some spatial correlation as most of the positive residuals as well as the negative residuals are showing some clustering.



# Landscape of Statistical Methods...

## Dealing with (auto-)correlated errors

1. Spatial correlation – (semi-)variogram

Variogram:

Semivariance:

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

Experimental variogram:

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

*(Semi)Variogram*

An alternative to the informal approach of making a bubble plot of the residuals and judging whether there is spatial dependence is to make a variogram of the residuals. A *variogram* (or semivariogram) is an alternative to the ACF we used above. The variogram is defined by:

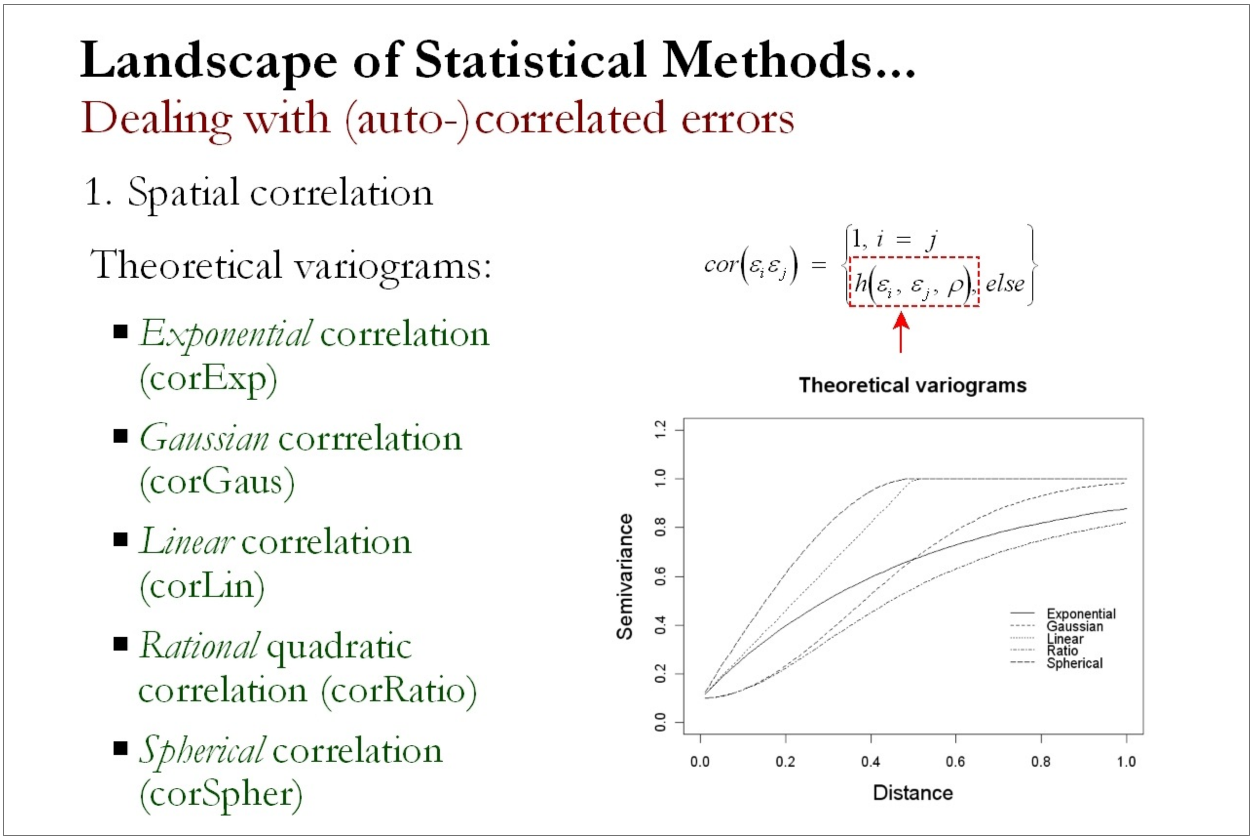
$$\gamma\left(x_i, x_j\right) = \frac{1}{2} E\left[\left(Z(x_i) - Z(x_j)\right)^2\right]$$

This is a function that measures the spatial dependence between two sites with coordinates $x_i$ and $x_j$. *Gamma* is called the *semivariance* because it measures half the variance between the two sites. Recall that variance is the squared difference between a value and its expected value. If these two sites are located close to each other, then we would expect the values of the variables of interest (residuals in this case) to be similar. A low value of *semivariance* indicates that this is indeed the case – the values are dependent, whereas a large value indicates independence. Under some assumptions that we will ignore for now, this leads to the sample (or experimental) variogram:

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

where *h* represents a distance class and *N*(*h*) represents the number of points that fall in the *h*th distance class.. The 'hat' on the *gamma* refers to the fact that it is an estimator based on sample data.

The variogram depicts the estimated semivariance (*gamma*) against lag distance (*h*). In the variogram of residuals against lag distance, spatial dependence shows itself as an increasing band of points, which then levels off at a certain distance. The point along the x-axis at which this pattern levels off is called the *range*, and the y-value at the range is the *sill*. The *nugget* is the y-value when the distance is 0; it represents the discontinuity of the variable caused by spatial processes at distances less than the minimum distance between points.



# Landscape of Statistical Methods…

## Dealing with (auto-)correlated errors

1. Spatial correlation

Theoretical variograms:

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

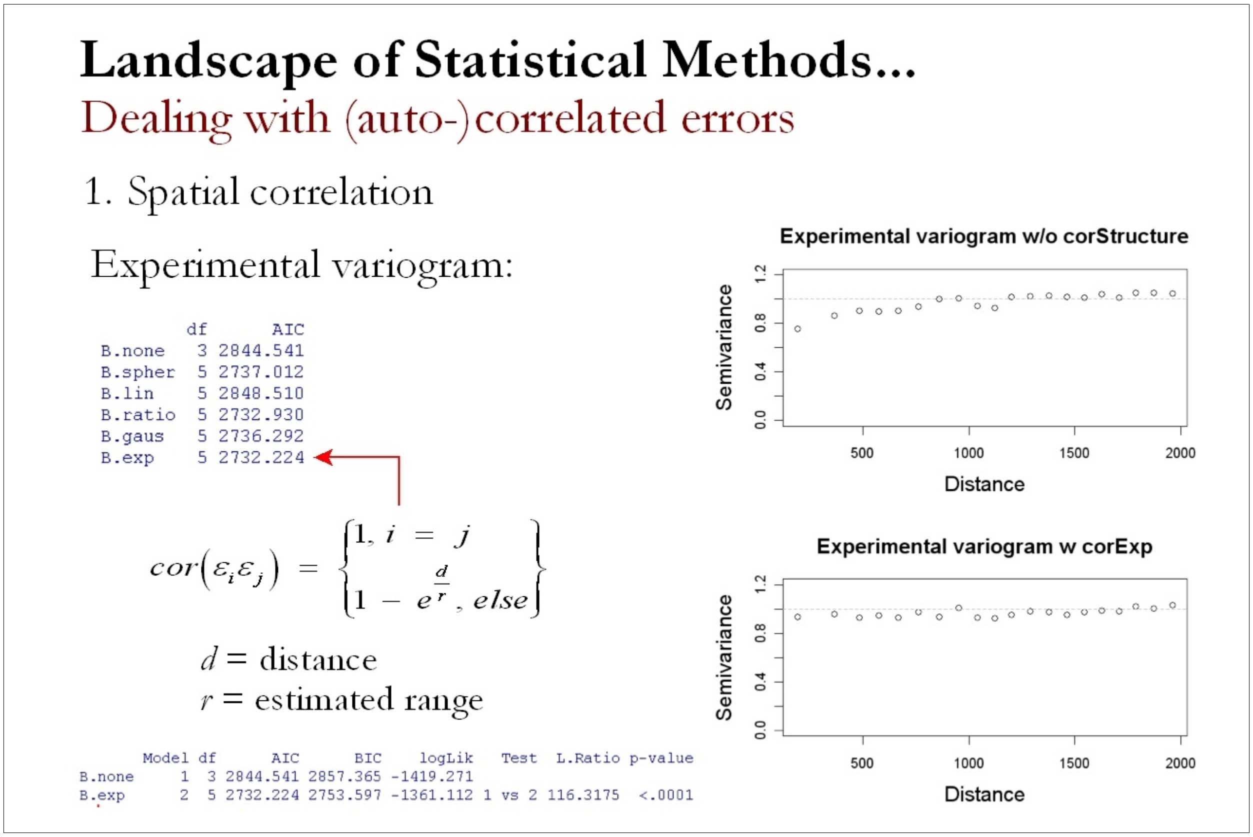
$$cor(\varepsilon_i \varepsilon_j) = \begin{Bmatrix} 1, \; i = j \\ \boxed{h(\varepsilon_i, \varepsilon_j, \rho),} \; else \end{Bmatrix}$$

Theoretical variograms

The question is how do we include a spatial residual correlation structure into the model. We need to do the same trick we used with the time series, but this time, based on the shape of the variogram, we need to choose a parameterization for the correlation function *h*(.). Not surprisingly, there are several common correlation structures that we can choose from, including:

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

The formulas for these can be a bit intimidating so we will not concern ourselves with them here, but the graphs of theoretical variograms using the various correlation structures gives a good picture of the kinds of patterns that each structure addresses. As before, selecting an adequate structure can be accomplished by trying a number of structures and using AIC to select the best. Note, these correlation structure can also be used for temporally correlated errors, just as the previous correlation structures we discussed can be used for spatially correlated errors.



