Analysis of presence-type data

SSA 200

Presence only vs. Presence-absence data

Example 1: A National Park office often gets reports from people who have seen wolverines in the park. They keep track of the locations and dates of these sightings.

Presence only

Example 2: A researcher conducts point counts for birds in the same forest plot every month from May – July each year. During point counts they make an effort to detect and record every species within the survey window.

Presence-absence

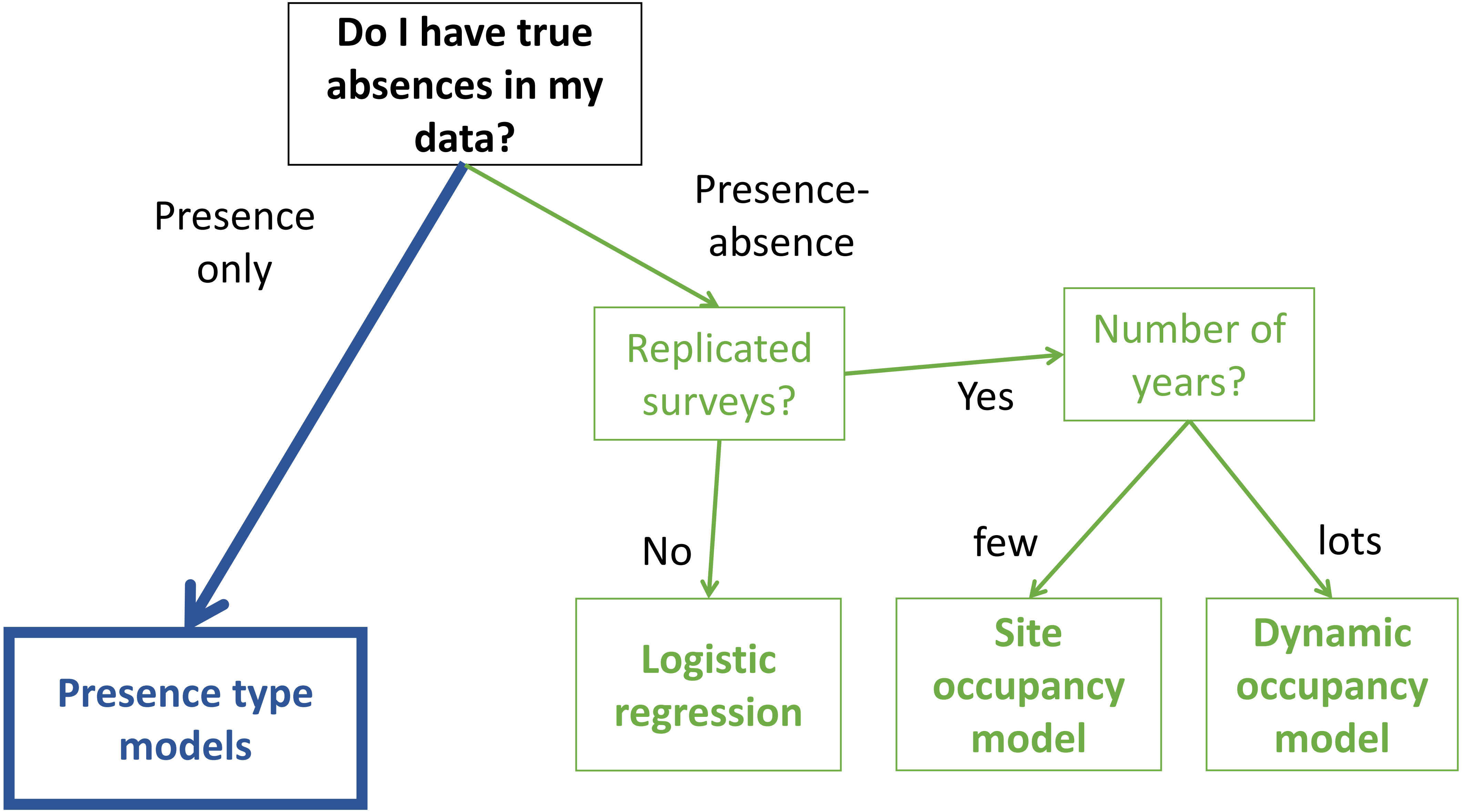
Example 3: A natural history museum has plant specimens that contain information about the location and date of collection.

Presence only

Example 4: iNaturalist has over 29,000 recorded observations of Monarch butterflies from across their range dating back to 2009.

Presence only

alabama cooperative fish and wildlife research unit



Always check specific model assumptions!

Presence only data

Example data sources

- Information about where/when a species was detected but NOT where/when it was looked for but not found
  - Opportunistic reporting by the public
  - Some eBird checklists (**complete checklists = presence-absence data*)
  - Historical records (e.g. museum specimens, field notes)
  - Others?

