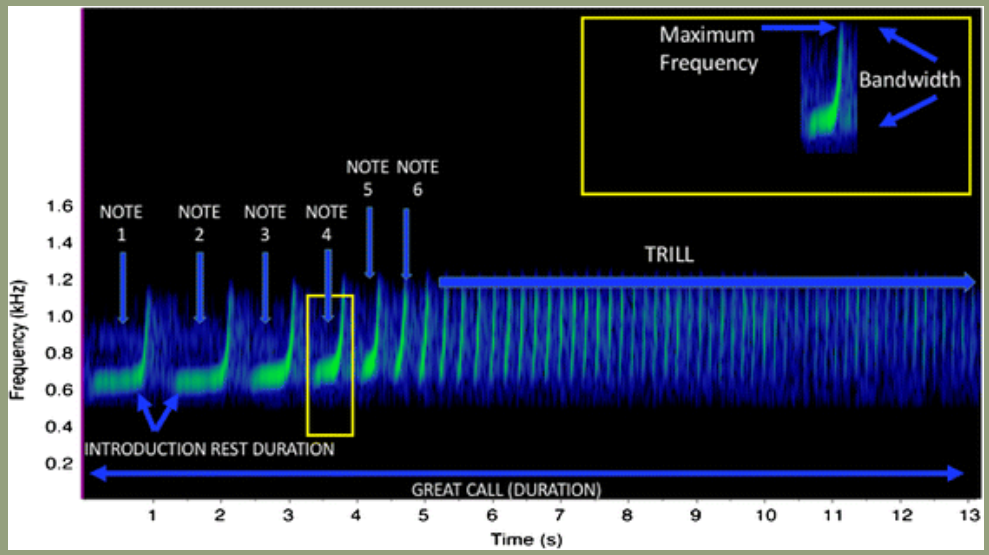
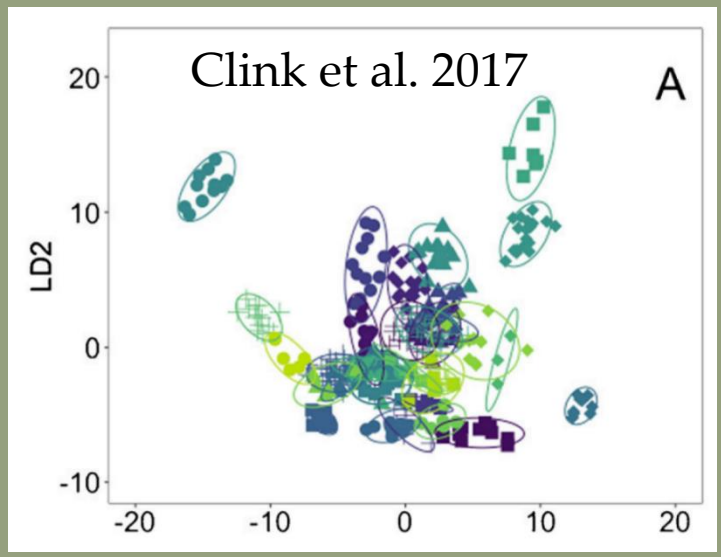
Coupled Classification Occupancy Models

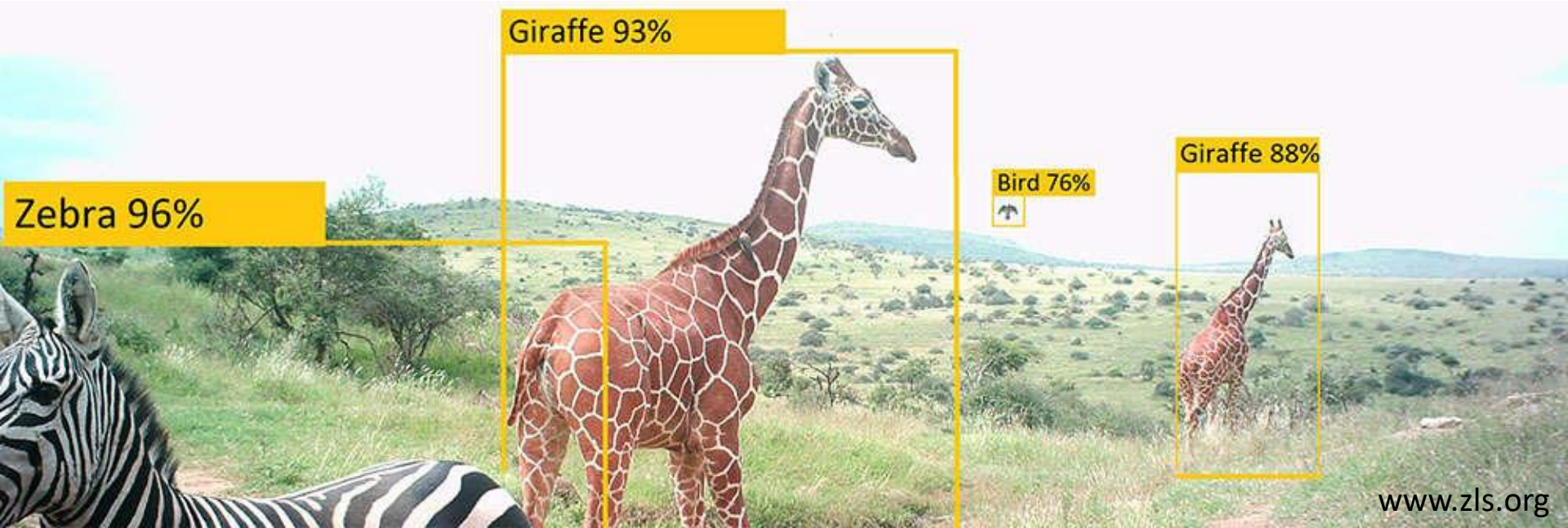




USGS

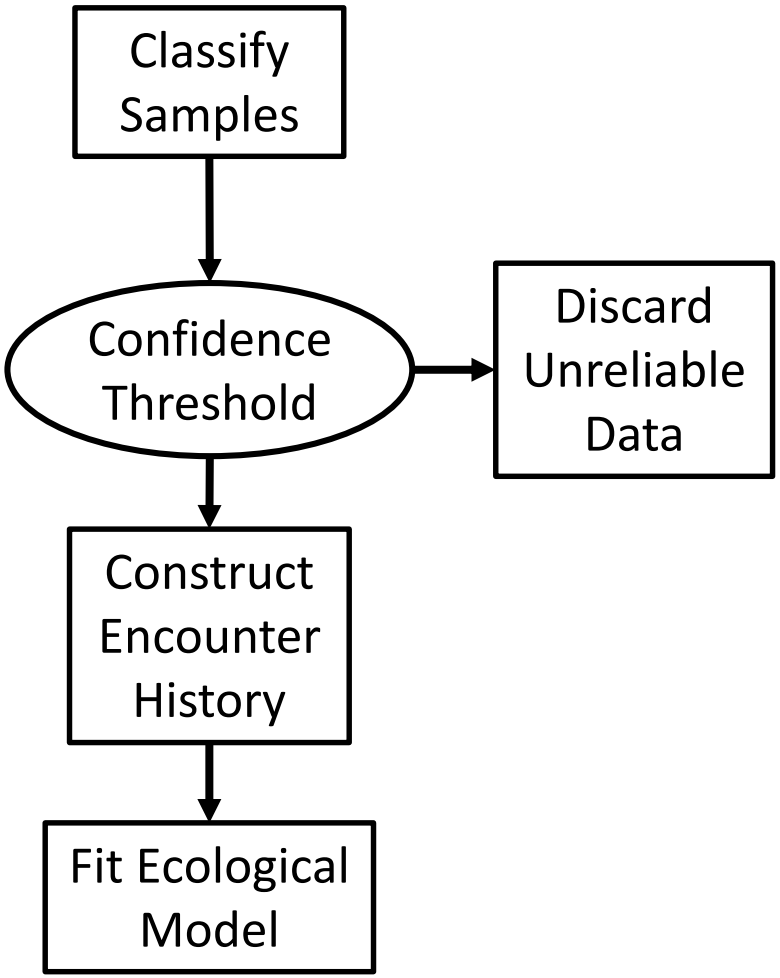
Coupled Classification

- Many ecological models rely on perfect classification of samples
  - Individual ID
  - Species ID
  - Disease state
  - Etc.



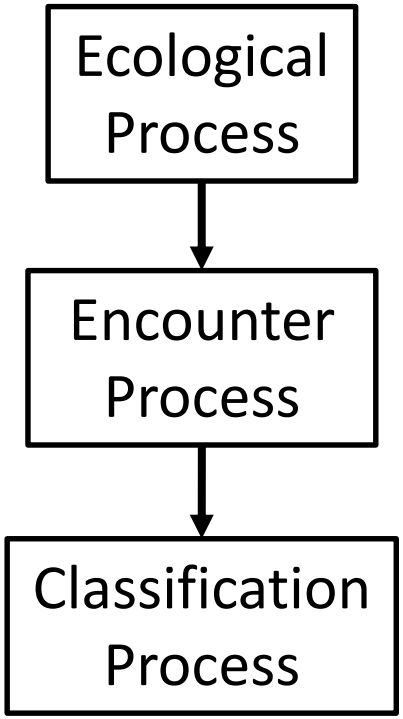
Coupled Classification

- Typical approach: “uncoupled classification”
- Uncertainty in classification not propagated to ecological parameters
  - Retained data treated as error free
  - Data loss
- Solution: coupled classification in a hierarchical ecological model



Coupled Classification

- Hierarchical Ecological Models
  - Model the processes that produced our data
- Coupled Classification
  - Classification process determines data we observe
  - Joint estimation of all parameters
- Regard ecological classes in encounter history as variables to be estimated
- Bayesian estimation: sample from joint posterior for encounter history and model parameters via Markov Chain Monte Carlo (MCMC)



Coupled Classification Occupancy

- Base model: multispecies occupancy model with count detections
  - Independent occupancy process (no sps interactions)
  - "species" can be any class, really.
    - Background noise, higher taxonomic orders, disease state, etc.