# Landscape of Statistical Methods...

## Dealing with (auto-)correlated errors

1. Spatial 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
```

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

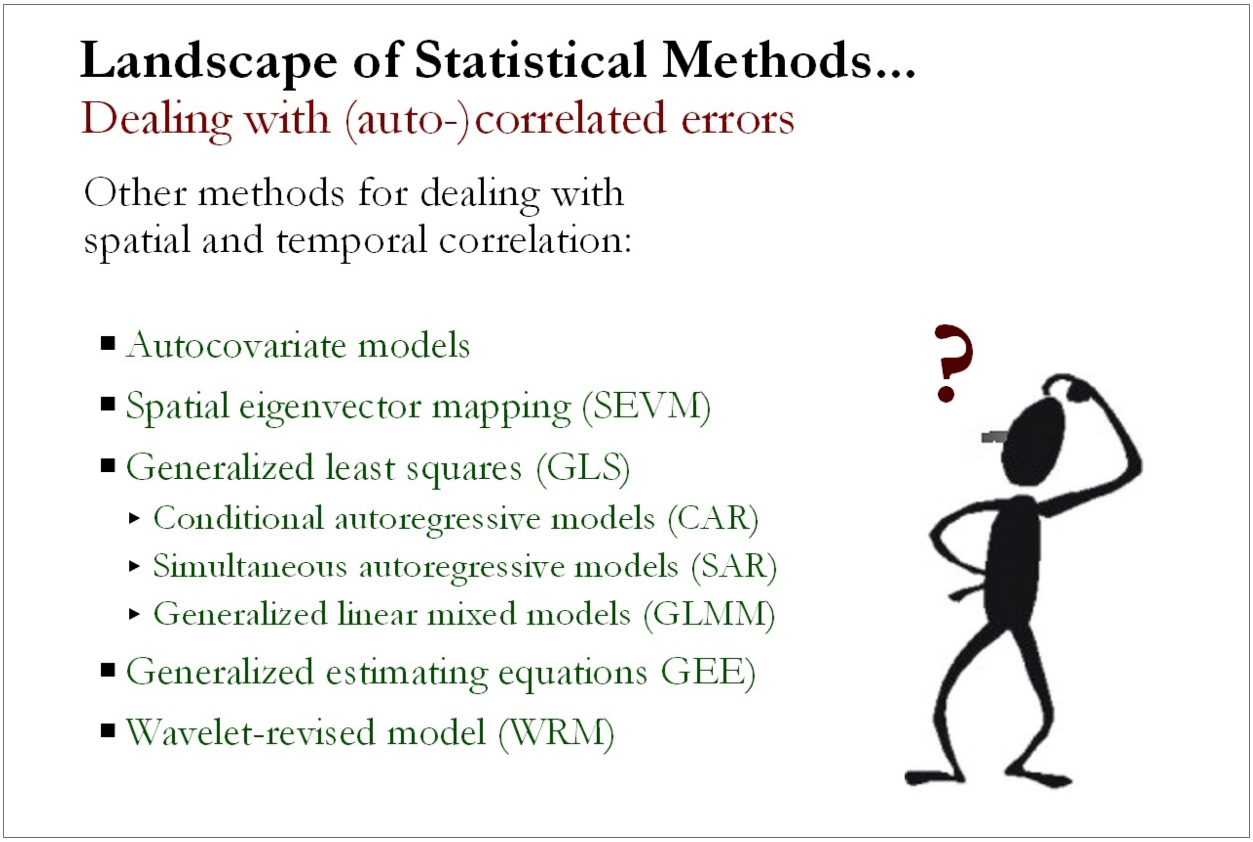
$d$ = distance

$r$ = 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
```

The experimental variogram for the residuals of the linear regression model applied on the boreality data reveals a clear spatial correlation up to a distance of about 1000m. This variogram assumes isotropy; i.e., the strength of the correlation is the same in all directions. We can verify this by making experimental variograms in multiple directions (not shown). In this case, it seems that isotropy is a reasonable assumption as the strength, and pattern, of the spatial correlation seems to be broadly the same in all four directions.

Both the bubble plot and experimental variogram indicate that there is a spatial correlation in the residuals and the multi-directional variograms seem to indicate that isotropy is a reasonable assumption. Given this, we can attempt to account for spatial correlation in the residuals in our model. Because there is no compelling reason to choose one correlation structure over another, we fit the model with each of the correlation structures and let AIC choose the best. In this case, the exponential correlation structure is the best, although the rational quadratic is a close second. The anova likelihood ratio test confirms that the model with the exponential correlation structure is significantly better than the original model that did not account for the correlated errors.



## Landscape of Statistical Methods...

Dealing with (auto-)correlated errors

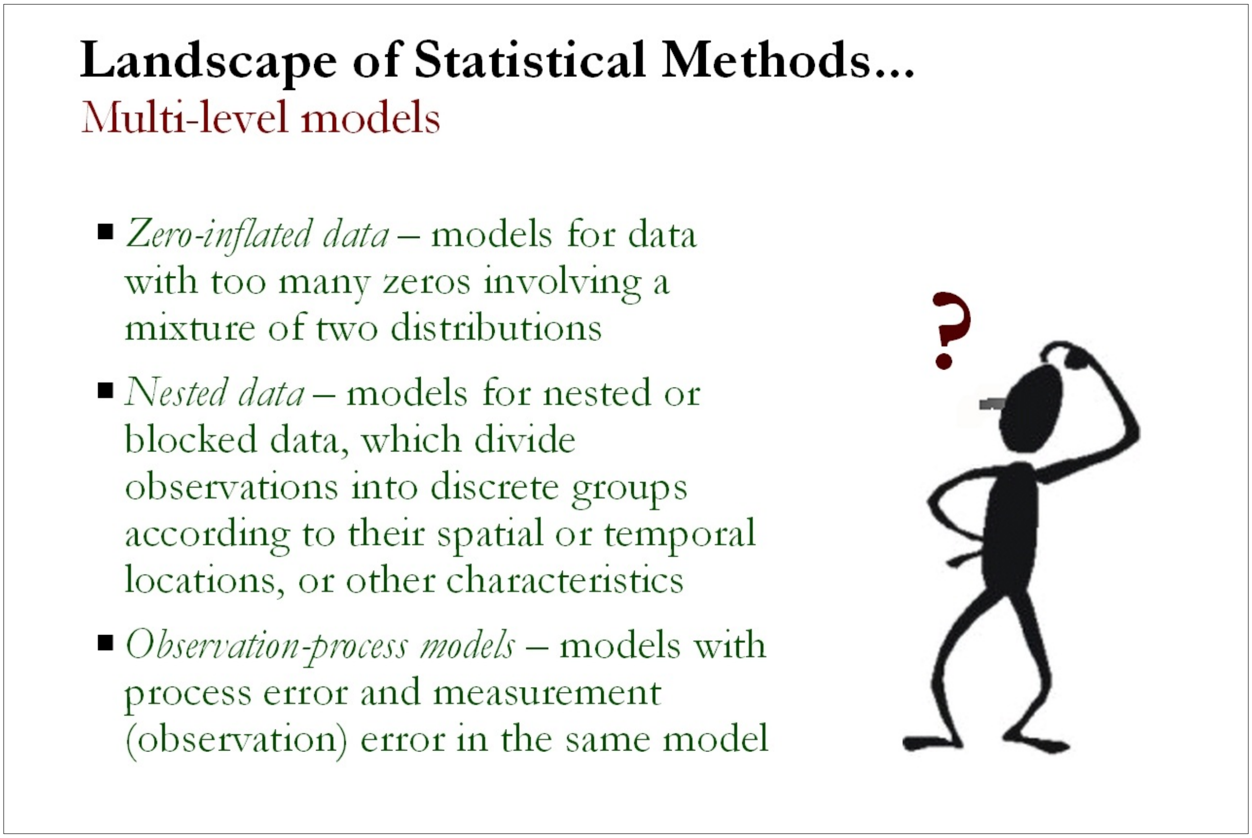
Other methods for dealing with spatial and temporal correlation:

- 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)

### 1.3 Correlation in complex models

If only it were always this easy. The bad news is that incorporating correlation structure into models is not always this easy. It is relatively simple with generalized least squares (GLS) and generalized nonlinear least squares (GNLS) models, as well as with linear mixed effects (LME) and nonlinear mixed effects (NLME) models (which we will discuss in the next section), because of the built-in correlation function, but all of these models are limited to *normally distributed* errors. Generalized additive models (GAMs) also can accommodate correlation structures, but it is unclear whether they can handle correlated errors for families other than the guassian distribution. Generalized estimating equations (GEEs) allow for correlated errors (and random effects) in Poisson and binomial GLMs and GAMs. For more complicated models that cannot be analyzed using one of these approaches and especially for data with nonnormal responses, such presence/absence, proportion and count data, it is going to be considerably more complex to account for autocorrelation in the model and will require delving into one or more of the other more advanced procedures. Dormann et al. (2007) provide an excellent review of available methods and Carl et al. (2008) provide an overwiew of the most promising newest method based on wavelet analysis.

- Dormann CF, et al. 2007. Methods to account for spatial autocorrelation in the analysis of species distributional data: a review. Ecography 30: 609-628.
- Carl G, Dormann CF, and Kühn I. 2008. A wavelet-based method to remove spatial autocorrelation in the analysis of species distributional data. Web Ecol. 8: 22–29.



**Landscape of Statistical Methods…**

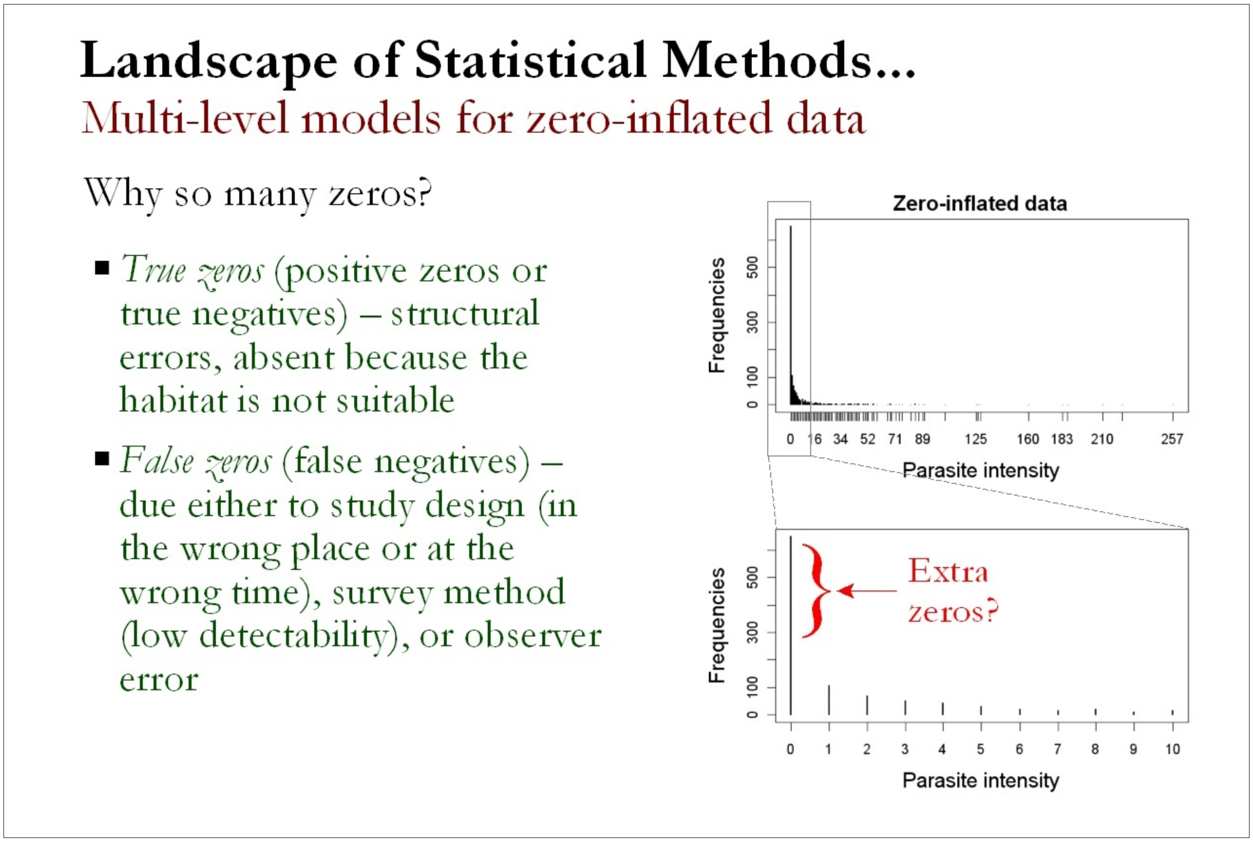
Multi-level models

- *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

## 2. Multi-level models (mixed effects models)

Thus far, we have been discussing models that incorporate only one type of variability in the model. However, there are many situations when we would like to incorporate more than one type of variability in the same model; when we do so, these models are variously referred to as *mixed*, *multi-level*, *multi-stratum*, or *hierarchical* models. Although these terms may connote specific variations when used in a particular context, we will refer to all such models as multi-level models. There are many different situations that can invoke the use of multi-level models; consequently, the analysis of multi-level models is a vast, rapidly growing, and increasingly complex subject. We will limit our consideration to three common applications:

- *Zero-inflation/altered models* – dealing with a mixture of two distributions to account for the inflated number of zeros in a data set, a situation quite common with count data in which there may be two different processes acting to produce zeros resulting in an inflated number of zeros.
- *Nested data* – dealing with nested or blocked data, which divide observations into discrete groups according to their spatial/temporal locations, or other characteristics. These models are typically referred to as mixed effects models because they distinguish between fixed and random effects, where the nesting or grouping variable (e.g., site or individual) is considered a random effect.
- *Observation-process models* – dealing with process error and observation error in the same model, a situation quite common in plant and animal studies with imperfect detection of individuals.