Example questions

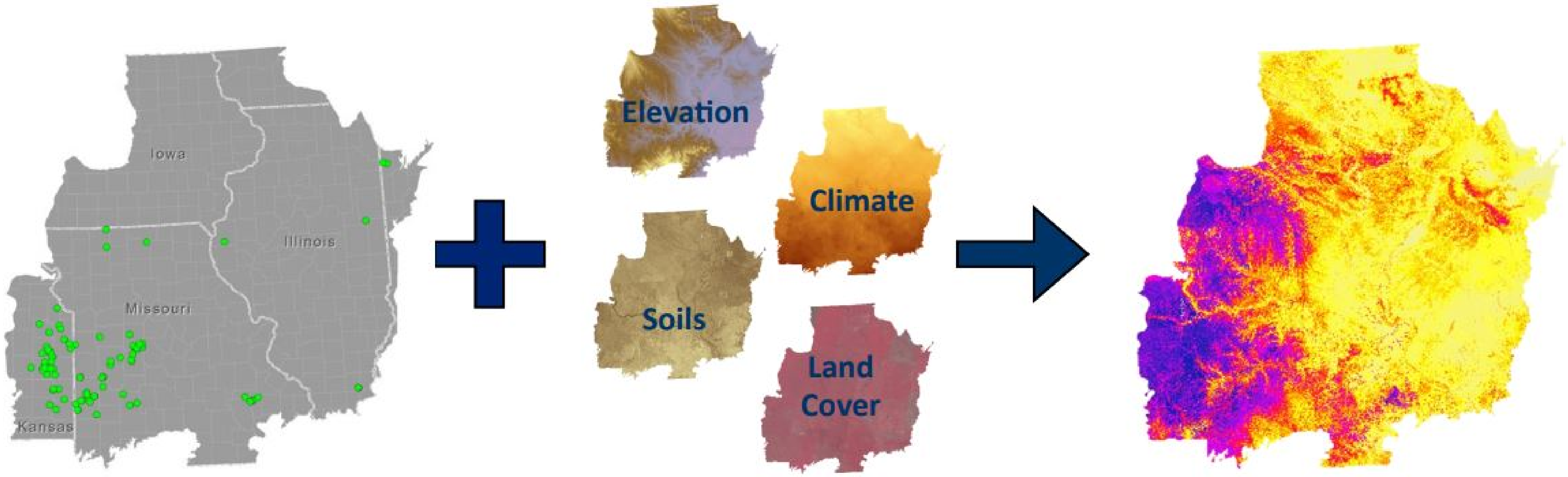
- Where is this species found? *(historic and current condition)*
- What ecological factors best predict its occurrence? *(ecological needs)*

Common analysis approaches

- Species Distribution Modeling
- Paired points

Species distribution modeling

- Habitat suitability models
- When your data covers a broad spatial scale/most of known range



Species Observation Points for *Asclepias Meadii, a threatened milkweed found in agricultural lands in the Midwest*

Examples of input layers used to characterize the environmental setting

Inputs

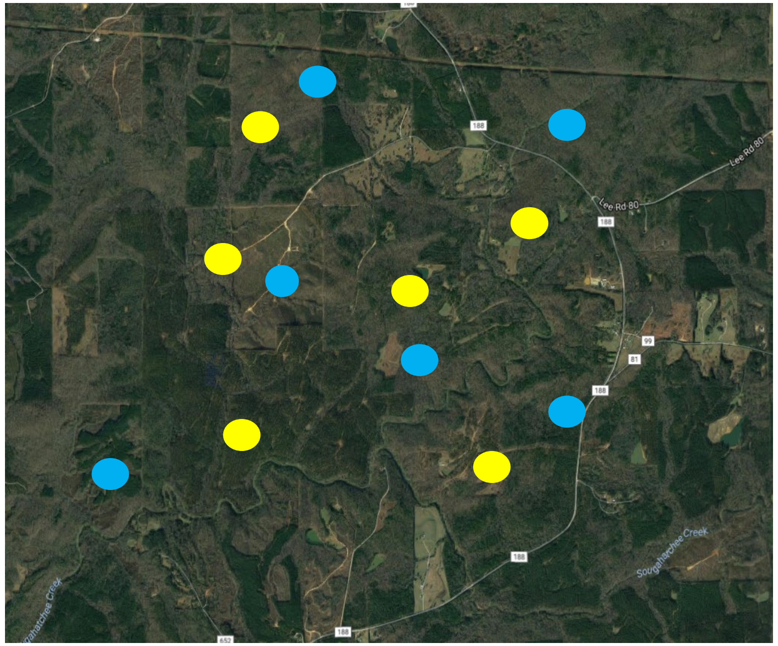
Outputs

Presence data models

Method(s)	Model/software name	Data type
Climatic envelope	BIOCLIM	Presence-only
Gower Metric	DOMAIN	Presence-only
Ecological Niche Factor Analysis (ENFA)	BIOMAPPER	Presence/background
Maximum Entropy	MAXENT	Presence/background
Genetic algorithm	GARP	Presence/pseudo-absence
Regression: Generalized linear model (GLM) and Generalized additive model (GAM)	GRASP	Presence/absence
Artificial Neural Network (ANN)	SPECIES	Presence/absence
Classification and regression trees (CART), GLM, GAM and ANN	BIOMOD	Presence/absence
Boosted decision trees	*(implemented in R)*	Presence/absence
Multivariate adaptive regression splines (MARS)	*(implemented in R)*	Presence/absence

Common analysis approaches

- Species Distribution Modeling
- Paired points – when you have a more limited spatial coverage of observations
    - Randomly select nearby points as “pseudo-absences” – places where the species could have been reported, but wasn’t



Enhance your data!