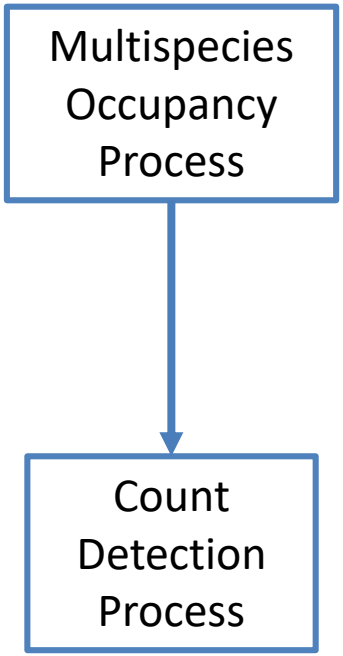
species *i, …, N*
site *j*, …, J
occasion *k, …, K*

$\psi_{i,j}$: P(site *j* occupied by species *i*)

$$z_{i,j} \sim \text{Bernoulli}(\psi_{i,j})$$

$\lambda_{i,j,k}$: Detection rate of species *i* at site *j* on occasion *k*

$$[y_{i,j,k}|z_{i,j}] \sim \text{Poisson}(\lambda_{i,j,k} * z_{i,j})$$



USGS science for a changing world

Coupled Classification Occupancy

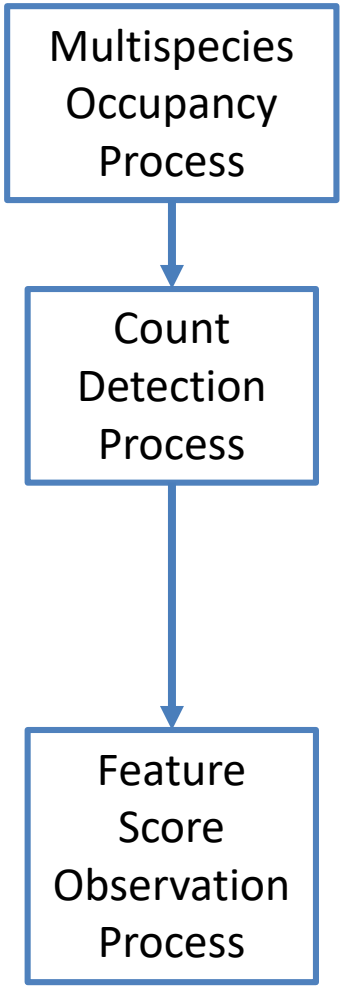
- Unknown species ID extension
  - Observe a species "feature score" instead
  - Correlates with species ID
  - γ_l: species ID of sample *l*

$$z_{i,j} \sim \text{Bernoulli}(\psi_{i,j})$$

$$[y_{i,j,k}|z_{i,j}] \sim \text{Poisson}(\lambda_{i,j,k} * z_{i,j})$$

Disaggregate $y_{i,j,k}$ to individual detections, index by sample number *l*

$$[g_l|\gamma_l] \sim \text{Normal}(\mu_{\gamma_l}, \sigma_{\gamma_l})$$




```
g[l] ~ dnorm(mu=mu[gamma[l]],sd=sigma[gamma[l]])
```

USGS science for a changing world

Coupled Classification Occupancy

- Unknown species ID extension
  - Observe a species "feature score" instead
  - Correlates with species ID
  - γ_l: species ID of sample *l*

$$z_{i,j} \sim \text{Bernoulli}(\psi_{i,j})$$

$$[y_{i,j,k}|z_{i,j}] \sim \text{Poisson}(\lambda_{i,j,k} * z_{i,j})$$

$$[g_l|\gamma_l] \sim \text{Normal}(\mu_{\gamma_l}, \sigma_{\gamma_l})$$

Multispecies Occupancy Process

Count Detection Process

Feature Score Observation Process

Example

$y_{1,3,2} = 2$

$y_{2,3,2} = 3$

Disaggregate $y_{i,j,k}$ to individual detections, index by sample number *l*

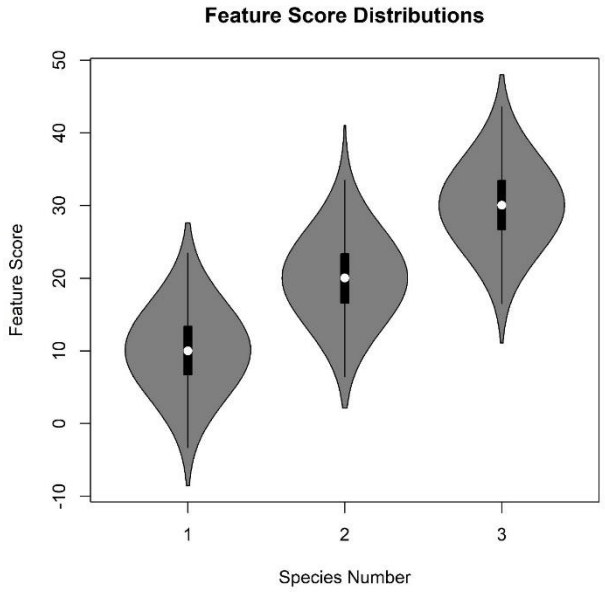
γ_1= 1, site$_1$= 3, occ$_1$ = 2
γ_2= 1, site$_2$= 3, occ$_2$ = 2
γ_3= 2, site$_3$= 3, occ$_3$ = 2
γ_4= 2, site$_4$= 3, occ$_4$ = 2
γ_5= 2, site$_5$= 3, occ$_5$ = 2

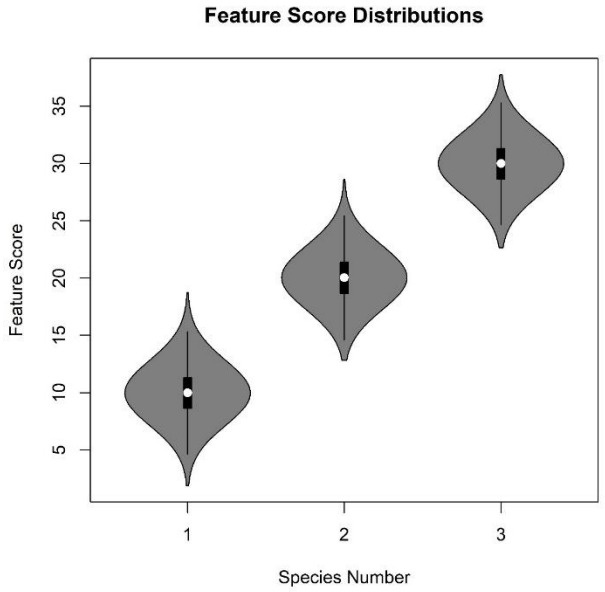
```
g[l] ~ dnorm(mu=mu[gamma[l]],sd=sigma[gamma[l]])
```

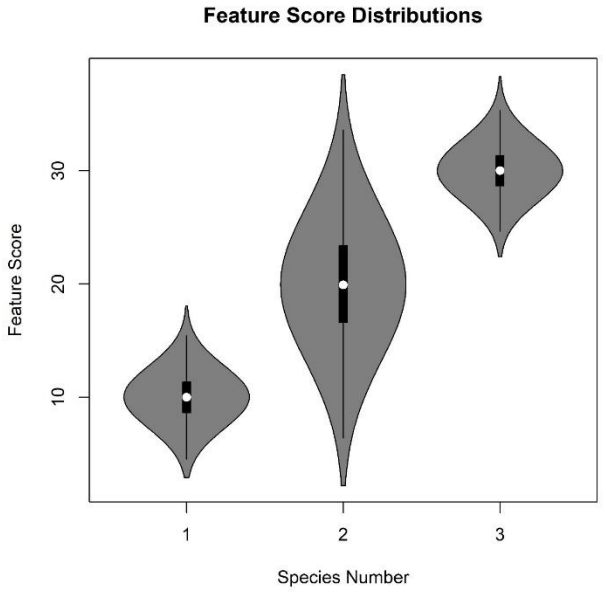
USGS science for a changing world

Coupled Classification Occupancy

- Hypothetical Feature Score Visualization
  - 3 species, Normal Distribution







$$[g_l|\gamma_l] \sim \text{Normal}(\mu_{\gamma_l}, \sigma_{\gamma_l})$$

```
g[l] ~ dnorm(mu=mu[gamma[l]],sd=sigma[gamma[l]])
```

USGS science for a changing world

Coupled Classification Occupancy

- Factor count likelihood to fit model
  - Only works with Poisson count model

$$[y_{i,j,k}|z_{i,j}] \sim \text{Poisson}(\lambda_{i,j,k} * z_{i,j})$$

Site by occasion contribution

$$y_{.,j,k} \sim \text{Poisson}(\Lambda_{j,k})$$

$$y_{.,j,k} = \sum_i y_{i,j,k}$$

$$\Lambda_{j,k} = \sum_i \lambda_{i,j,k} * z_{i,j}$$

Species contribution

$$\gamma_l \sim \text{Categorical}\left(\frac{\lambda_{1:N,j,k} * z_{1:N,j}}{\sum_i \lambda_{i,j,k} * z_{i,j}}\right)$$

"Ecological Prior" for species ID γ_l

```
for(l in 1:n.samples){
  gamma[l] ~ dcat(sps.prob[1:N,G.site[l],G.occ[l]])
  g[l] ~  dnorm(mu=mu[gamma[l]],sd=sigma[gamma[l]])
}
```

```
for(j in 1:J){
  for(k in 1:K){
    bigLam[j,k] <- sum(lambda[1:N,j,k]*z[1:N,j])
    y2D[j,k] ~ dpois(bigLam[j,k])
    sps.prob[1:N,j,k] <- (lambda[1:n.species,j,k]*z[1:N,j])/bigLam[j,k]
  }
}
```