# Landscape of Statistical Methods…

## Multi-level models for zero-inflated data

Why so many zeros?

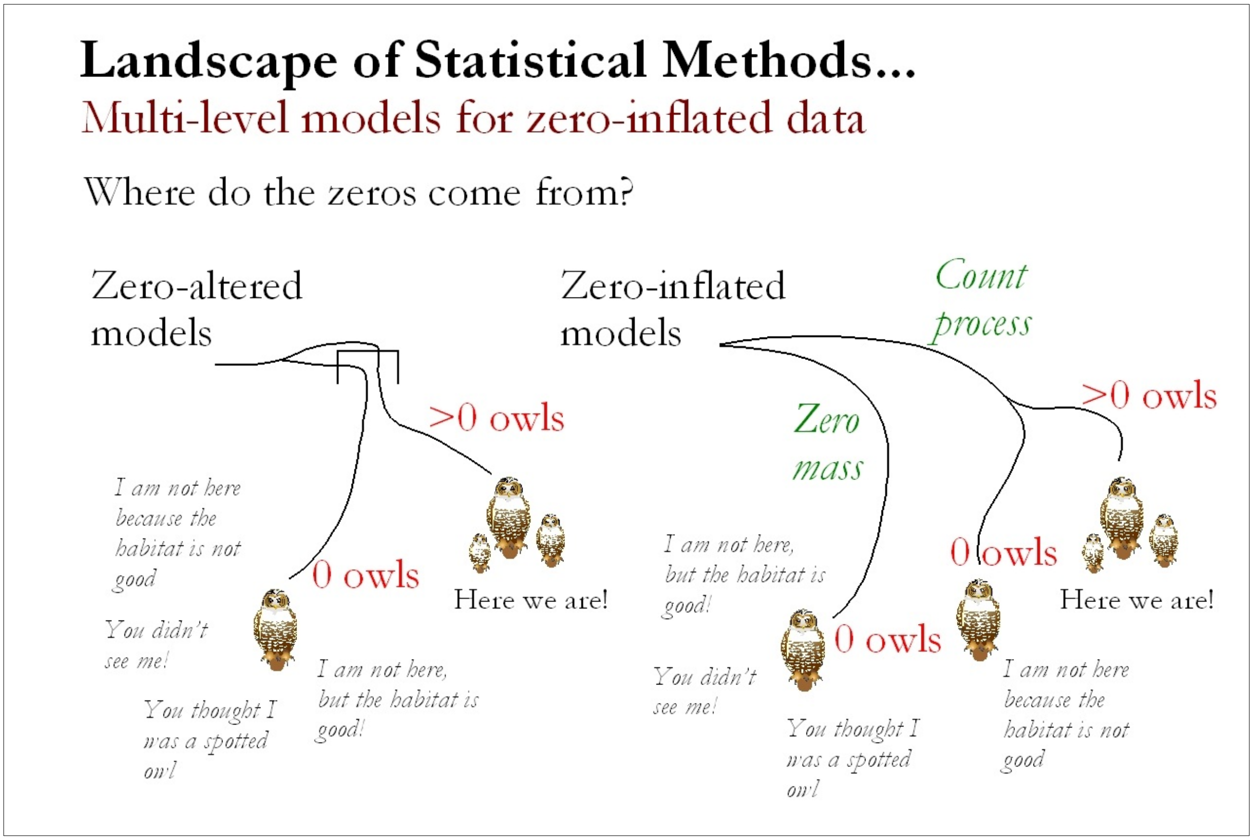
- *True zeros* (positive zeros or true negatives) – structural errors, absent because the habitat is not suitable
- *False zeros* (false negatives) – due either to study design (in the wrong place or at the wrong time), survey method (low detectability), or observer error

### 2.1 Mixture and two-part models for zero-inflated data

Our basic statistical model assumes that the errors are generated by a single process, or at least that we can adequately account for the errors in our model using a single error distribution, in addition to other ususal assumptions of independence and homogeneity. However, in ecological data sets we often encounter count data that has too many zeros – a problem called zero inflation – presumably arising from two different ecological processes. Zero-inflated data occur when there are more zero's than would be expected for the usual Poisson or negative binomial distribution. Ignoring zero inflation can have two consequences; firstly, the estimated parameters and their standard errors may be biased, and secondly, the excessive number of zeros can cause overdispersion (too much variance).

What is the source of the zeros?

- True zeros (positive zeros or true negatives) – these zeros reflect structural errors, wherein the organism is absent because the habitat is not suitable.
- False zeros (false negatives) – these zeros reflect false zeros due either to study design (surveying in the wrong place or at the wrong time), survey method (ineffective at detecting the organism when it is present), or observer error (failure to detect the organism when it is present). In a perfect world, we would not have false zeros.



## Landscape of Statistical Methods…

### Multi-level models for zero-inflated data

Where do the zeros come from?

Zero-altered models

>0 owls

Here we are!

0 owls

*I am not here because the habitat is not good*

*You didn't see me!*

*You thought I was a spotted owl*

*I am not here, but the habitat is good!*

Zero-inflated models

*Count process*

>0 owls

Here we are!

0 owls

*I am not here because the habitat is not good*

*Zero mass*

0 owls

*I am not here, but the habitat is good!*

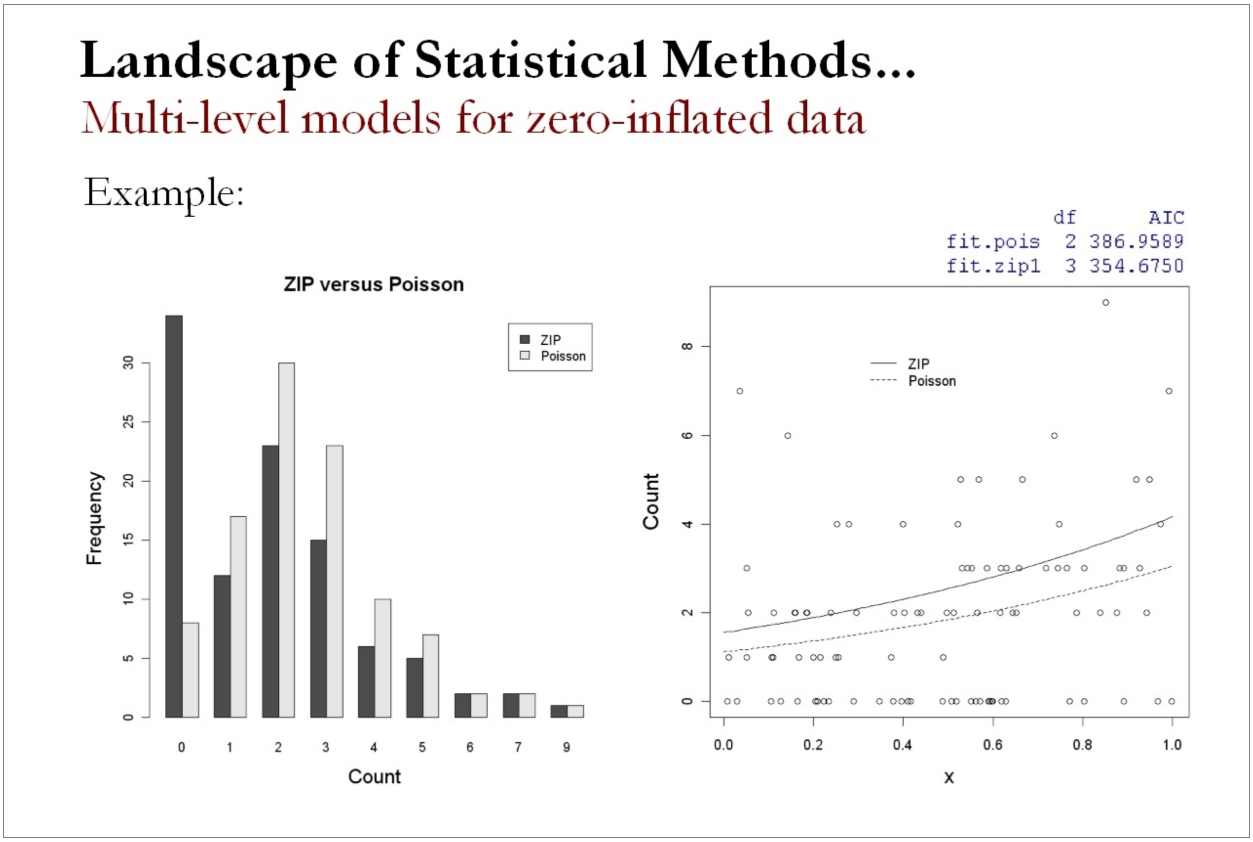
*You didn't see me!*

*You thought I was a spotted owl*

There are two types of models for handling zero-inflation: 1) zero-inflated mixture models, and 2) zero-altered two-part models. The difference is in how they deal with the different types of zeros.

*Zero-altered models* (ZAP/ZANB).–Zero-altered (ZA) models involving Poisson errors (P) and negative binomial errors (NB) work in two stages. In the first stage, the data are considered as zeros and non-zeros and a binomial model is used to model the probability that a zero value is observed. It is possible to use covariates in this model, but an intercept-only model is also an option. In the second stage, the non-zero observations are modeled with a *truncated* Poisson (ZAP) or *truncated* negative binomial (ZANB) model, and a (potentially different) set of covariates can be used. Because the distributions are zero-truncated, they cannot produce zeros. These zero-altered models are also referred to as "hurdle" models because whatever the mechanism that is causing the absence of the organism, it has to cross a hurdle before values become non-zero. The important point is that the model does not discriminate between the different types of zeros - true and false ones.