- Several methods to generate "pseudo-absences"
  - Entire range/study area
  - Within some specified distance of each observation
  - Constrained by environmental variables
- Place random points from the background around your observations
  - Points similar based on the climatic/biological datasets, but where no observation occurred

- Now you have a dataset of 0s and 1s – use logistic regression to estimate effects of environmental covariates

Logistic regression

- A type of **generalized linear model** where the *y* variable is drawn from the Binomial distribution
  - Response variable consists of 1s and 0s (“successes” and “failures”)
- Forms the basis for many more complex models
  - Occupancy analysis
  - Survival analysis
  - Resource selection
- **Assumes perfect detection of the species where it occurs**

alabama cooperative fish and wildlife research unit

U.S. FISH & WILDLIFE SERVICE

Some caveats about presence-only models*

- Temptation for overfitting
- Potential observation bias
- Be careful with extrapolation (both spatial and temporal)
- Omission & commission errors

**not just presence-only models, these apply to many different modeling cases!*

Some caveats about presence-only models

- Temptation for overfitting
  - The number of terms in the model should not exceed the number of observations
  - Consider ecologically-relevant environmental variables to include

Overfitting

Simpler

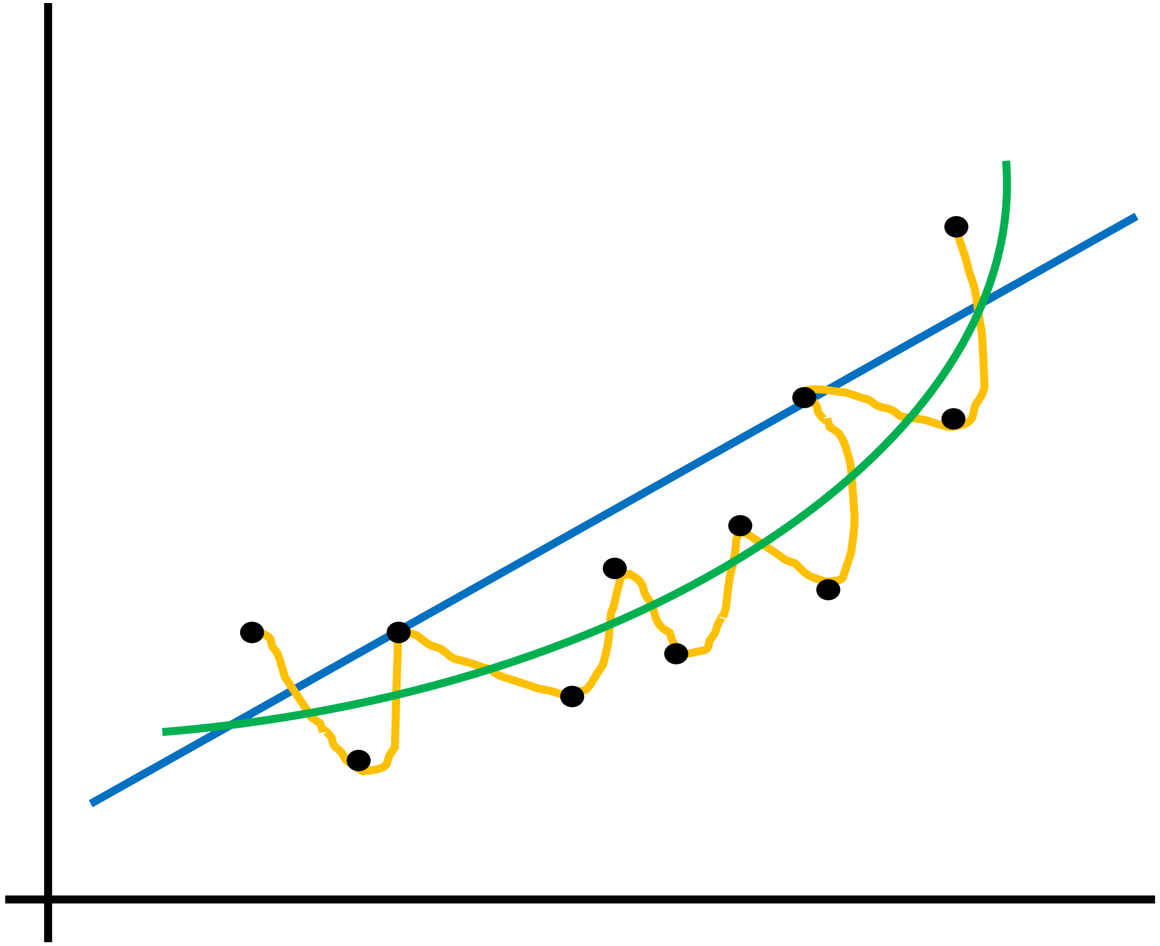
More complex

Worse fit

Best fit

More useful

Less useful



Overfitting

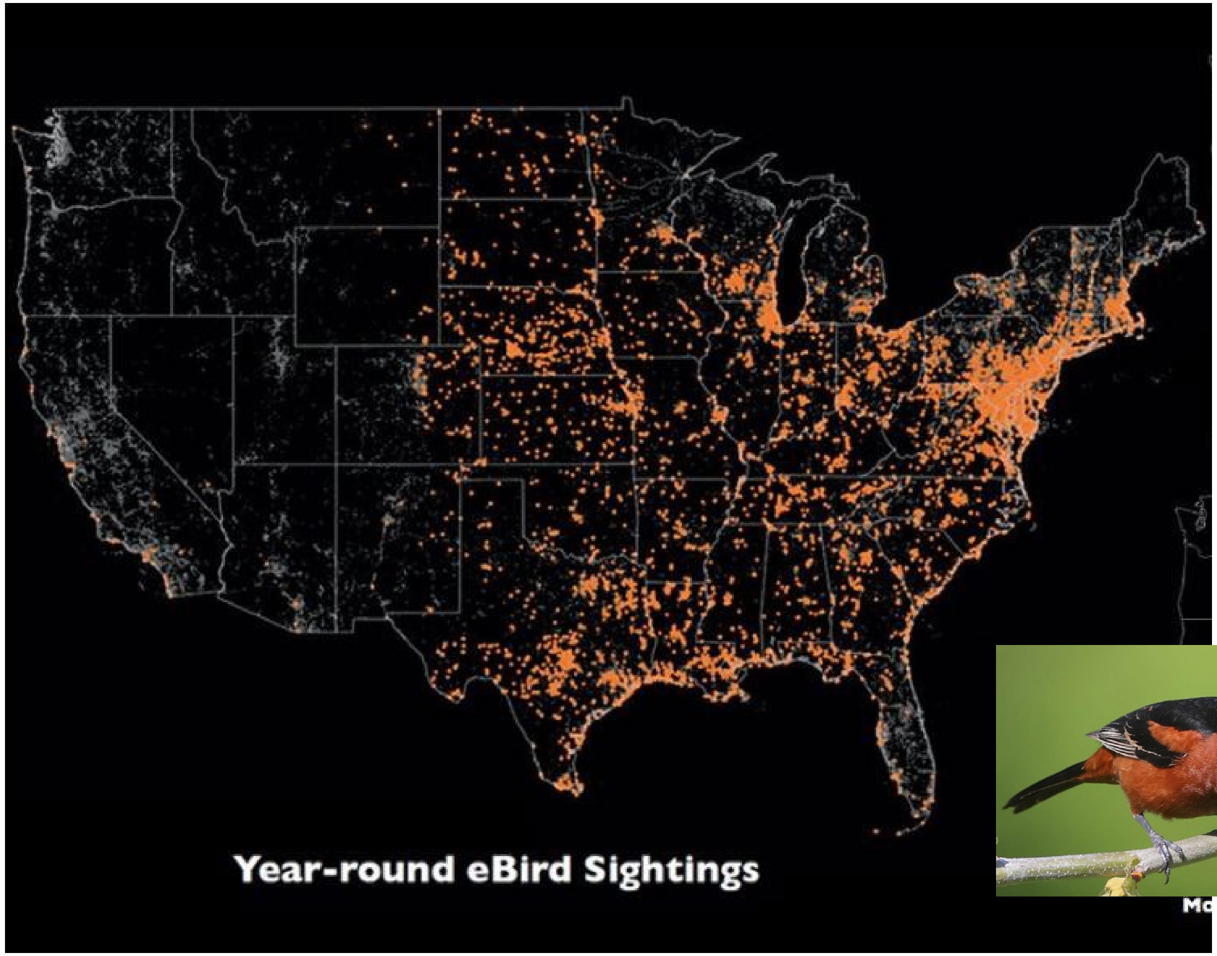
alabama cooperative fish and wildlife research unit