Coupled Classification Occupancy

- "Availability" process
  - Optional, but a good idea!
  - Site by occasion-level zero-inflation

$$z_{i,j} \sim \text{Bernoulli}(\psi_{i,j})$$

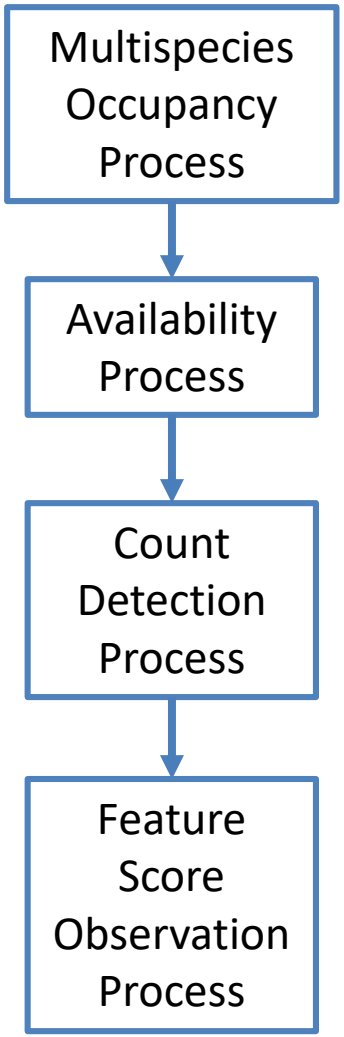
$$[w_{i,j,k}|z_{i,j}] \sim \text{Bernoulli}(\theta_{i,j,k} * z_{i,j})$$

$$[y_{i,j,k}|w_{i,j,k}] \sim \text{Poisson}(\lambda_{i,j,k} * w_{i,j,k})$$

$$[g_l|\gamma_l] \sim \text{Normal}(\mu_{\gamma_l}, \sigma_{\gamma_l})$$

Modified "ecological prior"

$$\gamma_l \sim \text{Categorical}\left(\frac{\lambda_{1:N,j,k} * w_{i:N,j,k}}{\sum_i \lambda_{i,j,k} * w_{i,j,k}}\right)$$



Coupled Classification Occupancy

- Ecological "Prior Information"
  - Samples at site *j* are more likely to belong to
    - Species more likely to occupy site *j*
    - Species with higher detection rate| occupancy at site *j*

$$\gamma_l \sim \text{Categorical}\left(\frac{\lambda_{1:N,j,k} * z_{1:N,j}}{\sum_i \lambda_{i,j,k} * z_{i,j}}\right)$$

With Availability Process

$$\gamma_l \sim \text{Categorical}\left(\frac{\lambda_{1:N,j,k} * w_{i:N,j,k}}{\sum_i \lambda_{i,j,k} * w_{i,j,k}}\right)$$

  - Function of species-specific occupancy, availability and detection intercepts and possible site and occasion covariate relationships or random effects.

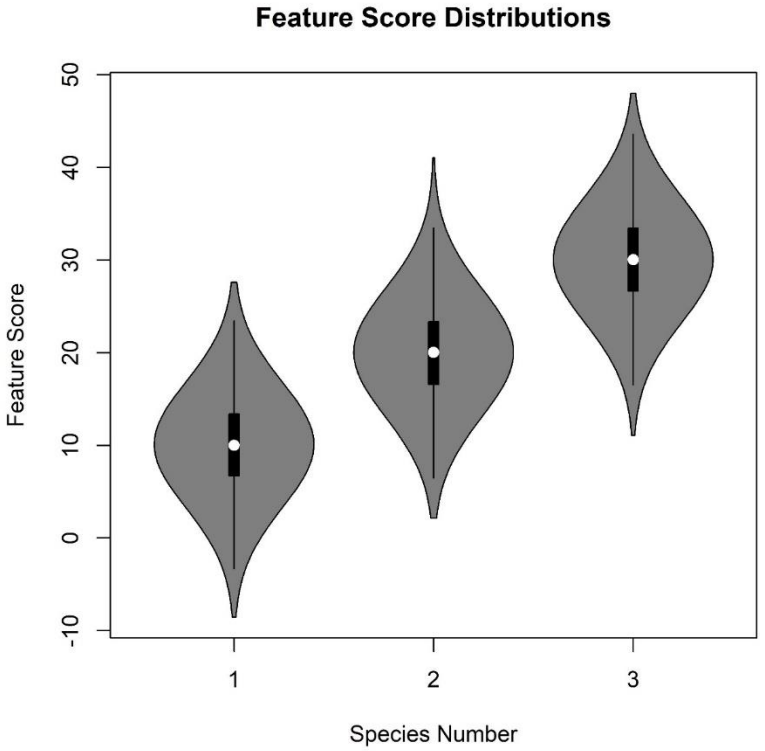
$$\text{logit}(\psi_{i,j}) = \beta_{0,i} + \beta_{1,i} * \text{cov}_j$$

$$\text{logit}(\theta_{i,j,k}) = \alpha_{0,i} + \alpha_{1,i} * \text{cov}_j$$

$$\log(\lambda_{i,j,k}) = \eta_{0,i} + \eta_{1,i} * \text{cov}_j$$

Validation Data

- Can fit the model without validation data*
  - Sometimes, not recommended!
  - Limited to most simple models?
  - Multimodality present
    - Like Royle and Link (2016), but better able to tell which species is which?



- Types of validation data
  - Independent: not from focal survey
    - Requires “transferability”
  - Random validation of focal survey detections
    - Bonus: some z and w states known

- Alternatively, can use informative prior

Coupled Classification

- Parameter Estimation via MCMC
  - Default algorithm sometimes will not converge/fully explore posterior
  - Difficulty sampling latent counts conditioned on latent indicator variables
    - Example in AHM book should work fine with default algorithm (background z always 1)
  - Can marginalize likelihood over all latent variables, **z**, **w**, $\boldsymbol{\gamma}$
    - Marginalize for every parameter update on each iteration (could fit in Stan this way)
    - Sample **z**, **w**, $\boldsymbol{\gamma}$ from marginal distributions (once per iteration)
    - Must calculate the likelihood for all possible combinations of z_{ij} and w_{ijk} at each site
    - Number of combinations is 2^N

```
> 2^(1:15)
 [1]     2     4     8    16    32    64   128   256   512  1024  2048  4096  8192 16384 32768
```

  - Implemented in Nimble via custom update
    - github.com/benaug/Coupled-Classification-Occupancy

Coupled Classification Occupancy

- Some possible Feature Score Distributions
  - **Any parametric distribution**

Feature Score Observation Process

$[g_l|\gamma_l] \sim \text{Normal}(\mu_{\gamma_l}, \sigma_{\gamma_l})$

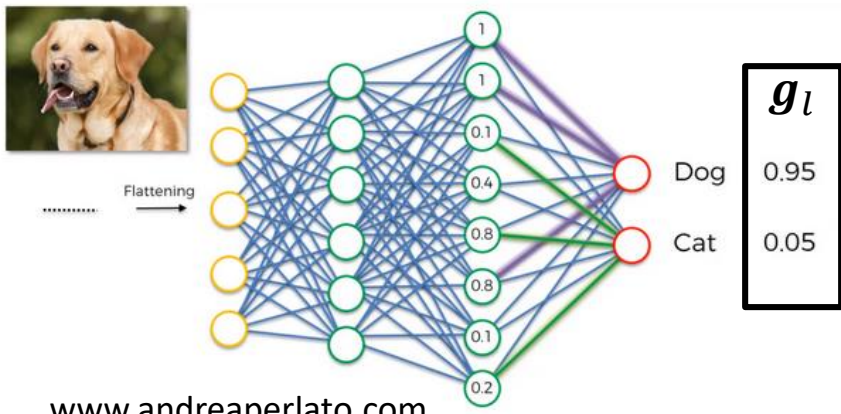
$[g_l|\gamma_l] \sim \text{Categorical}(\pi_{\gamma_l})$ (Equivalent to Wright et al. 2020)

$[\boldsymbol{g}_l|\gamma_l] \sim \text{Dirichlet}(\boldsymbol{\pi}_{\gamma_l})$

Vectors of length N

$\boldsymbol{\pi}_{\gamma_l}$ sums to 1

Softmax vector output from CNN



www.andreaperlato.com

Coupled Classification Owl Example

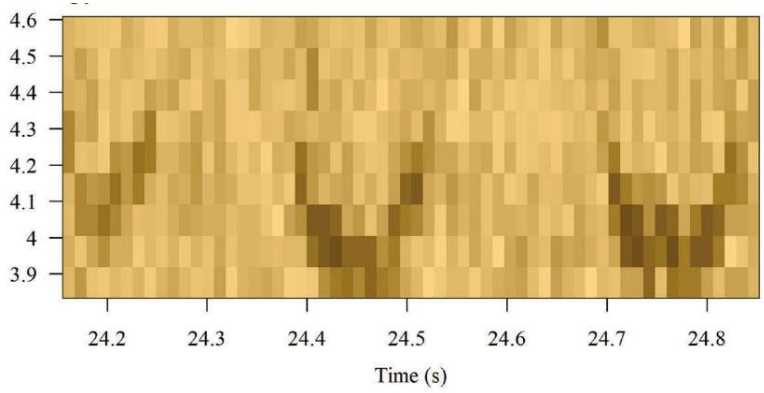
- Northern Spotted Owl (NS0) acoustic occupancy data set
- Collected by Connor Wood, Zach Peery, and others
- 45 sites, 3 weekly occasions
- 2-3 ARUs per site
- Variable Effort: ARU-nights
- Acoustic events identified by correlation with NSO template
- Statistically, owl calls difficult to distinguish from “background noise”