*Zero-inflated models* (ZIP/ZINB).–Zero-inflated (ZI) models involving Poisson errors (P) and negative binomial errors (NB) work in a single stages, but involve the mixture of two distributions. They are called mixture models because the zeros are modeled as coming from two different processes: the binomial process and the count process. As with the hurdle models, a binomial model is used to model the probability of measuring a zero and covariates can be used in the this model. The count process is modeled by a Poisson (ZIP) or negative binomial (ZINB) model. The fundamental difference with the hurdle models is that the count process can produce zeros – it is not truncated.



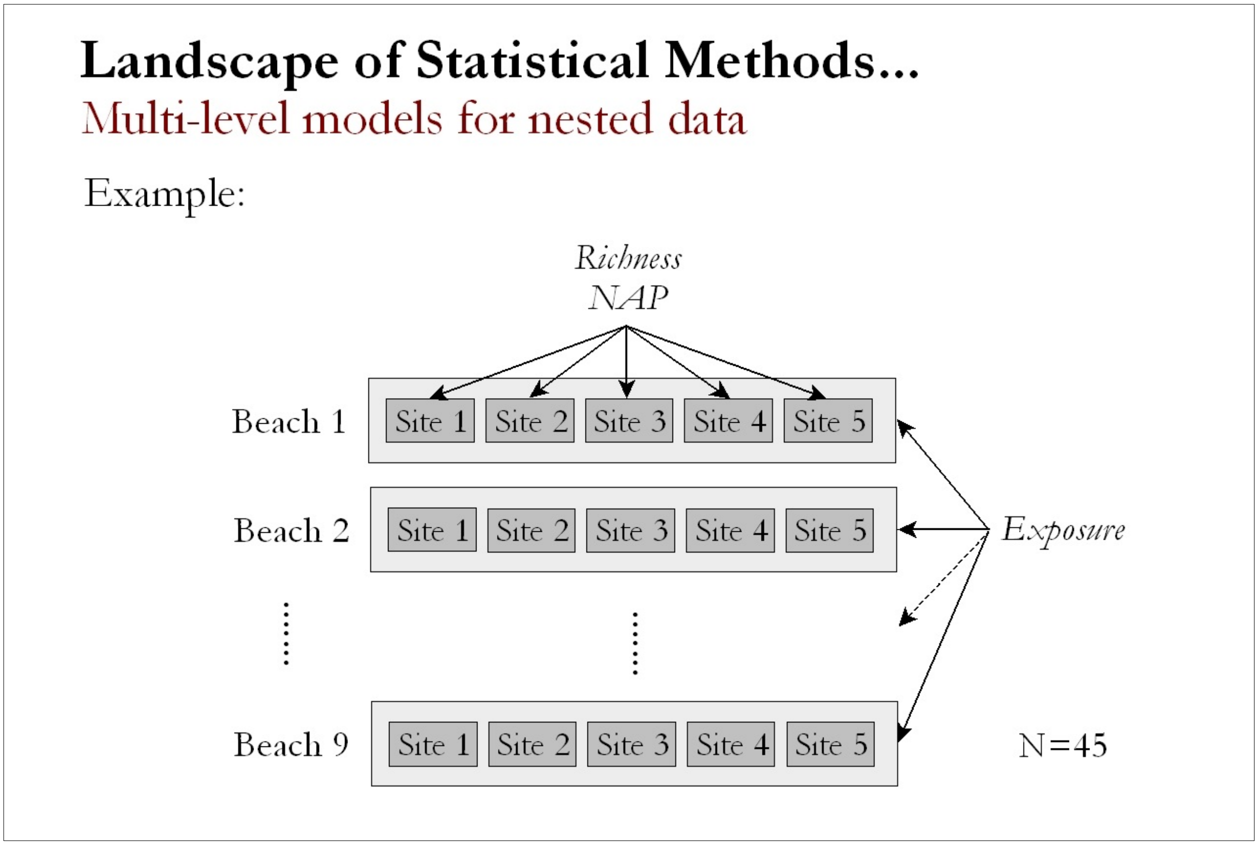
# Landscape of Statistical Methods…

## Multi-level models for zero-inflated data

Example:

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

A simple example illustrates how important it can be to account for zero inflation if it occurs. In this hypothetical example we are interested in the relationship between the count of individuals and an environmental variable measured at 100 sites. The histogram depicts the frequency of counts and shows the distribution of counts produced by a Poisson process (gray bars) alone and the distribution inflated by a separate binomial process that produces extra zeros. In addition, the scatter plot reveals that the mean (and variance) of the Poisson count process, denoted by the parameter lambda, increases with increasing values of x. If we fit the zero-inflated data (black bars in the histogram) with a single-level Poisson regression using generalized linear modeling (GLM) procedures, we get the fitted line depicted by the dotted line. If instead we fit the same data with a zero-inflated Poisson model, we get the fitted line depicted by the solid line. The fitted models are significantly different, which is confirmed by a likelihood ratio test (not shown), and the model AIC's are substantially different, indicating that there is no weight of evidence for the Poisson GLM.



# Landscape of Statistical Methods...

## Multi-level models for nested data

Example:

*Richness*
*NAP*

Beach 1 Site 1 Site 2 Site 3 Site 4 Site 5

Beach 2 Site 1 Site 2 Site 3 Site 4 Site 5

Beach 9 Site 1 Site 2 Site 3 Site 4 Site 5

*Exposure*

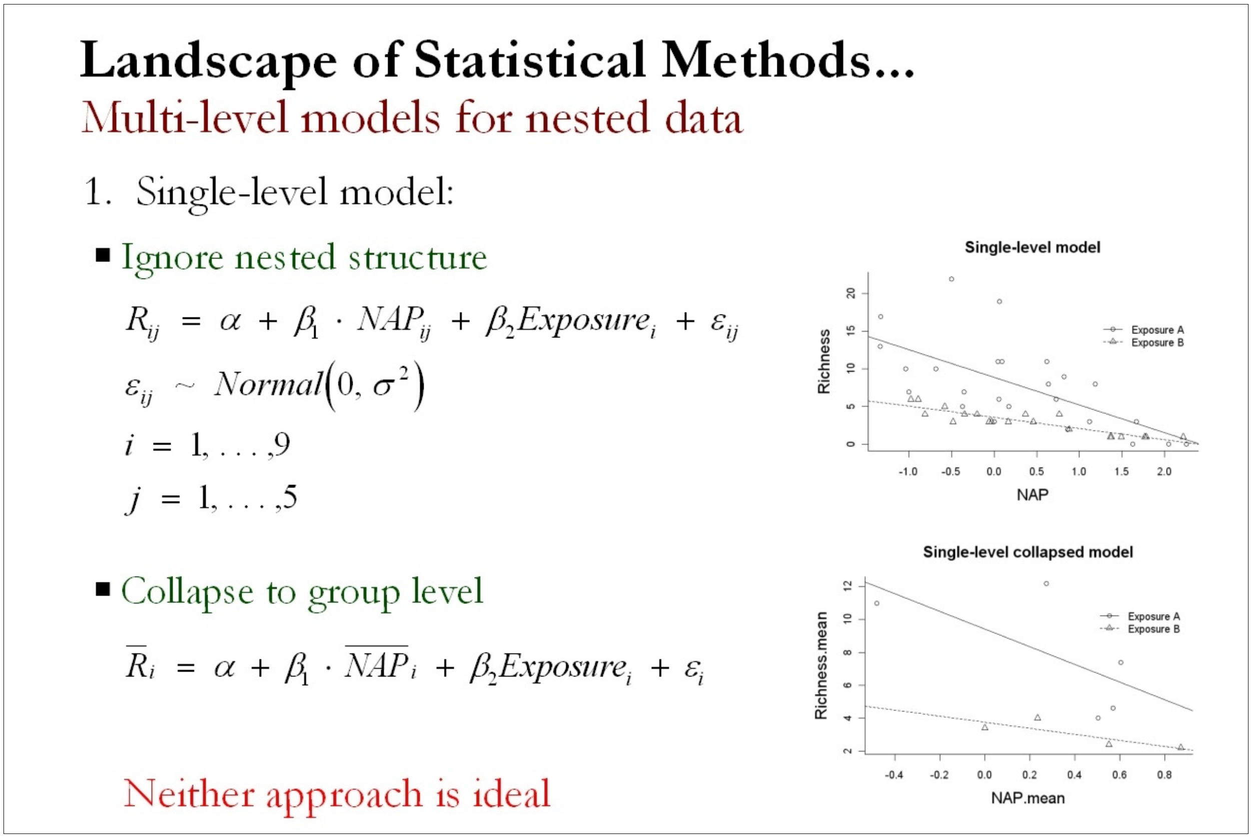
N=45

### 2.2 Nested data

The most common multi-level model involves nested or blocked data, which is best understand via an example. This example is taken from Zur et al. (2009, chapter 5) and involves marine benthic data from nine inter-tidal areas along the Dutch coast. In each inter-tidal area, denoted by 'beach', five samples were taken at different sites, and the species richness of macro-fauna was measured, denoted by 'R', as was an environmental variable representing the height of the sampling station compared to mean tidal level, denoted by 'NAP', and another environmental variable representing an index of exposure (nominally scaled with two levels) for each beach. Note, species richness and NAP vary among sites, but exposure varies among beaches; sites are grouped by beach. The basic question for these data is whether there is a relationship between species richness, beach exposure, and NAP.

There are three basic ways to analyze this data:

1. *Single level model* – conventional linear or generalized linear model ignoring the nested structure of the data
2. *Two-stage model* – break the model into two stages; in the first stage, model the relationship between richness and NAP for each beach separately, and then in the second stage, model the relationship between the estimated regression coefficients from the first stage and exposure.
3. *Multi-level model* – combine the two-stage model above into a single integrated model, with advantages to be discussed below.