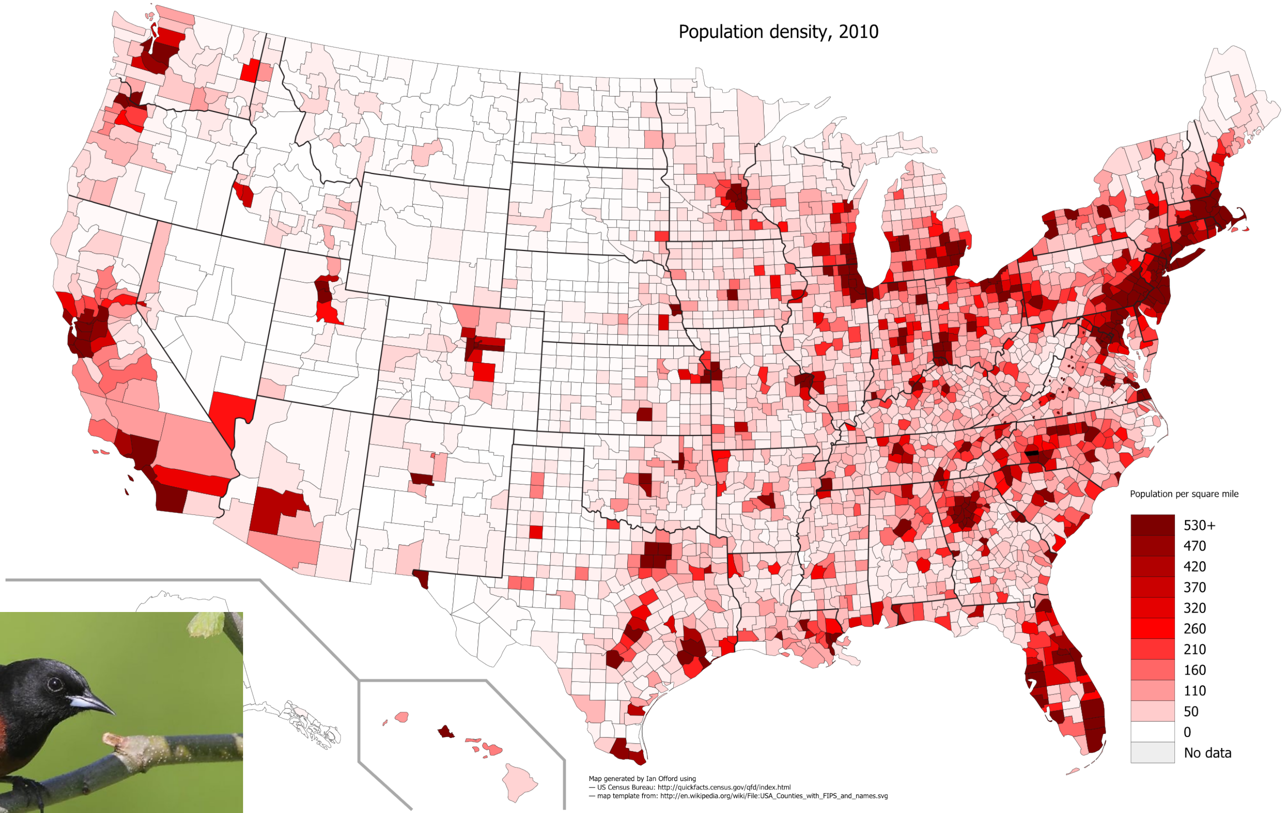
U.S. FISH & WILDLIFE SERVICE

Some caveats about presence-only models

- Temptation for overfitting
- Potential observation bias

Observation bias





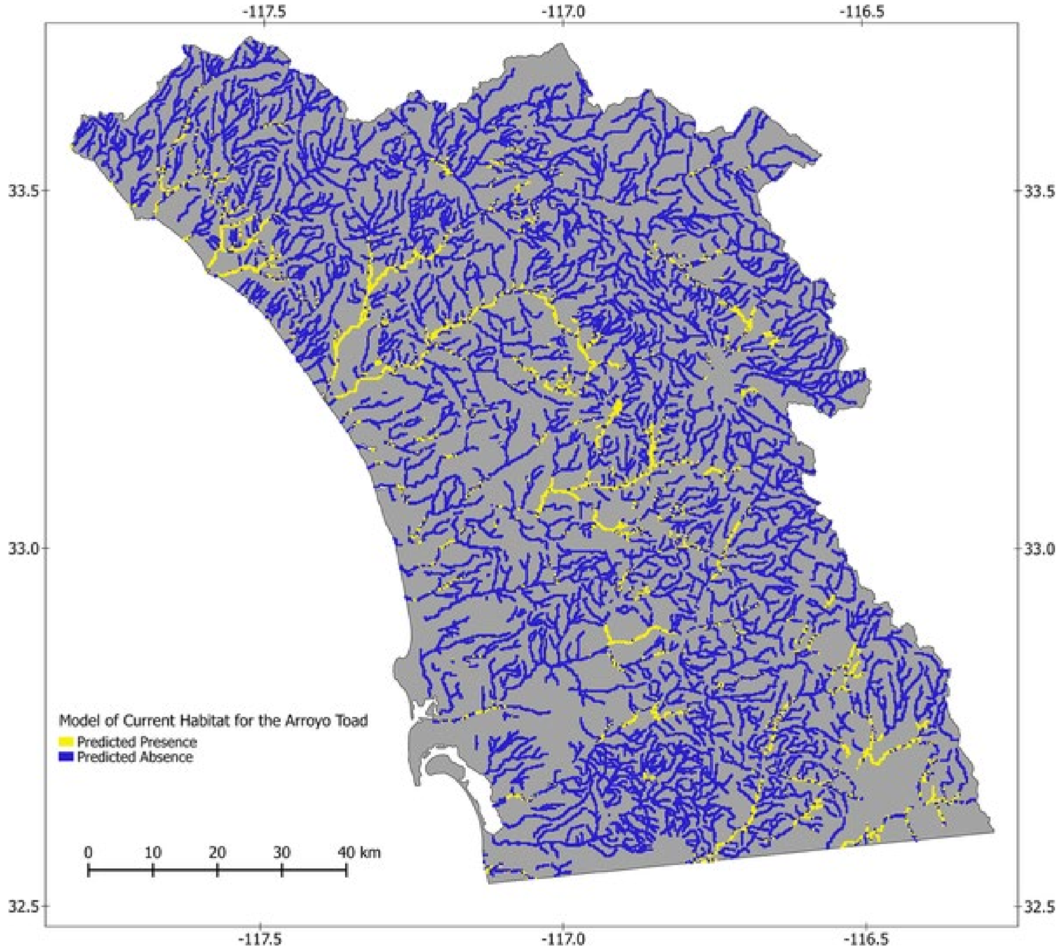
Population density, 2010 US Census

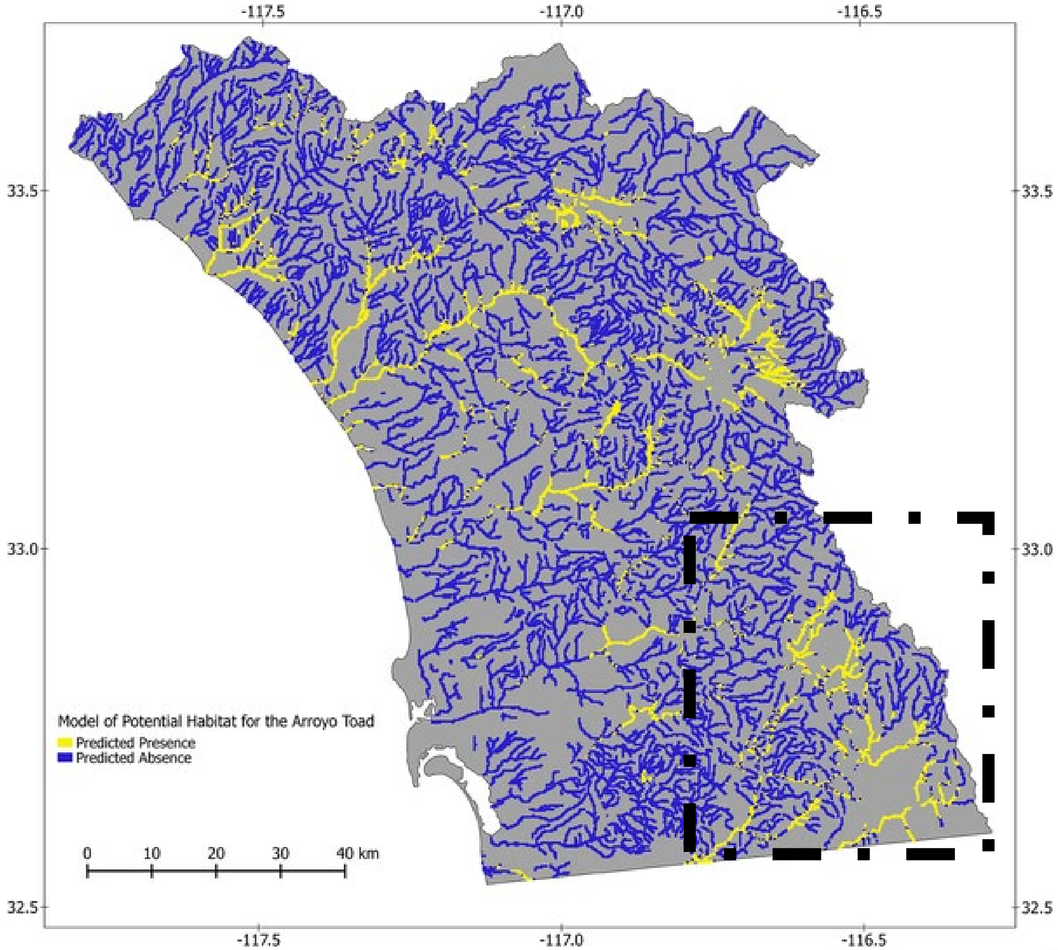
alabama cooperative fish and wildlife research unit

Some caveats about presence-only models

- Temptation for overfitting