USGS science for a changing world

Coupled Classification Owl Example

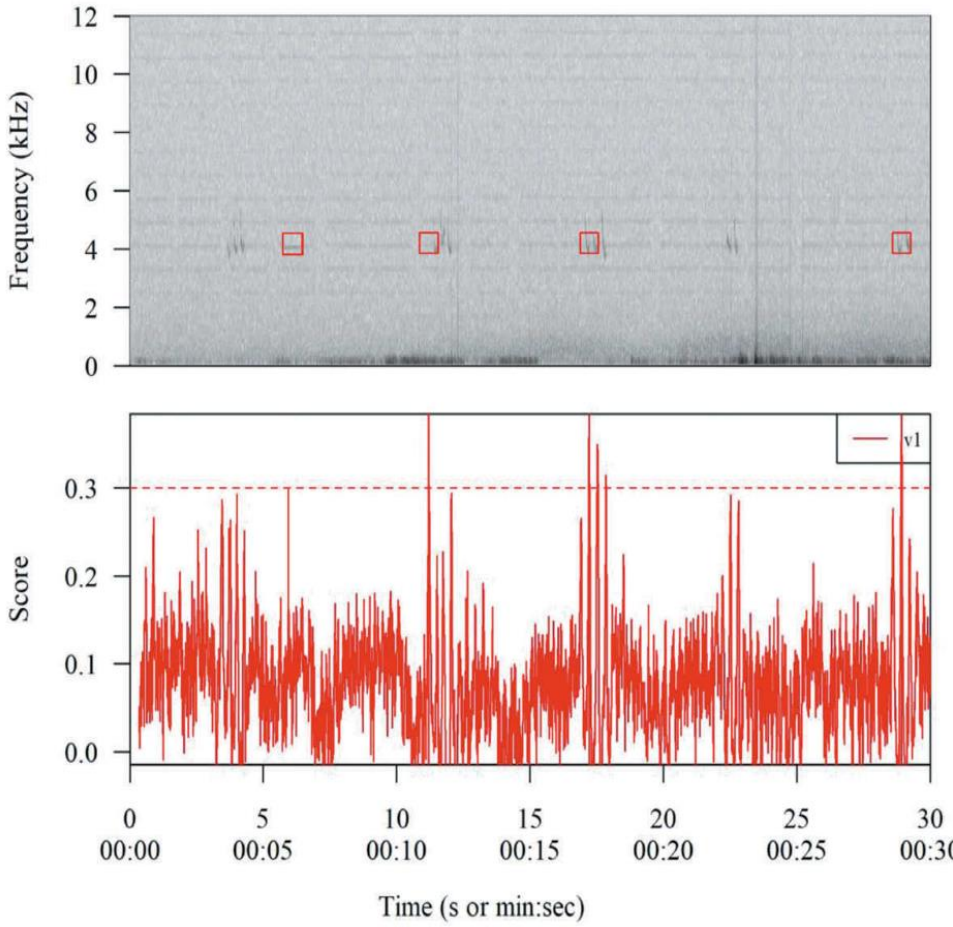
- Acoustic events identified by correlation with NSO template
  - Events with correlation < 0.8 discarded
  - Humans classified 12777 all events
    - Owl vs. background
  - Event “features” measured