**Landscape of Statistical Methods…**

Multi-level models for nested data

1. 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

*Single level model*

A candidate single level model for this data (analogous to the random intercept model below) is:

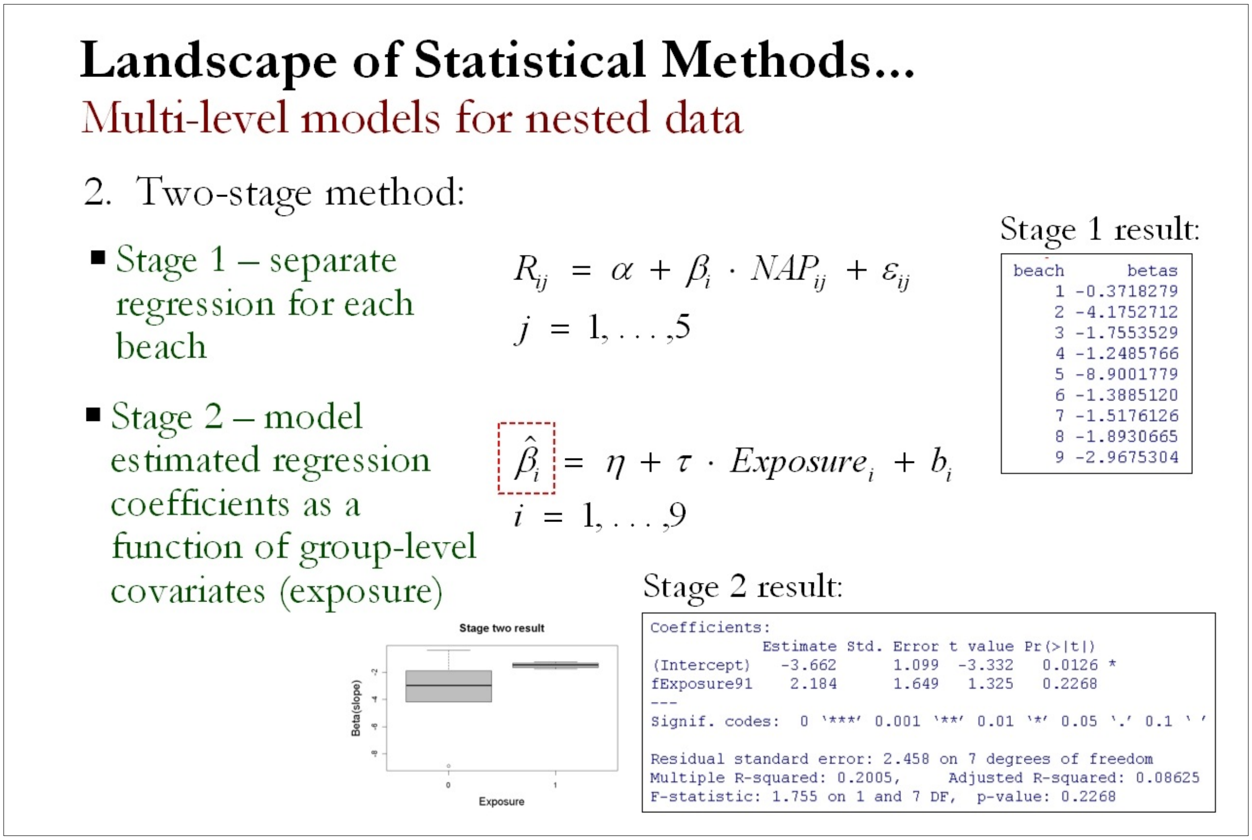
$$R_{ij} = \alpha + \beta_1 \cdot NAP_{ij} + \beta_2 Exposure_i + \varepsilon_{ij}$$
$$\varepsilon_{ij} \sim Normal\left(0, \sigma^2\right)$$

where $R_{ij}$ = species richness at site *j* on beach *i*; $NAP_{ij}$ = the corresponding *NAP* value, $Exposure_i$ = the exposure on beach *i*, and $\varepsilon_{ij}$ = the unexplained error. This linear regression model can be analyzed easily using ordinary least squares estimation under the usual assumptions. However, as we have five sites per beach, the richness values at these sites are likely to be more related to each other than to the richness values from sites on different beaches. This violates the assumption of independence among observations in the linear model. Ignoring this structure is a problem, unless we can convince ourselves that the between-group variation is unimportant both statistically and ecologically.

Another candidate single model for this data:

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

Note, hereafter, for brevity sake, we leave off the description of the error distribution. In this model, we collapse the data across sites (treated as subsamples) for each beach, by taking the average richness and average NAP, and then analyze the model as before but with only nine observations (beaches) instead of 45 sites. This approach will be disappointing if we are hoping to glean information about the within-group variance, but it is simple.