- Potential observation bias
- Qualify any extrapolation you make
    - Potential translocation site, historic records from that area, etc.

Extrapolation





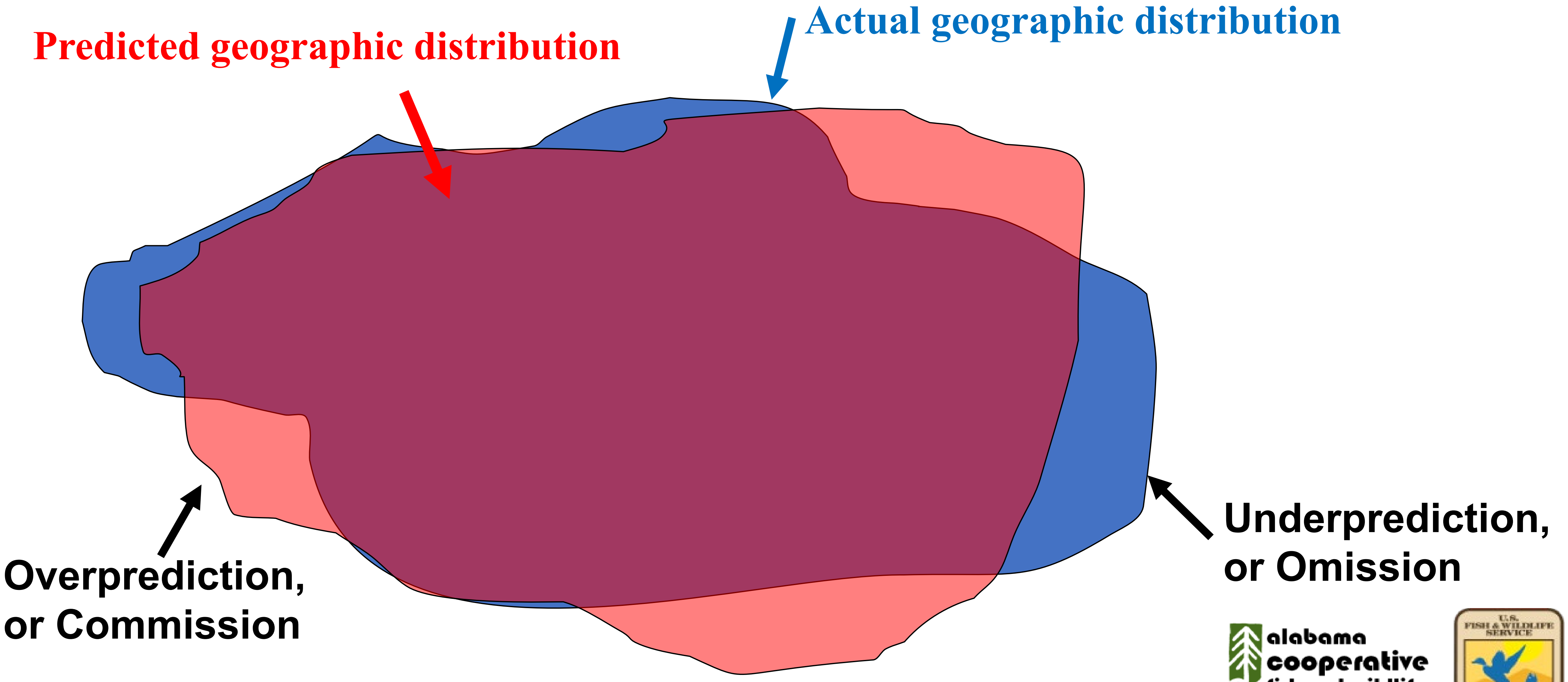
alabama cooperative fish and wildlife research unit

U.S. FISH & WILDLIFE SERVICE

Some caveats about presence-only models

- Temptation for overfitting
- Potential observation bias
- Qualify extrapolation
- Omission & commission errors
    - Area Under the receiving operator Curve (AUC) used to score model omission/commission (AUC > 0.9 considered good)

Omission and commission errors



Presence-absence data

Example data sources

- Information about where/when a species was detected AND where/when it was looked for but not found
  - Transect surveys
  - Point counts
  - Any other systematic survey effort
  - Complete eBird checklists
  - Others?