Template



Balantic and Donovan 2020

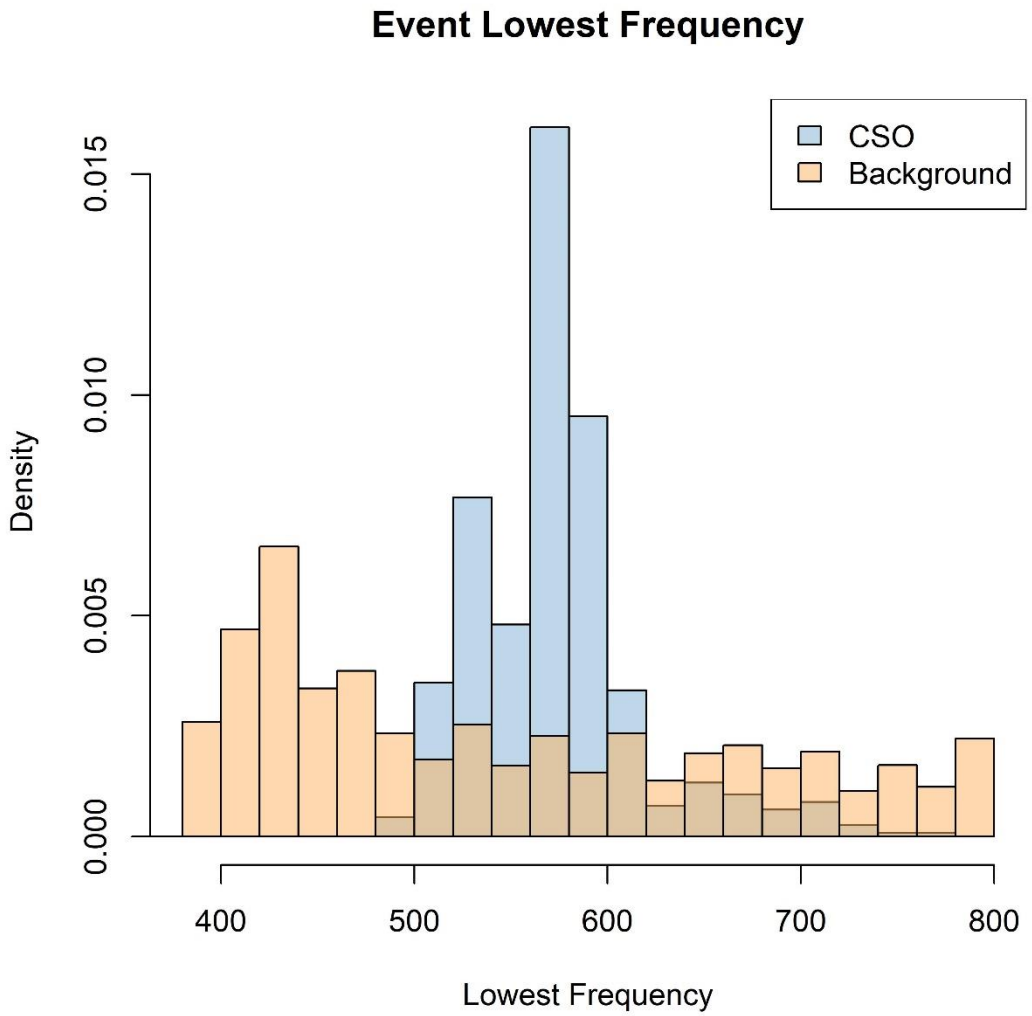
Full Recording



USGS science for a changing world

Coupled Classification Owl Example

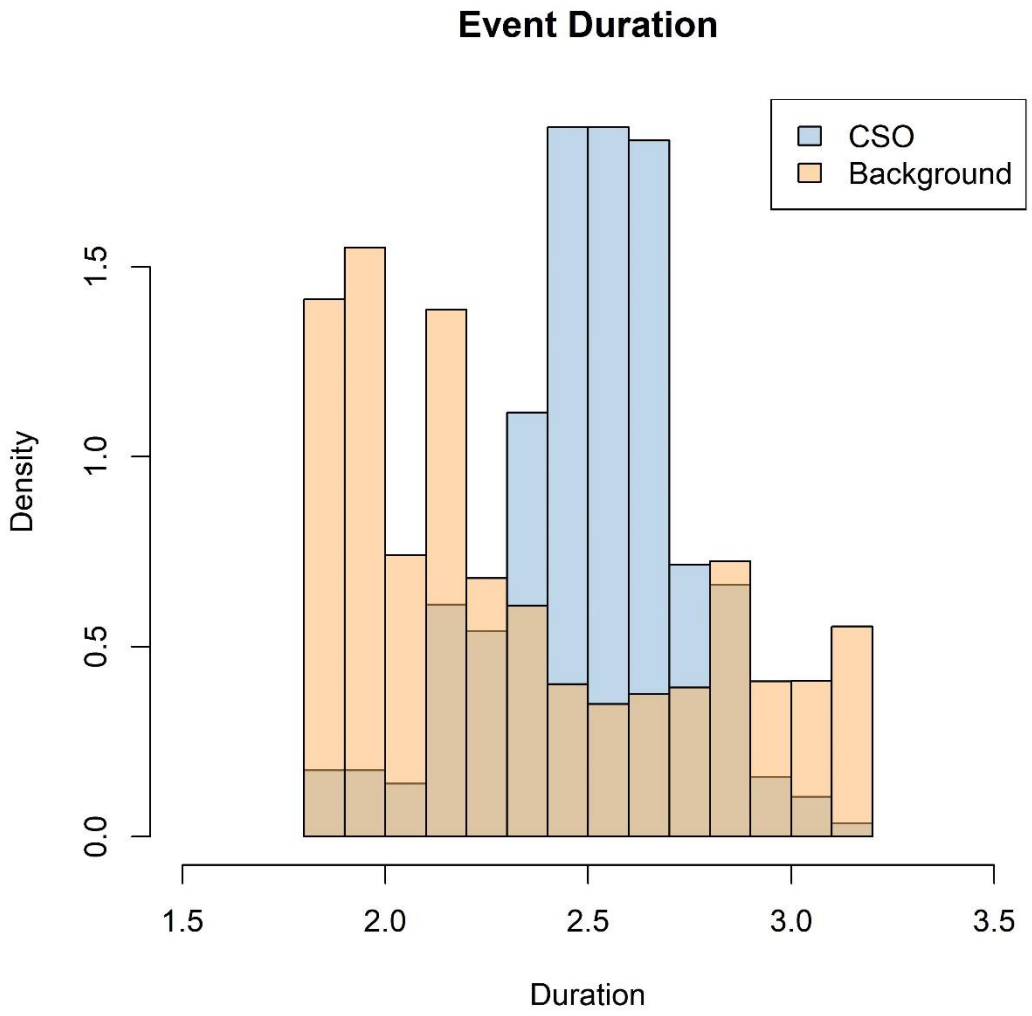
- Potential Feature Scores



USGS
science for a changing world

Coupled Classification Owl Example

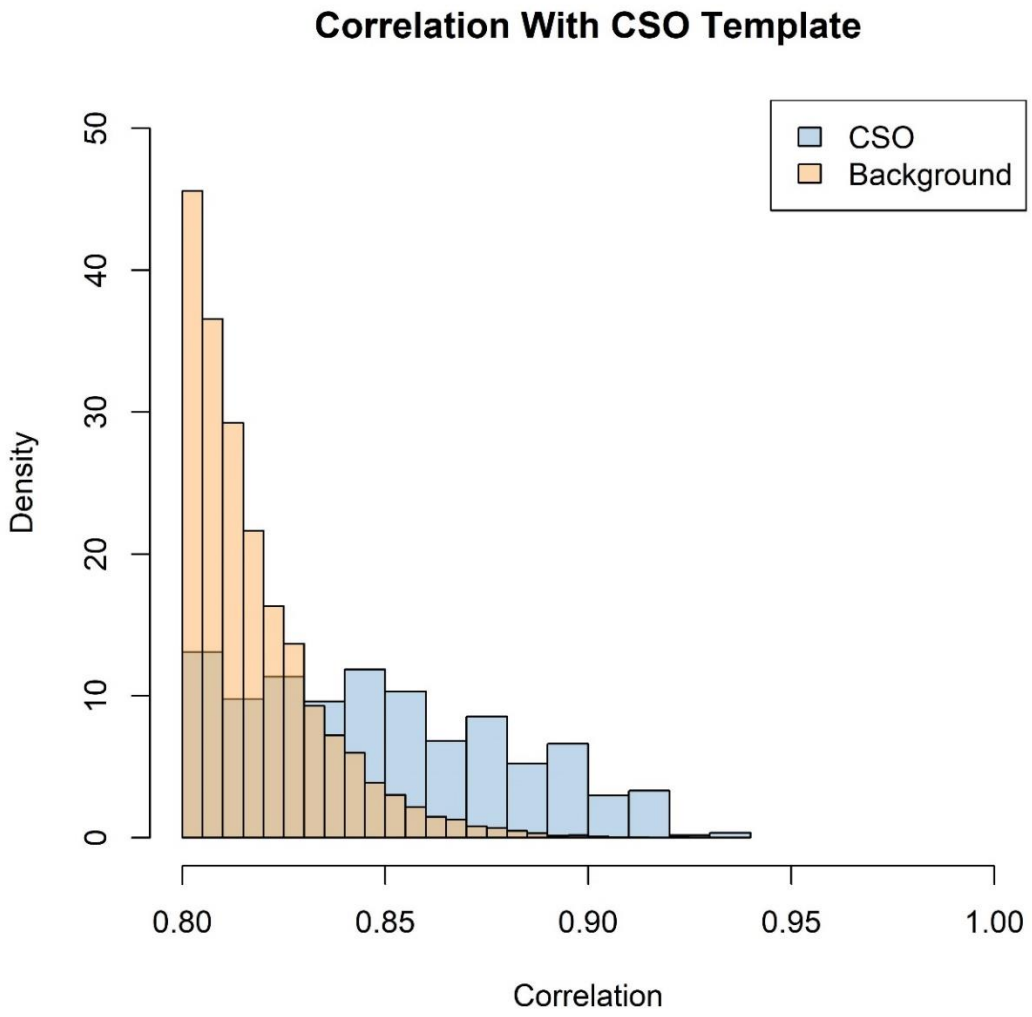
- Potential Feature Scores



USGS
science for a changing world

Coupled Classification Owl Example

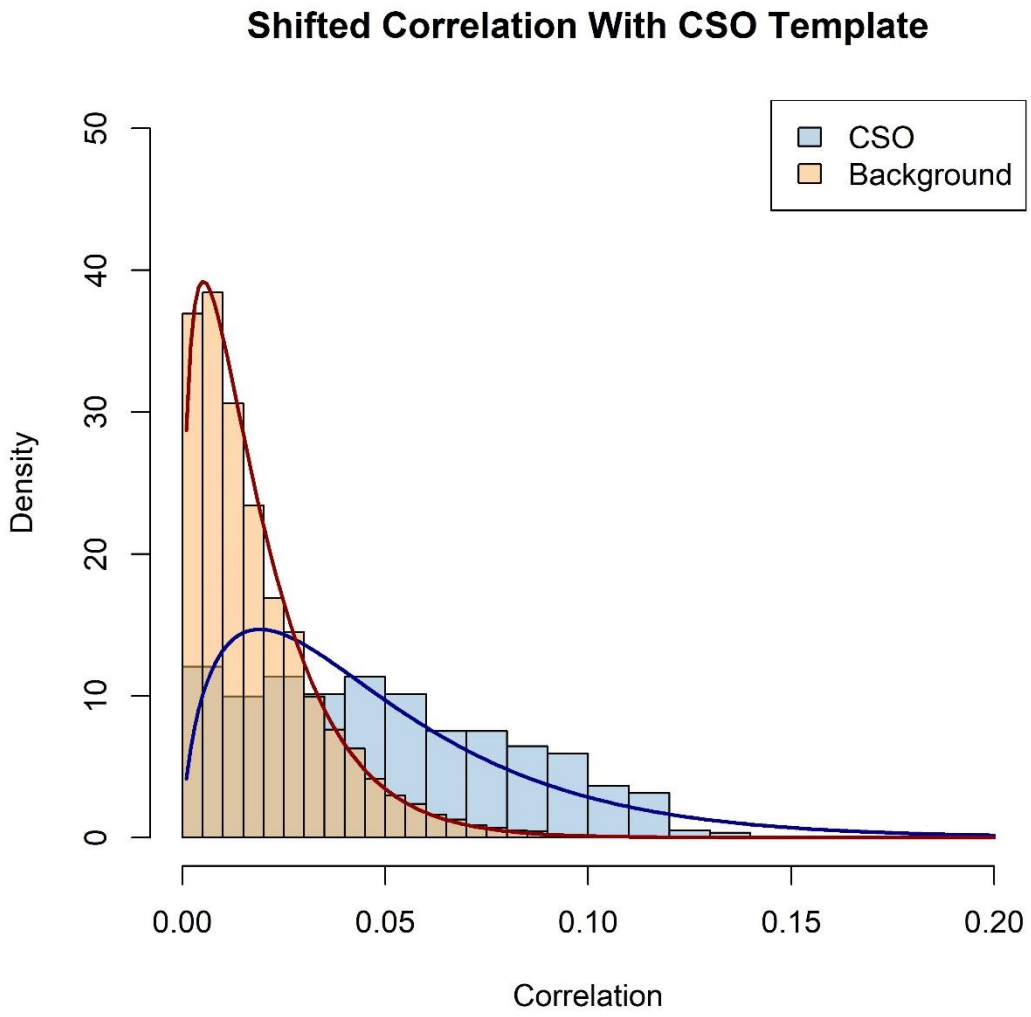
- Potential Feature Scores



USGS science for a changing world

Coupled Classification Owl Example

- Potential Feature Scores
  - Treat as Gamma RV
  - Other choices?
    - Truncated Normal?



USGS science for a changing world

Coupled Classification Owl Example

For species *i*, site *j*, occasion *k*

- Model Specification
  - Occupancy

$$\text{logit}(\psi_{i,j}) = \beta_{0,i} + \beta_{1,i} * \text{slope}_j$$

$$z_{i,j} \sim \text{Bernoulli}(\psi_{i,j})$$

  - Availability

$$\text{logit}(\theta_{i,j}) = \alpha_{0,i} + \alpha_{1,i} * \text{slope}_j$$

$$[w_{i,j,k}|z_{i,j}] \sim \text{Bernoulli}(\psi_{i,j} * z_{i,j})$$

  - Detection
    - Site by occasion random effect for each species

$$\log(\lambda_{i,j,k}) \sim \text{Normal}(\mu_i, \sigma_i)$$

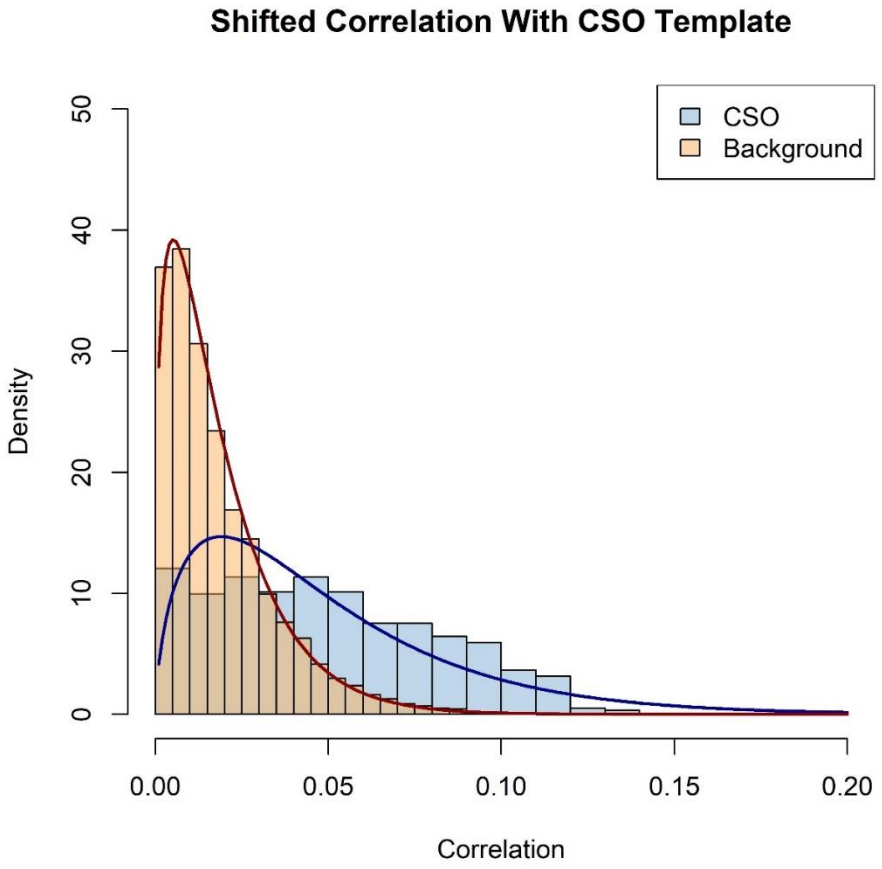
$$[y_{i,j,k}|w_{i,j,k}] \sim \text{Poisson}(\lambda_{i,j,k} * w_{i,j,k} * \text{effort}_{i,j,k})$$