## Landscape of Statistical Methods…

### Multi-level models for nested data

2. 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 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 – 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 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
```

*Two-stage model*

In the two-stage approach, we separate the two levels of the model, site and beach, into separate models, with the first-stage result serving as the input to the second stage. Briefly, in the first stage, a linear regression model is applied to the data from one beach to assess the relationship between richness and NAP:

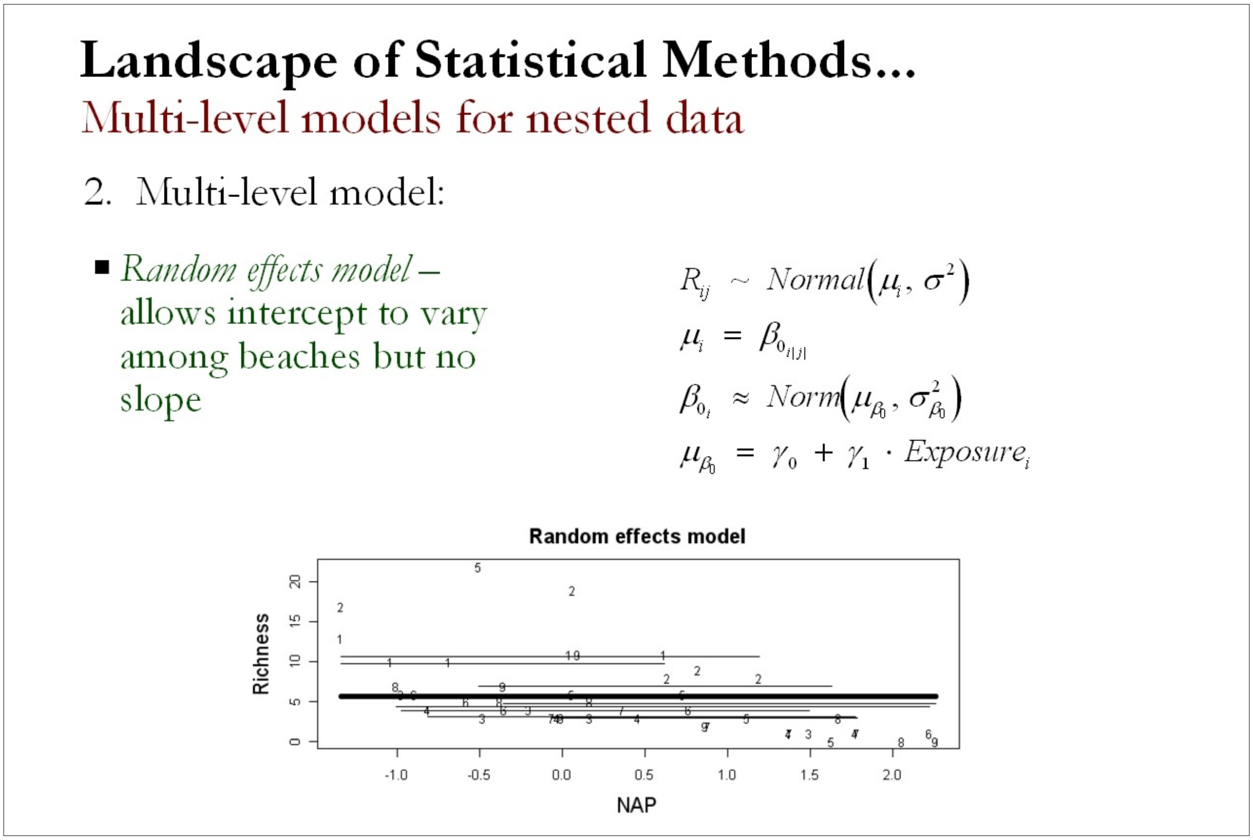
$$R_{ij} = \alpha + \beta_i \cdot NAP_{ij} + \varepsilon_{ij}$$
$$j = 1, \ldots, 5$$

We repeat this for each of the beaches separately, each time producing an estimate of the beta coefficients for the corresponding beach. In the second stage, the estimated regression coefficients are modeled as a function of exposure:

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

Note, this is just a one-way ANOVA. The response variable is the estimated slopes from stage 1, Exposure is the (nominal) explanatory variables, *tau* is the corresponding regression parameter (slope), *eta* is the intercept, and $b_i$ is random noise. It is common to assume that the residuals $b_i$ are normally distributed with mean 0 and variance $D$.

The second stage can be seen as an analysis of a summary statistic; in this case, it is the slope representing the strength of the relationship between species richness and NAP on a beach. The two-stage analysis has various disadvantages. Firstly, we summarize all the data from a beach with one parameter. Secondly, in the second step, we analyze regression parameters, not the observed data. Hence, we are not modeling the variable of interest directly. Finally, the number of observations used to calculate the summary statistic is not used in the second step. In this case, we had five observations for each beach. But if we had 5, 50, or 50,000 observations, we still end up with only one summary statistic.



# Landscape of Statistical Methods…

## Multi-level models for nested data

2. Multi-level model:

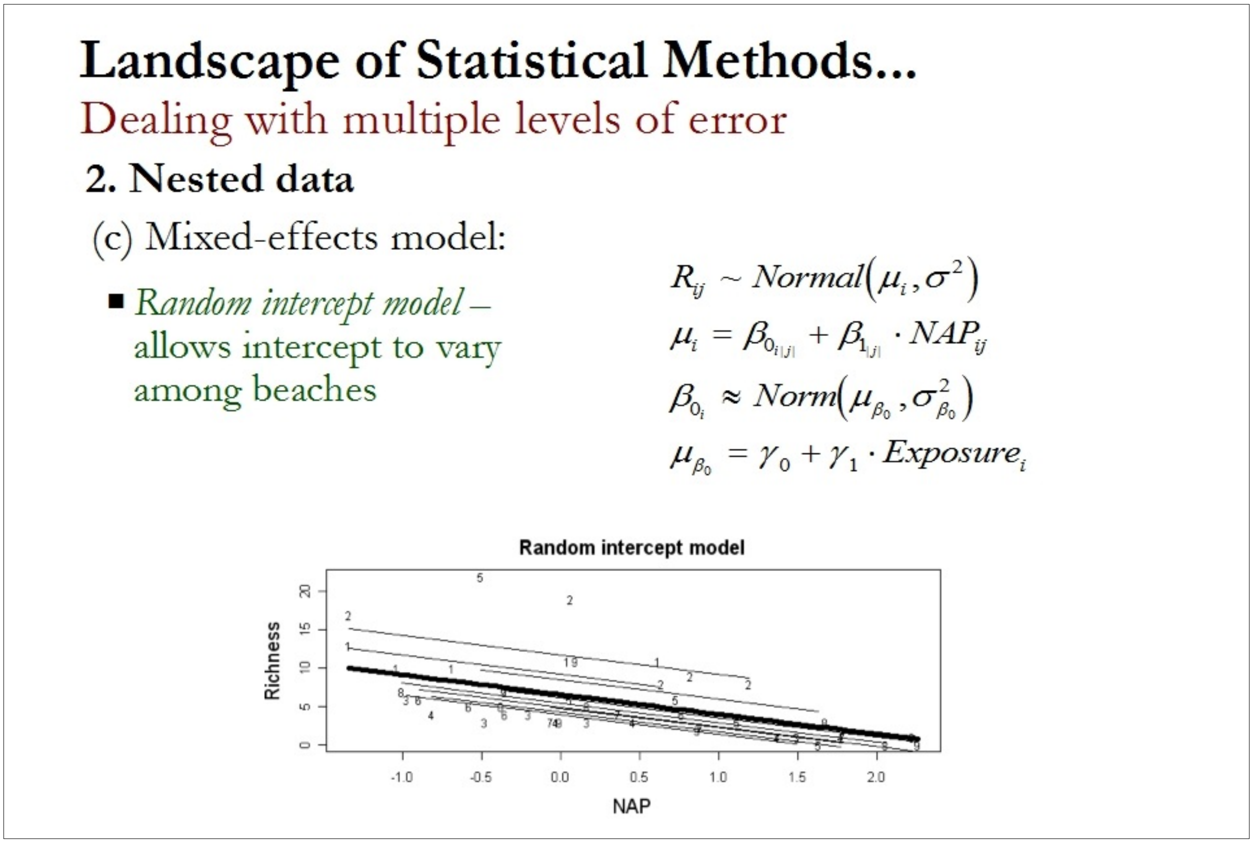
- *Random effects model* – allows intercept to vary among beaches but no slope

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

*Multi-level model*
The multi-level model approach combines both of the stages above into a single model. The details of the analysis go way beyond the scope of our survey, but briefly, there are at least three different models potentially of interest to us: 1) random effects model, 2) random intercept model, and 3) random intercept and slope model:

Random effects model.– the random effects model does not contain any coefficients for the richness-NAP relationship (i.e., no slope estimate(s)). Richness is modeled as an intercept plus a random term that is allowed to differ among beaches. The model is:

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



# Landscape of Statistical Methods…

## Dealing with multiple levels of error

### 2. Nested data

(c) Mixed-effects model:

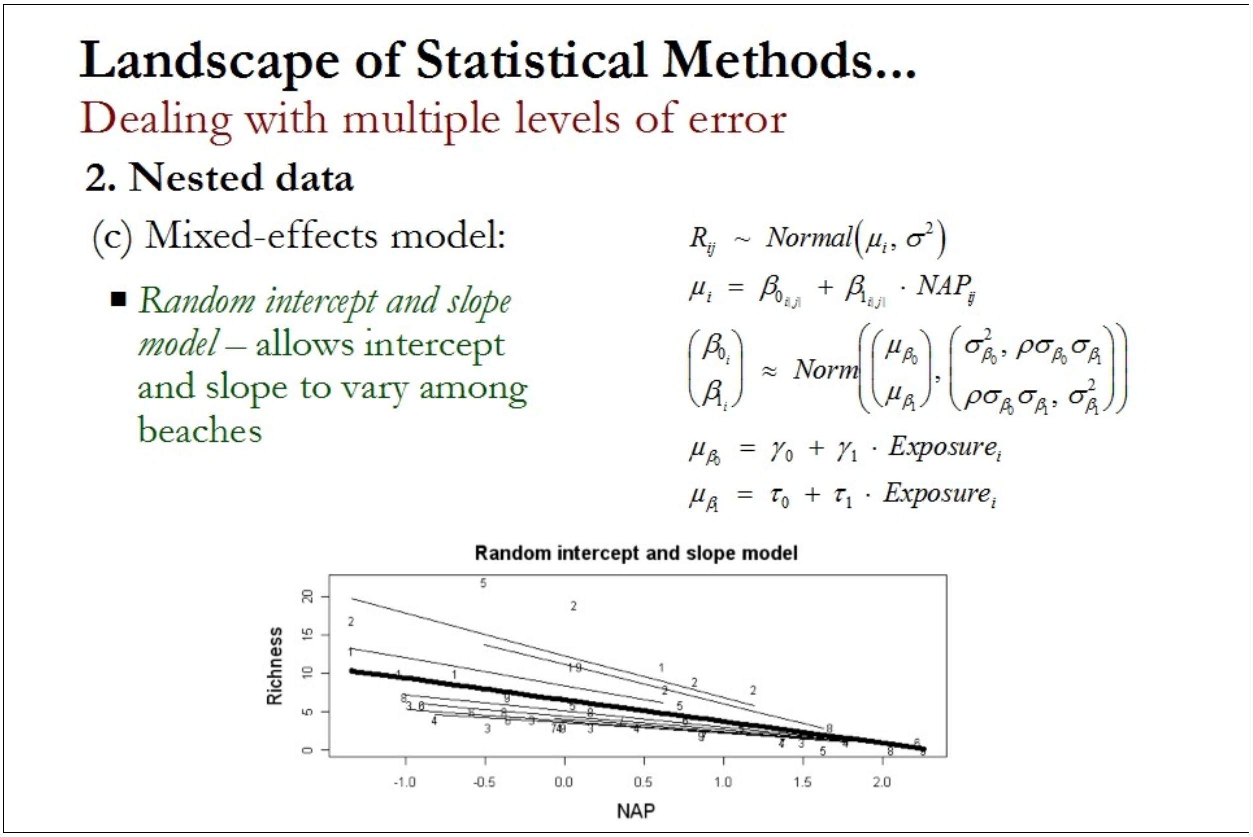
- *Random intercept model* – allows intercept to vary among beaches

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

Random intercept model

Random intercept model.–the random intercept model combines the two stages above into a single model and allows only the intercept to vary among beaches, but the slope of the richness-NAP relationship is not allowed to vary among beaches. This model is particularly relevant if we are most interested in the global relationship between richness and NAP. The model is:

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



# Landscape of Statistical Methods…

## Dealing with multiple levels of error

### 2. Nested data

(c) Mixed-effects model:

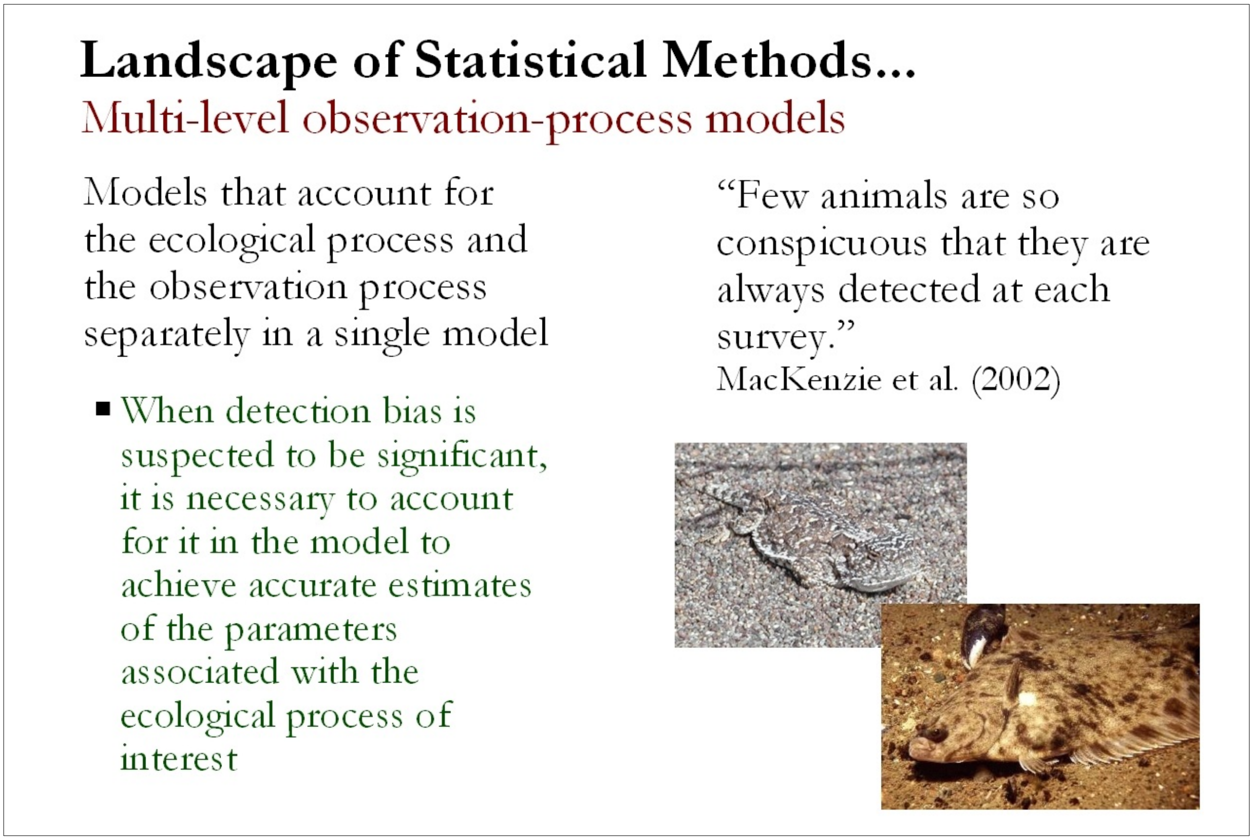
- *Random intercept and slope model* – allows intercept and slope to vary among beaches

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

Random intercept and slope model

Random intercept and slope model.–the random intercept and slope model combines the two stages above into a single model and allows both the intercept and slope coefficients (from stage 1) to vary among beaches. The model is:

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



**Landscape of Statistical Methods…**

Multi-level observation-process models

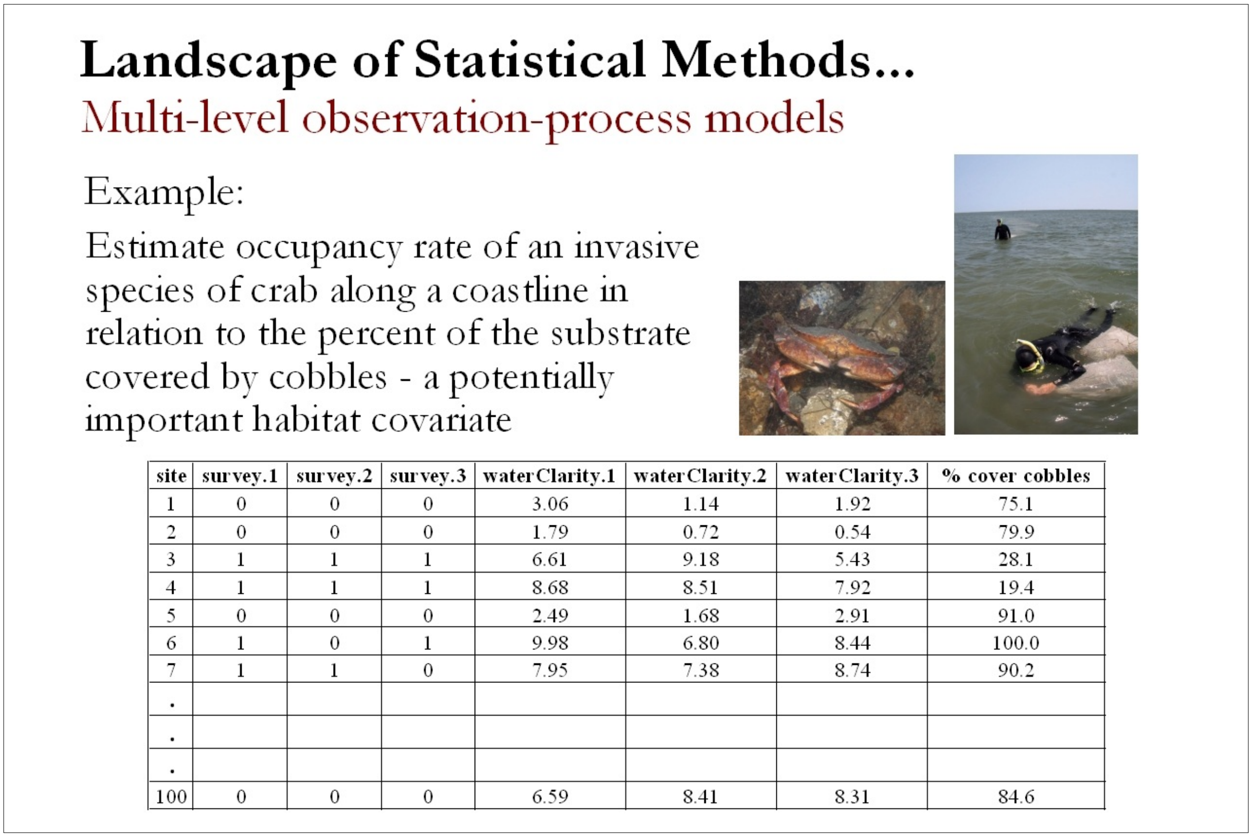
Models that account for the ecological process and the observation process separately in a single model

- 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

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