Example research questions

- What habitat characteristics are associated with species presence and absence? *(ecological needs, stressors)*
- What is the distribution of a species in a given area? *(Representation, Redundancy)*
  - How has that distribution shifted over time? (e.g. due to habitat loss, invasive species, etc.)
- What is the extent of the species range?
- To what extent does this species co-occur with other species?
  - Species interactions, exclusion, etc.
- How many species are found in this area?

alabama cooperative fish and wildlife research unit

U.S. FISH & WILDLIFE SERVICE

Common analysis approaches

- Logistic regression
  - **Assumes perfect detection**

Common analysis approaches

- Logistic regression
- Site-occupancy models

When is occupancy analysis appropriate?

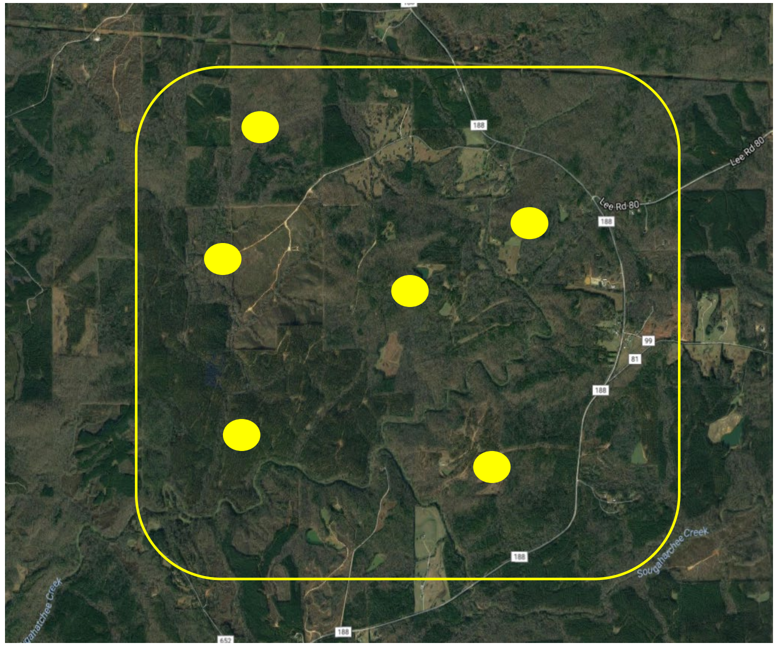
- Presence/absence data
- Multiple sites (*to estimate effects of ecological covariates*)
- Repeated visits in a **closed** period (*to estimate detection probability*)
    - Assume that true occupancy of a given site does not change between visits
    - Need to be collected within a short enough time-frame for this to be reasonable—depends on species of interest

alabama cooperative fish and wildlife research unit

U.S. FISH & WILDLIFE SERVICE

Sampling

- Replication is key
  - **Spatial** – multiple, randomly selected sites or sampling units within the area of interest
  - **Temporal** – repeated visits to each site