Coupled Classification Owl Example

- Model Specification
  - Feature Score for sample/detection *l*

$g_l \sim$ Gamma(shape$_1$, rate$_1$) if $\gamma_l = 1$

$g_l \sim$ Gamma(shape$_2$, rate$_2$) if $\gamma_l = 2$

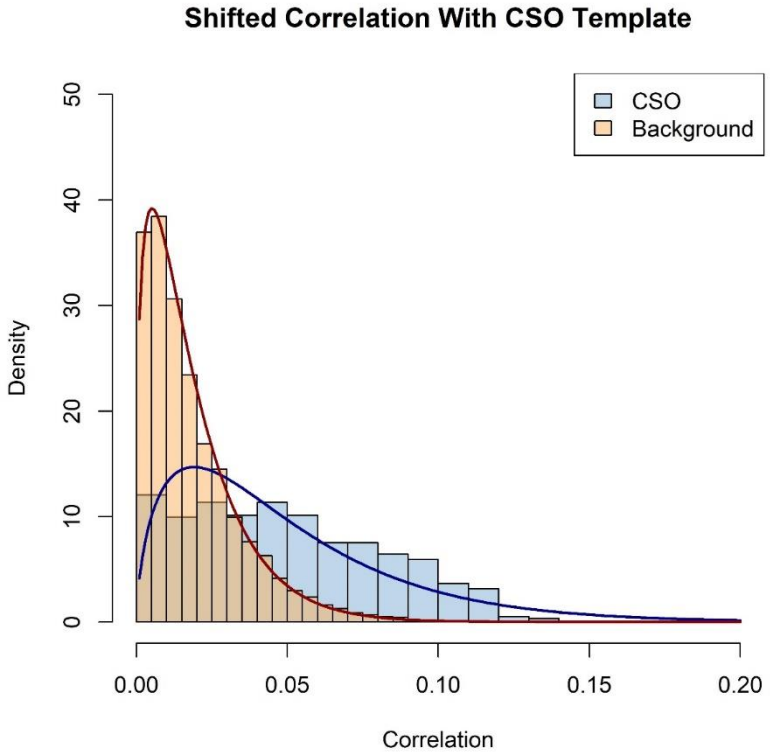



```
for(l in 1:n.samples){
  g[l] ~ dgamma(shape=G.shape[gamma[l]],rate=G.rate[gamma[l]])
}
```

USGS science for a changing world

Coupled Classification Owl Example

- Modeling Exercise
  - How do model estimates vary as more species IDs are provided?
  - Fit models with variable % known ID
    - 100% (regular occupancy model)
    - 50%
    - 25%
    - 10%
  - n=1 randomization of known ID samples
    - Better to try many possible ways to select X% to ID

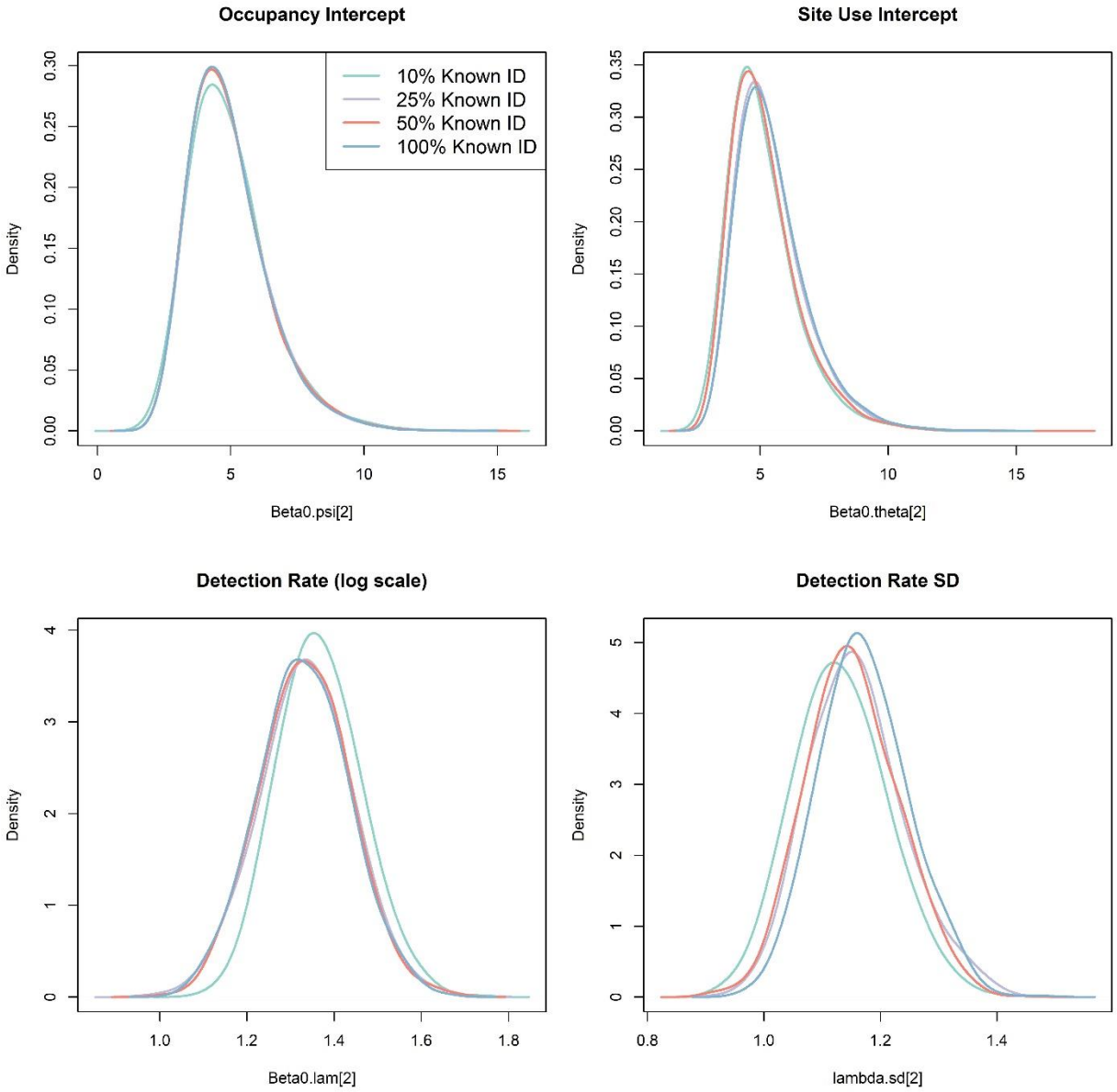



```
for(l in 1:n.samples){
  g[l] ~ dgamma(shape=G.shape[gamma[l]],rate=G.rate[gamma[l]])
}
```