## 2.3 Observation-process models

In conventional (single-level) statistical models, the stochastic component of the model includes both process error (i.e., random error due to the ecological process) and measurement or observation error (i.e., error due to the imperfect measurement or observation process). In situations where the observation error is high, combining these two sources of error can lead to biased estimates of the true ecological relationships under study. Fortunately, we now have the tools to be able to model these sources of error separately. Models that explicitly separate the observation process from the ecological process in the same model are sometimes referred to as observation-process models. These models are more generally referred to as hierarchical models or simply multi-level models. Observation-process models are best understood via an example - a very simple example, as these models can get considerably more complex by adding additional covariates and random effects to the model.



## Landscape of Statistical Methods...

### Multi-level observation-process models

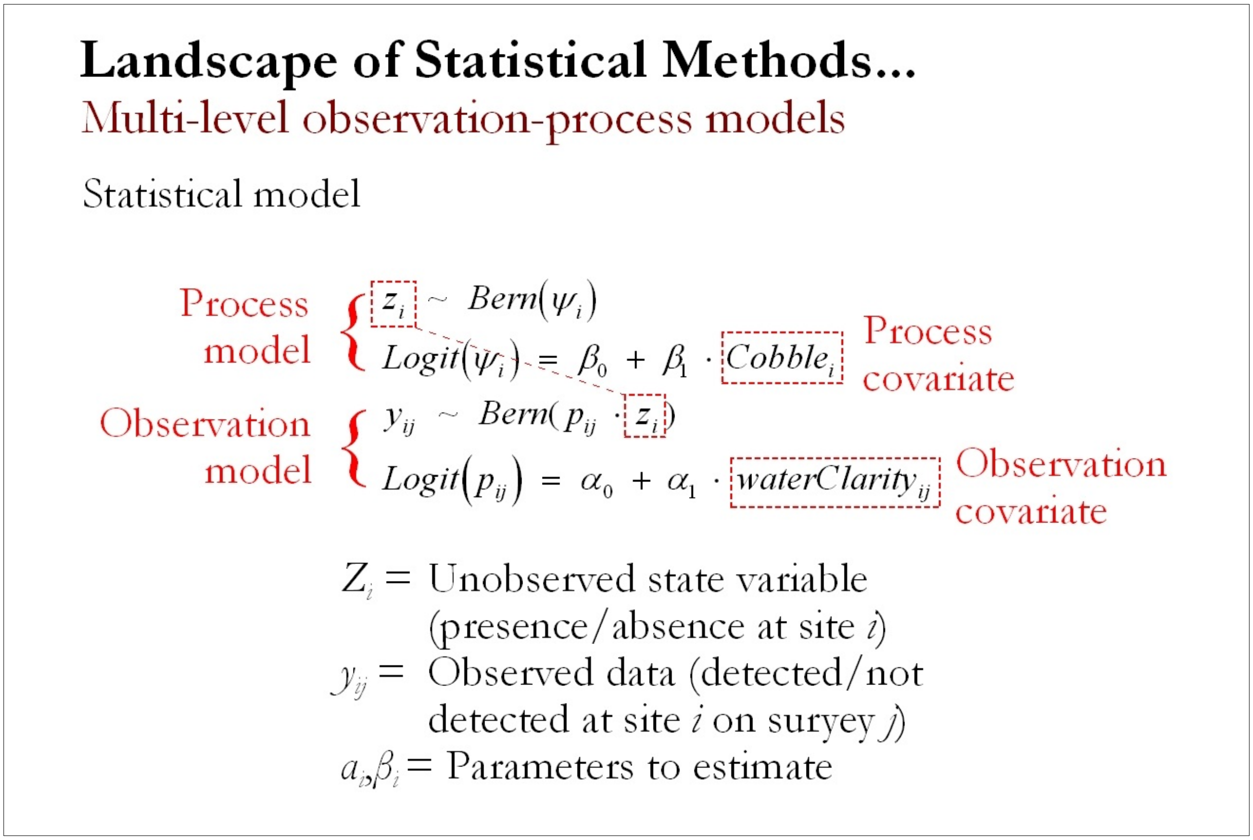
Example:

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 |

*Question, study design and data*

In this hypothetical study, let’s say we want to estimate the occupancy rate of an invasive species of crab along the coastline of Massachusetts and also assess habitat preferences of this species with respect to the percent of the substrate covered by cobbles, an *a priori* known habitat covariate of importance to this species. To accomplish this, we randomly selected 100 sampling locations along the Massachusetts coastline, where a sampling location consisted of a 50x50 m plot, and we surveyed each plot 3 times during the summer. During each survey, we recorded whether or not we observed the species. Each survey consisted of a 30 minute time-constrained snorkling survey of each plot. In addition to recording whether or not we observed the species during each survey at each plot, we also collected data related to the water clarity at each sampling plot during each survey, as we had reason to believe that our detectability of the species would decline with decreasing water clarity. We used a turbidity meter that allowed us to measure water clarity on a scale from 0 (lowest water clarity) to 10 (highest water clarity). Upon completion of all of the surveys we went back and measured the percent cover cobbles within each sample plot. Here is what the data looks like.



## Landscape of Statistical Methods…

Multi-level observation-process models

Statistical model

Process model

$$\begin{cases} \boxed{z_i} \sim Bern(\psi_i) \\ Logit(\psi_i) = \beta_0 + \beta_1 \cdot \boxed{Cobble_i} \end{cases}$$

Process covariate

Observation model

$$\begin{cases} y_{ij} \sim Bern(p_{ij} \cdot \boxed{z_i}) \\ Logit(p_{ij}) = \alpha_0 + \alpha_1 \cdot \boxed{waterClarity_{ij}} \end{cases}$$

Observation covariate

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

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

$a_i,\beta_i$ = Parameters to estimate

*Multi-level statistical model*

The data is indexed as follows:

- $y_{ij}$ = crab presence at $i^{th}$ sample location ($i$ = 1,...,100) during the $j^{th}$ survey ($j$ = 1,2,3).
- $waterClarity_{ij}$ = water clarity (1-10) at the $i^{th}$ sample location during the $j^{th}$ survey; this is a site- and time-specific covariate affecting detectability.
- $cobbles_i$ = percent cover of cobbles at the $i^{th}$ sample location; this is a site-specific covariate affecting occupancy (presense/absence).

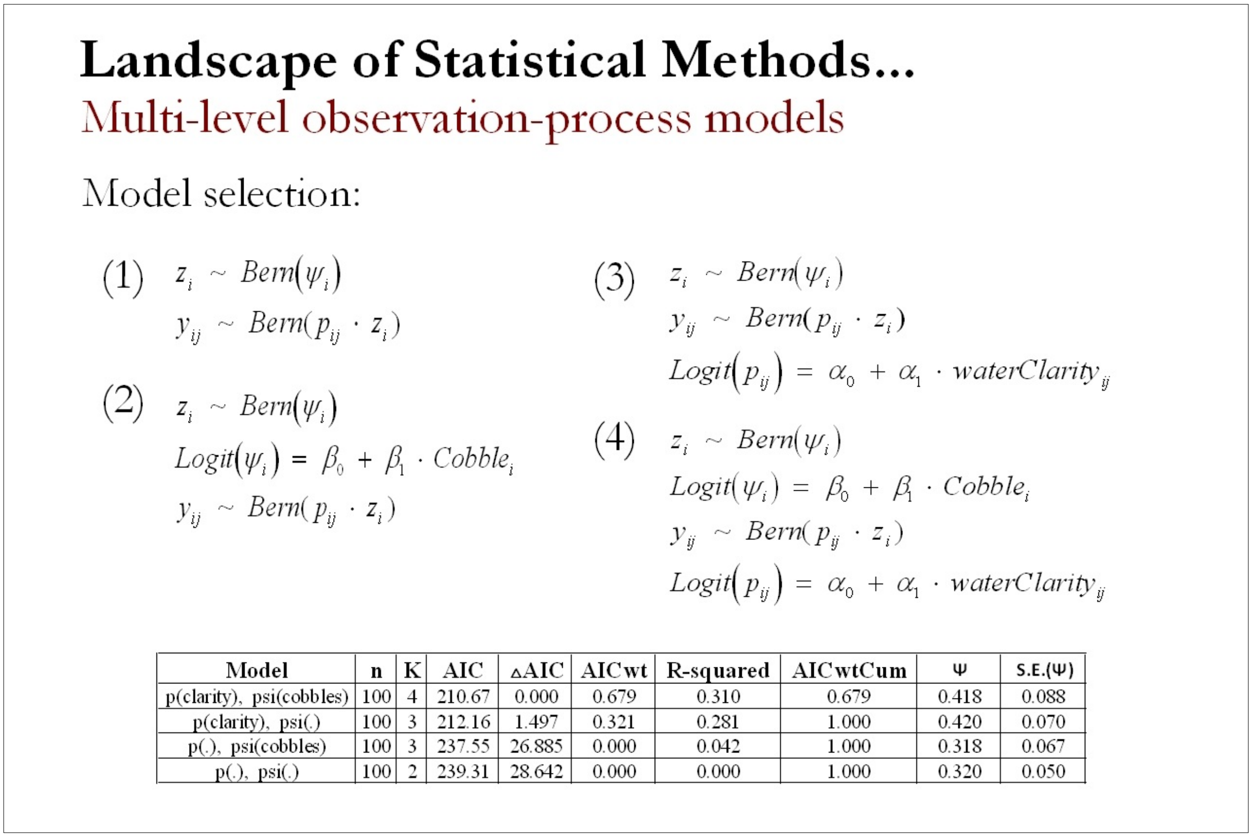
The full hierarchical model for this data set is as follows:

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

Process model: The state variable, $z$ (presence of the species), is distributed Bernoulli (equivalent to a binomial with size=1) with probability equal to $\psi$ *(psi)*. *Psi* is modeled as a logistic function of percent cover of cobbles at the sample location with parameters $b_0$ (intercept) and $b_1$ (slope).

Observation model: The observed data, *y* (detection of the species), is distributed Bernoulli with probability equal to *p***z*. Thus, if *z*=0 (species absent), the probability of detection = 0. If *z*=1 (species present), the probability of detection is equal to *p*, which is modeled as a logistic function of water clarity at the plot-survey level with parameters $a_0$ (intercept) and $a_1$ (slope).

This two-stage hierarchical model composed of a Bernoulli observation model and a Bernoulli process model is a fully estimable model, since there is replication within site (i.e., 3 survey occasions). With spatially and temporally replicate surveys for species' presence, a very simple hierarchical construction permits a formal rendering of the model into constituent observation and state process components. That is, species presence is decomposed into two components, one for the unobserved (latent) state variable occupancy *z* and another for the observed state (i.e., the data) *y* that has obvious interpretations in the context of the ecological sampling and inference problems. The hierarchical model has several advantages over other model formulations, but we will not discuss them here as part of our survey. Suffice it to say that when it is logical to think of the data as being comprised of two (or more) processes, one deriving from the ecological process and another deriving from the measurement or observation process, then a hierarchical model formulation is not only intuitive, but leads to better understanding of the constituent processes.



## Landscape of Statistical Methods…

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

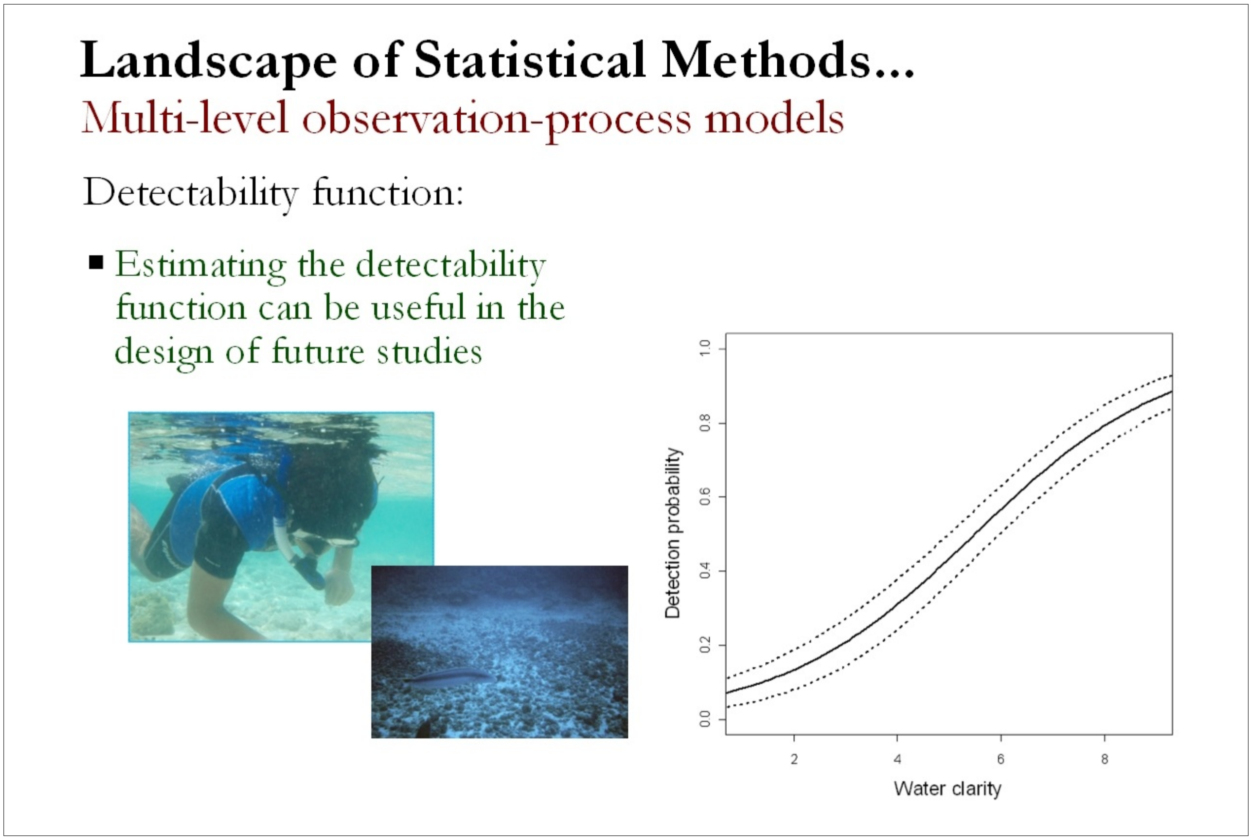
| Model | n | K | AIC | ΔAIC | AICwt | R-squared | AICwtCum | Ψ | S.E.(Ψ) |
|---|---|---|---|---|---|---|---|---|---|
| p(clarity), psi(cobbles) | 100 | 4 | 210.67 | 0.000 | 0.679 | 0.310 | 0.679 | 0.418 | 0.088 |
| p(clarity), psi(.) | 100 | 3 | 212.16 | 1.497 | 0.321 | 0.281 | 1.000 | 0.420 | 0.070 |
| 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 |

*Model selection*

Our first task was to estimate the occupancy rate of this species in our sample plots. To estimate this we can select several potential models to assess via a model selection procedure (e.g., AIC model selection). We have four candidate models to consider:

1. Null model – no influence of water clarity on detection probability and no influence of percent cover cobbles on occupancy [p(.), psi(.)];
2. No influence of water clarity on detection probability with an influence of percent cover cobbles on occupancy [p(.), psi(cobbles)];
3. Influence of water clarity on detection probability and no influence of percent cover cobbles on occupancy [p(clarity), psi(.)];
4. Influence of water clarity on detection probability and an influence of percent cover cobbles on occupancy [p(clarity), psi(cobbles)].

The AIC results suggest that two of these models contain considerable support as the best model. The model with the most support is the global model (i.e., the model that includes water clarity as a detection covariate and percent cover cobbles as an occupancy covariate), and the model with the second most and a considerable amount of support based on the ΔAIC value (i.e., ΔAIC< 2) is the model that includes water clarity as a detection covariate with no effect of percent cover cobbles on occupancy rate. The estimated occupancy rate using either of these models gives very similar results, with an estimated occupancy rate between 0.418 and 0.420, considerably higher that what was obtained from the null model (i.e., 0.320).



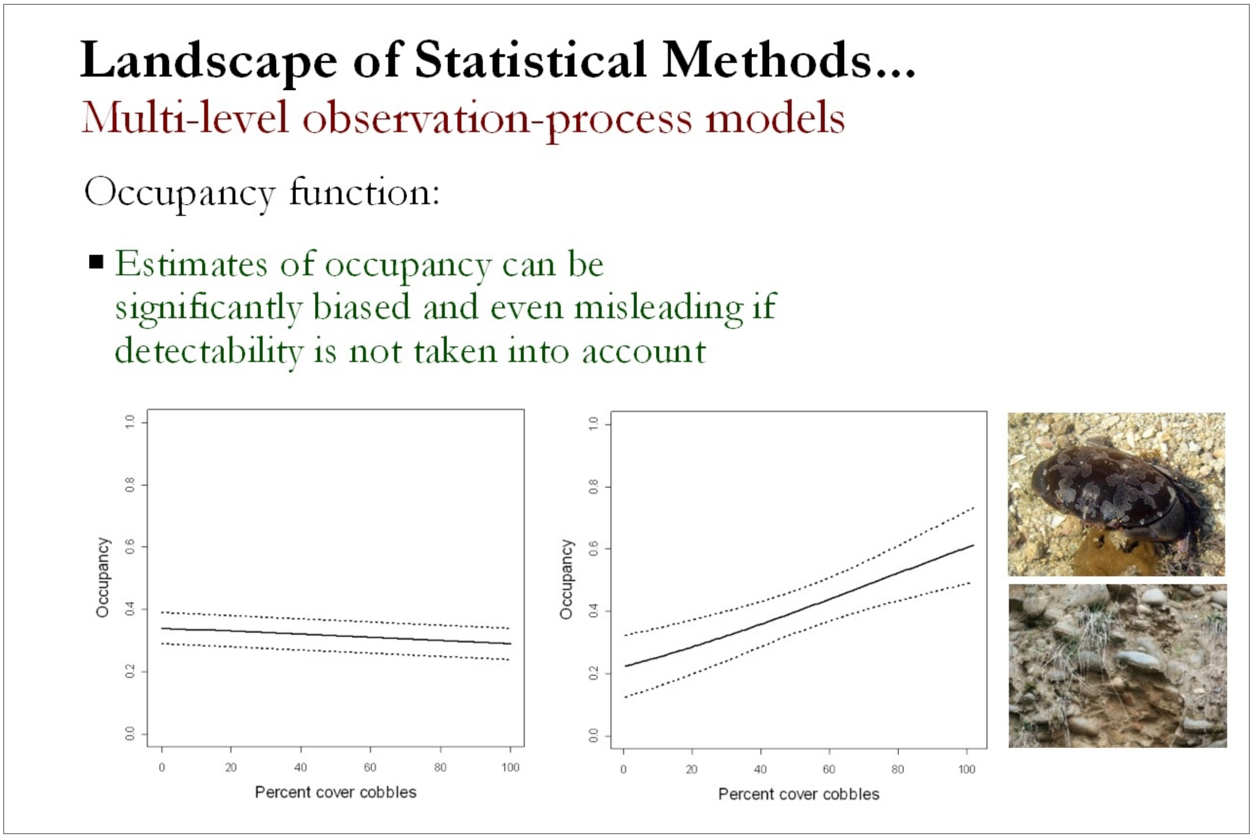
## Landscape of Statistical Methods…

### Multi-level observation-process models

Detectability function:

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

*Detectability function*
It is apparent from the two top models selected from the AIC model selection procedure that water clarity has a considerable effect on the detection probability of the crab species in our surveys. In addition to assessing this effect for the purposes of correcting our estimates of occupancy from our 100 sample plots, we also may want to assess this relationship in order to refine future inventory and monitoring surveys for this species such that surveys are only conducted when detection probability is relatively high for this species. Fortunately, being good ecologists, we designed our study so that we can estimate detection probability and account for it in the construction of our statistical model and in our estimates of occupancy (by employing multiple surveys per site and measuring data on a survey covariate, i.e. water clarity, that likely influences detection probability). With our data in hand, we are also able to plot detection probability as a function of water clarity. We see here that there is a strong, positive curvilinear relationship between water clarity and detection probability of the invasive crab species.



**Landscape of Statistical Methods…**

Multi-level observation-process models

Occupancy function:

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

*Occupancy function*

Our second task was to estimate the occupancy rate of this species in relation to the percent cover of cobbles within 50x50 m plots. If we had not accounted for the effect of water clarity on detectability of crabs in our survey (i.e., the [p(.), psi(cobbles)] model), we would have inferred that as percent cover cobbles increases, the occupancy rate of crabs decreases slightly. However, from our AIC model selection results (and from the previous detection probability plot) it is evident that we should account/correct for the effect of water clarity in the construction of our statistical model relating percent cover cobbles to occupancy rate. When we do account for the effect of water clarity on detectability we find that there is actually a positive (and considerably strong, based on the observed slope) relationship between percent cover cobbles and occupancy rate of the invasive crab species.