USGS science for a changing world

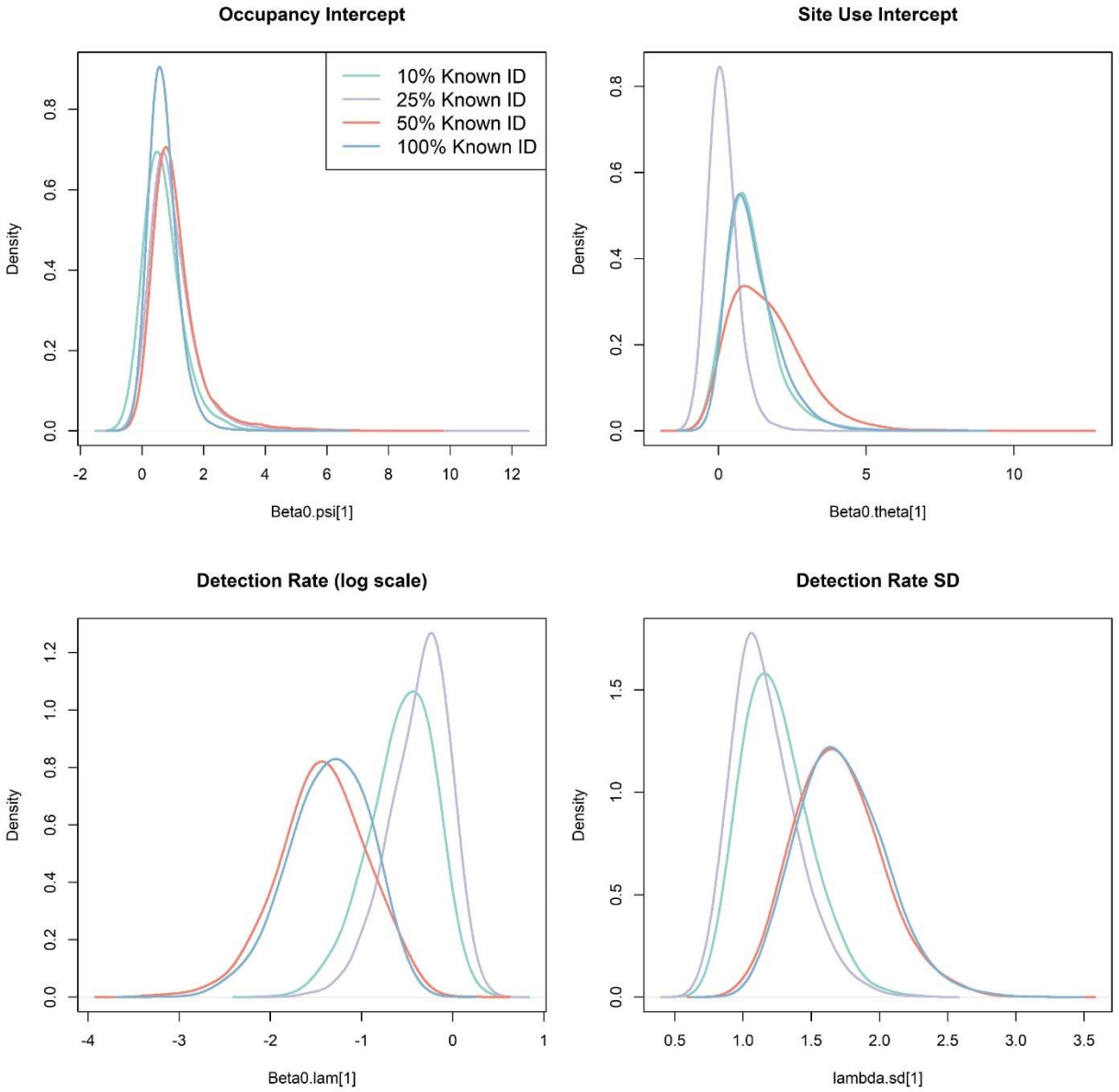
Coupled Classification Owl Example

- Compare (some) parameter estimates
  - Background



Coupled Classification Owl Example

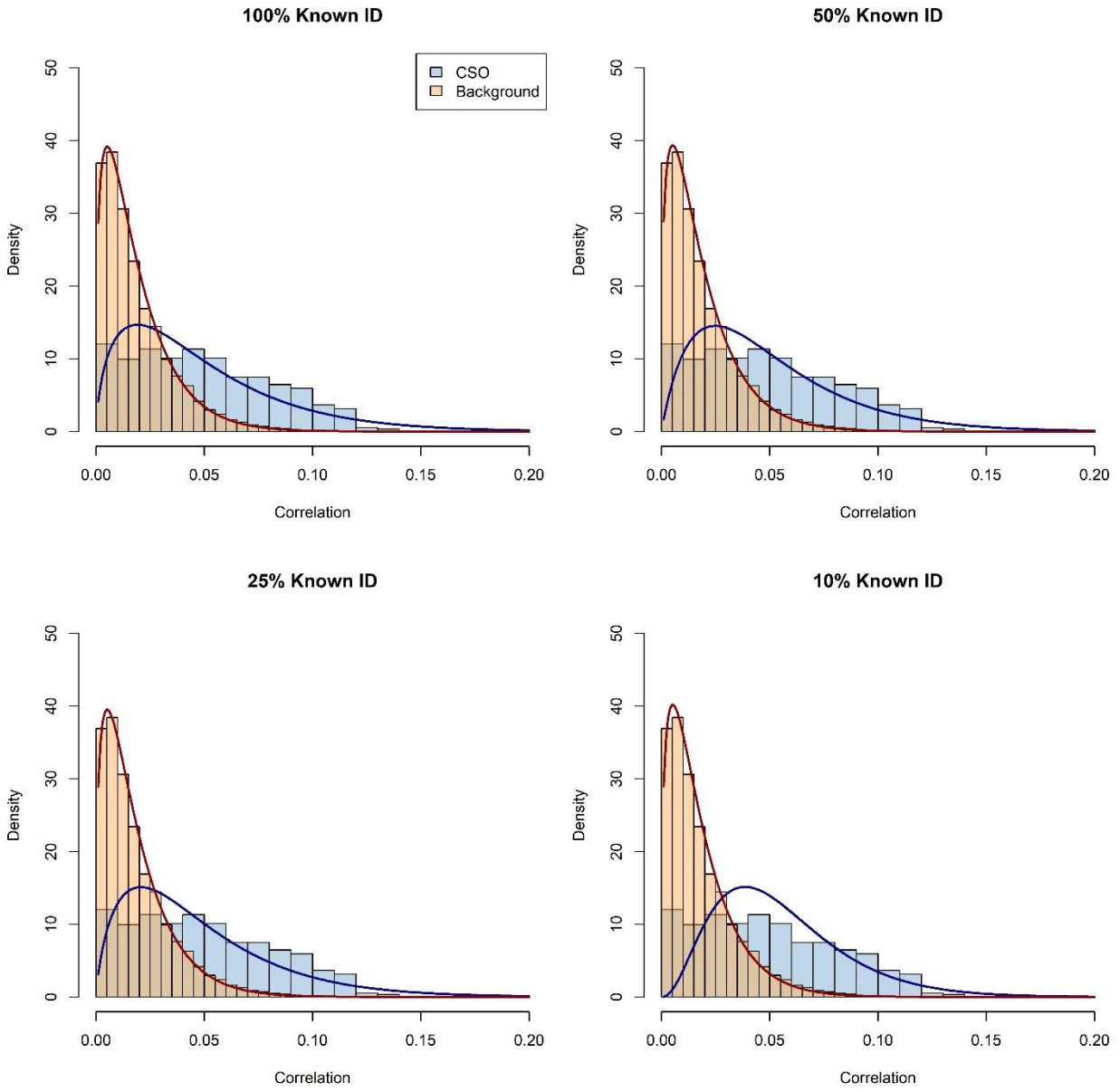
- Compare (some) parameter estimates
  - Owl



USGS science for a changing world

Coupled Classification Owl Example

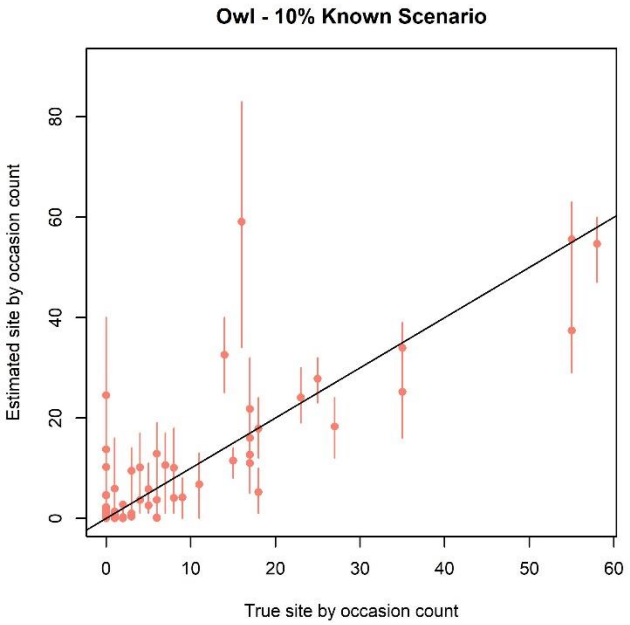
- Compare (some) parameter estimates
  - Gamma fit
  - Background fit deviates more with fewer known ID
  - Mostly in area of overlap
  - Not quite gamma?
  - Feature covariate doesn't separate classes well?
  - Site Heterogeneity?



USGS science for a changing world

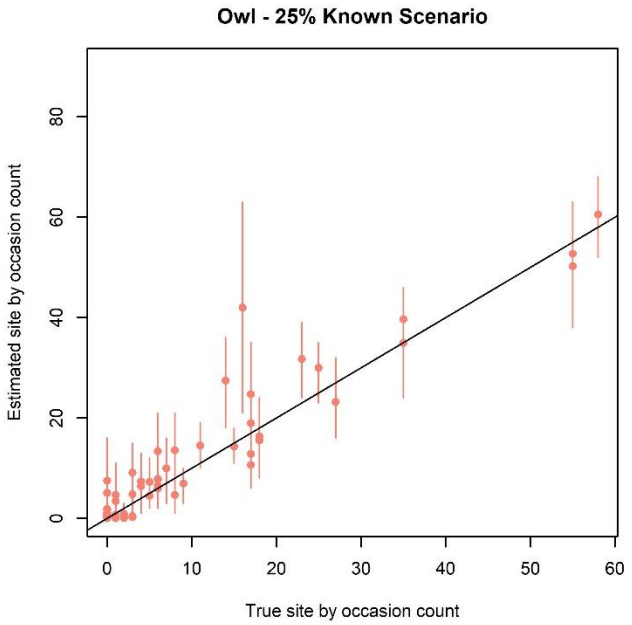
Coupled Classification Owl Example

- How do the models do estimating the true site by occasion detection counts?
  - Owl
  - y[1,j,k]



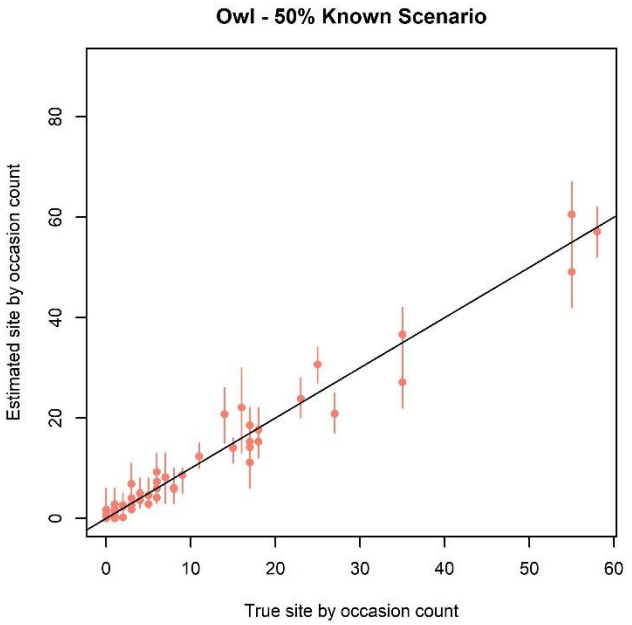
95% Coverage: 85%

64/573 Known



95% Coverage: 93%

147/573 Known



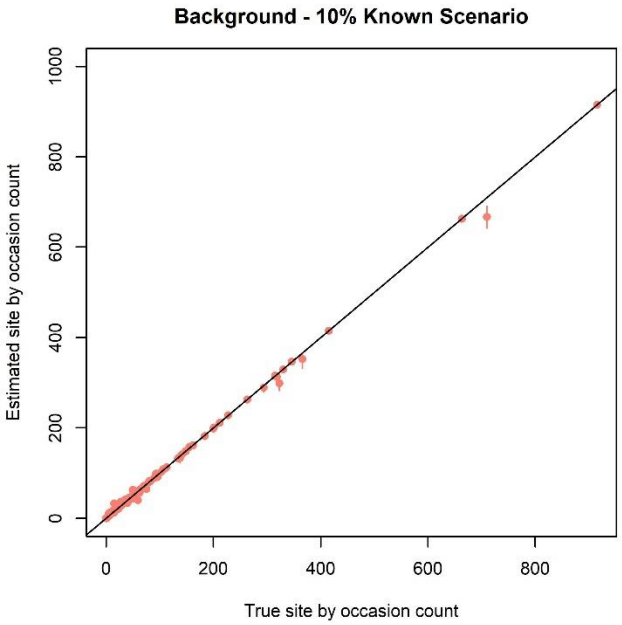
95% Coverage: 94%

286/12204 Known

USGS science for a changing world

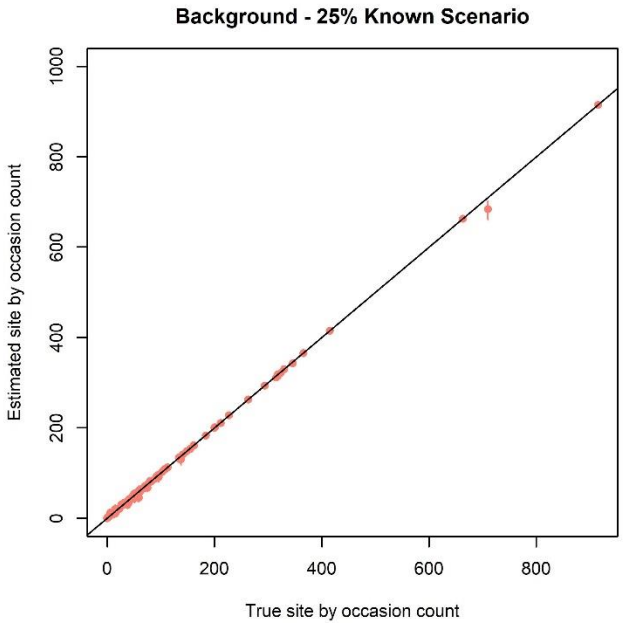
Coupled Classification Owl Example

- How do the models do estimating the true site by occasion detection counts?
  - Background
  - y[2,j,k]



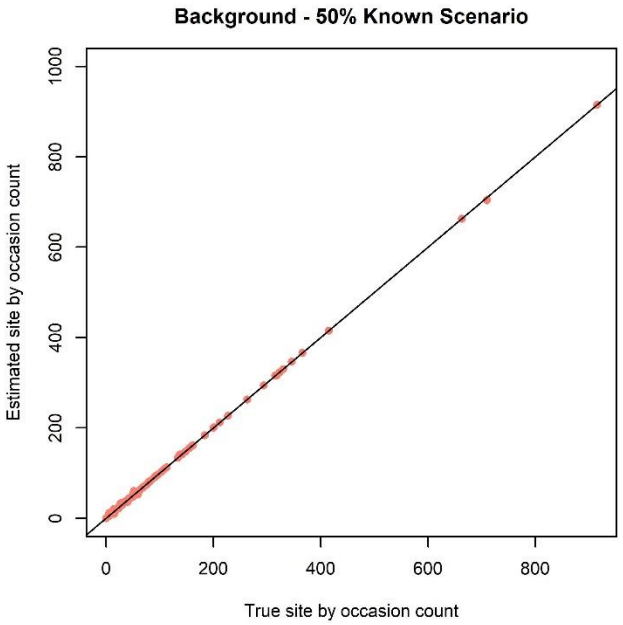
95% Coverage: 87%

1231/12204 Known



95% Coverage: 92%

2998/12204 Known



95% Coverage: 95%

6074/12204 Known

USGS science for a changing world

Coupled Classification Owl Example

- Can estimate recall and precision
  - Point estimates
  - Uncertainty estimates
  - Excluding known-ID samples
  - Recall: P(classify as owl|true owl), related to detection
  - Precision" P(true owl|classified as true owl), related to classification error

Recall

% Known	Owl	Background
10	0.56	0.99
25	0.62	0.99
50	0.56	0.99

Precision

% Known	Owl	Background
10	0.67	0.98
25	0.71	0.98
50	0.76	0.98

Some Concluding Remarks

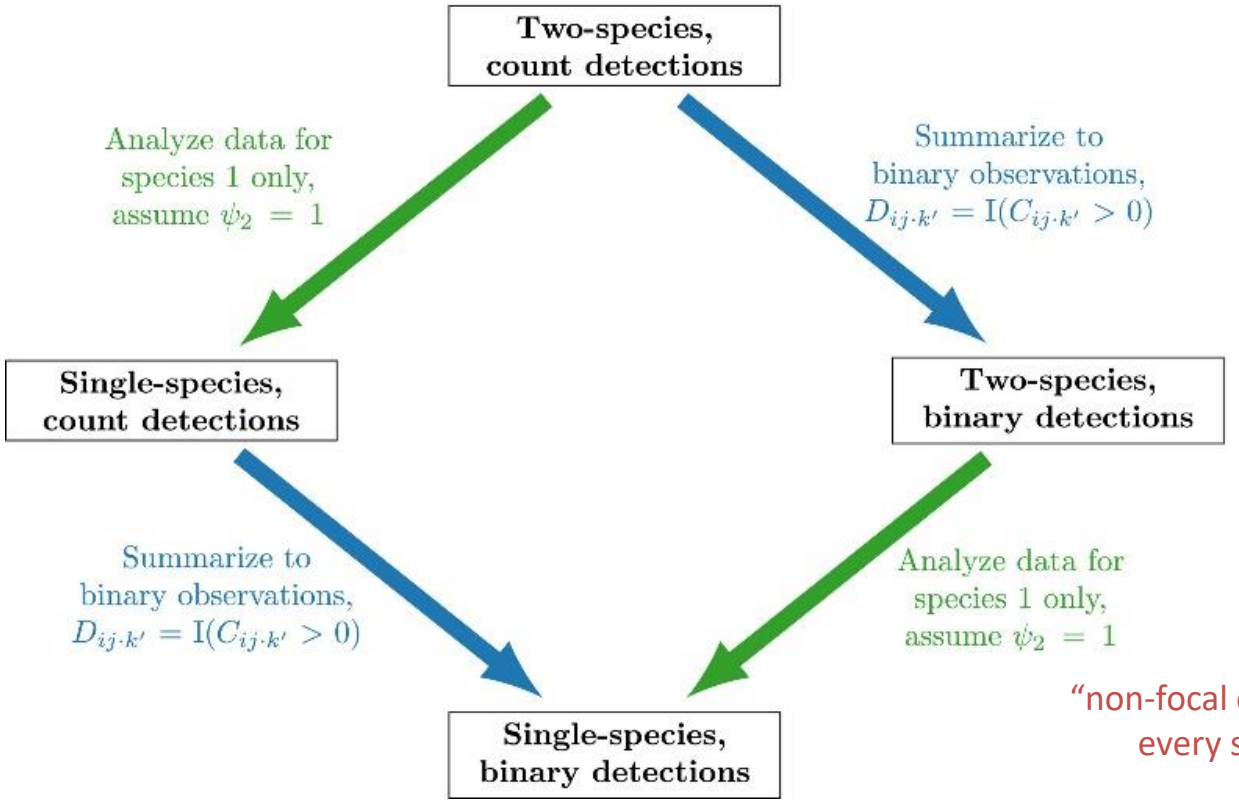
- A “more mechanistic” false positive model
  - Greater understanding of false positive sources
- Greater ability to model heterogeneity in false positive rates across sites
  - Depends on how frequently each species is confused with focal and their ecological parameters determining how frequently they are detected in space and time
- More likely that feature score distributions more similar across sites/studies that false positive probabilities
  - Still likely to be unmodeled heterogeneity in feature score distributions
- All (?) FP models lean more heavily on parametric assumptions than when no FPs
  - Definitely CC occupancy models

False Positive Model Landscape

- Species number
- Detection vs. count observations

Wright et al. (2020), Spiers et al. (2022)

>1 species



Chambert, Waddle, et al. (2018)

Chambert, Campbell Grant, et al. (2018)

"non-focal occupies every site"

Royle and Link (2006), Chambert et al. (2011), Chambert et al. (2015)

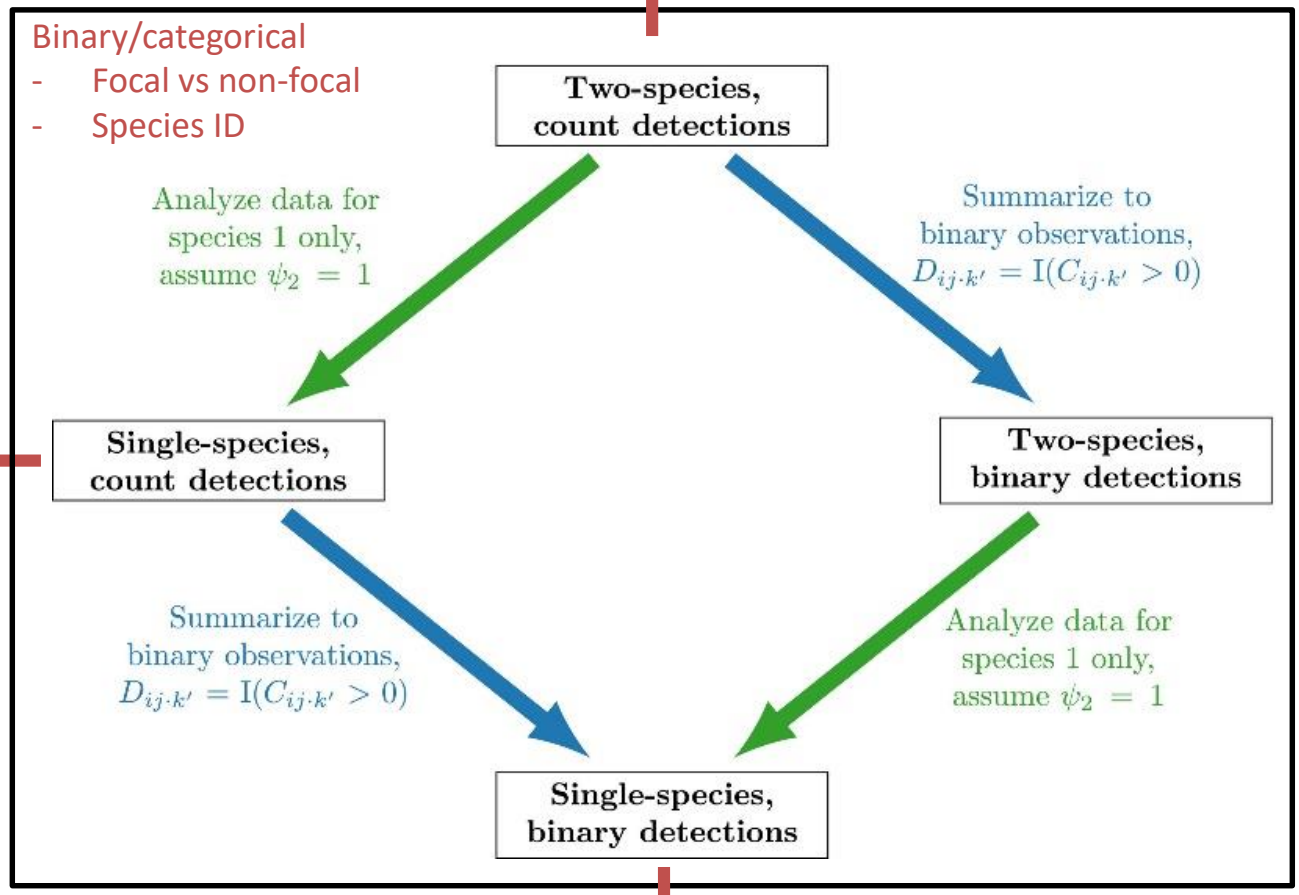
Figure modified from Wright et al. (2020)

False Positive Model Landscape

- Classification information observed/used

Augustine et al. (unpublished, Github) – Normal, Categorical, Dirichlet RV

>1 species



AHM2 Chapter 7 – Normal RV

Rhinehart et al. (2022) - Normal RV

Figure modified from Wright et al. (2020)