Visit 1

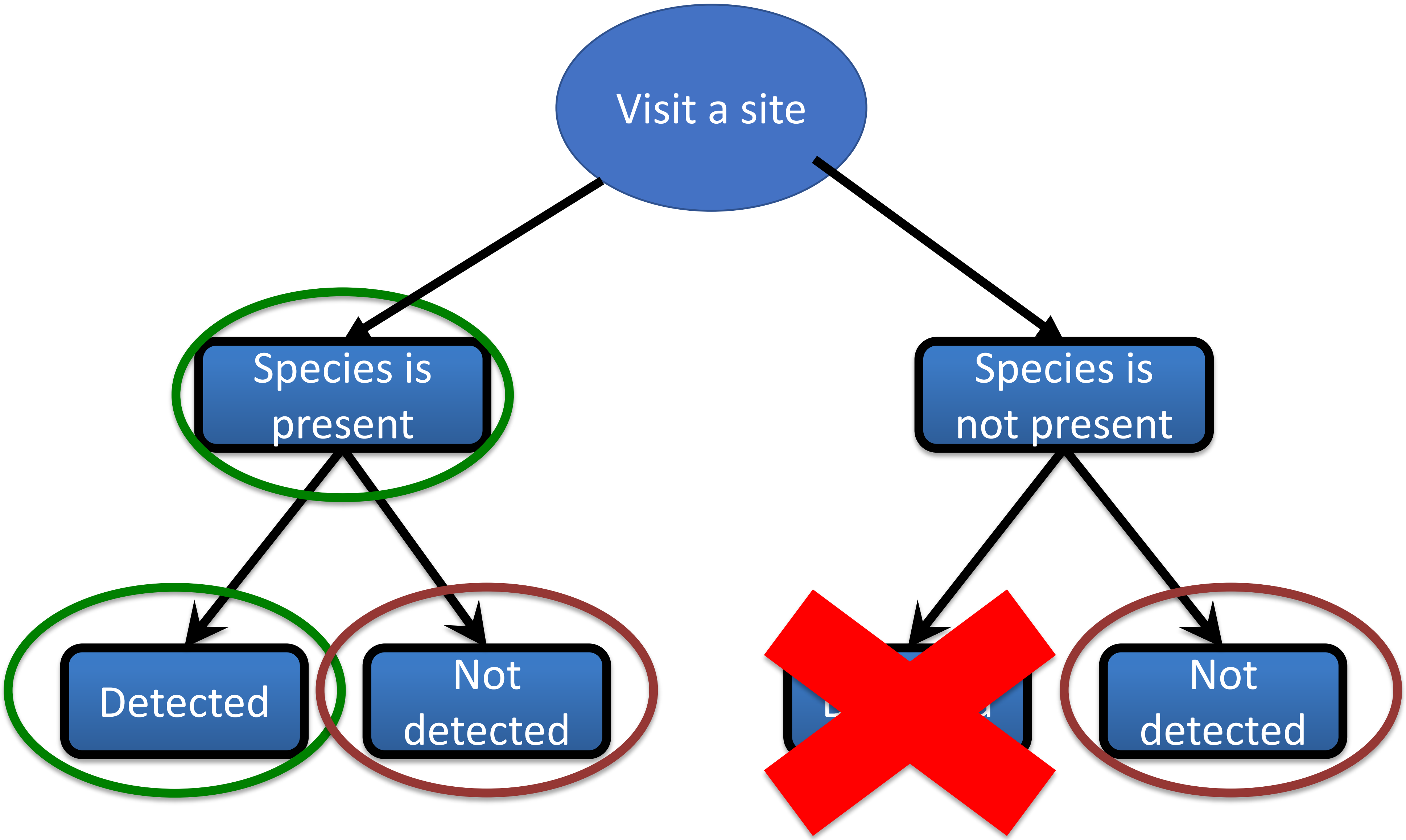
Visit 2

Visit 3

Example occupancy data set

Site	Date	Species detected?
A	6/27/2006	Yes
B	6/28/2006	Yes
C	6/25/2006	No
A	7/3/2006	No
B	7/5/2006	Yes
C	7/2/2006	No
A	7/12/2006	Yes
B	7/11/2006	Yes
C	7/13/2006	Yes

Site	Visit 1	Visit 2	Visit 3
A	1	0	1
B	1	1	1
C	0	0	1



alabama cooperative fish and wildlife research unit

U.S. FISH & WILDLIFE SERVICE

Site is **closed** during this time

True presence/absence of the species does not change

Site	Visit 1	Visit 2	Visit 3
A	1	0	1
B	1	1	0
C	0	1	1
D	0	0	0

Confirmed present

Likely absent?

$$\Pr(present|never\ detected) = (1 - 0.67)^3 = 0.04$$

Assumptions

- Sites are closed to changes in occupancy between sampling occasions
    - Appropriate duration between surveys
- Detection process is independent at each site
    - Appropriate distance between sites
- Both detection probability and occupancy probability are constant across all sites **OR** explained by covariates
    - For example, if we think rainfall influences our ability to detect the species, then rainfall should be included in the model

Model parameters

$\boldsymbol{\psi_i}$ = probability that site *i* is occupied

$\boldsymbol{p_{i,t}}$ = probability of detecting the species in site *i* at time *t*, given that the species is present

$\boldsymbol{\beta_k}$ = effect of covariate *k* on occupancy (or detection) probability

- Positive or negative?
- "significant" effect? – does the confidence interval contain 0?
- Importance of covariates often assessed by comparing models using AIC

* *Check text and captions for notation definitions within each paper – not always consistent!*

What influences occupancy probability?

- Potential stressors and threats included as **covariates**
  - Site characteristics (e.g. land cover, vegetation)
  - Weather (rainfall, temperature)
  - Distance to other occupied sites
  - … *etc.* …

- Effect of covariates often expressed as odds or odds ratios
  - Odds = how much more times as likely it is that a site is occupied for a one unit increase in the predictor

Parameter(predictor variables)

Relative support for each model (Relative to the model with the lowest AIC)
If ΔAIC > 2 then top model has the most support

Model weight – another way to assess relative support

Model	AIC	ΔAIC	Np	w_i
S(time) p(.)	684	0	5	0.98
S(.) p(.)	693	9	2	0.01
S(time) p(time)	698	14	10	0.01
S(time + sex) p(time)	710	26	12	0

Usually listed in decreasing order

Number of parameters (sometimes called *k*)

model (Table 2). Consistent with our predictions, this model indicated that per-visit detection probabilities were higher for conspecific surveys ($\hat{p} = 0.66$, SE = 0.03, 95% CI = 0.61–0.71) than for spotted owl surveys ($\hat{p} = 0.48$, SE = 0.04, 95% CI = 0.39–0.56) and that occupancy was positively influenced by the amount of public ownership in the sampling unit ($\beta = 4.67$, SE = 1.69, 95% CI = 1.36–8.00). Using single-visit estimates of detection probability from the best-supported model, the overall probability of

Table 2. Ranking of single-season occupancy models used to examine variation in the probability of detection (p) ar owls in western Oregon, USA, 2009.

Model[a]	No. parameters	AIC_c[b]	ΔAIC_c[b]
$\{\psi(\text{ownership})\ p(\text{survey type})\}$	4	776.28	0.00
$\{\psi(.)\ p(\text{survey type})\}$	3	782.48	6.20
$\{\psi(.)\ p(\text{stage + survey type})\}$	4	783.70	7.42
$\{\psi(\text{ownership})\ p(.)\}$	3	787.86	11.58
$\{\psi(.)\ p(\text{stage} \times \text{survey type})\}$	14	788.86	12.42
$\{\psi(.)\ p(t + \text{survey type})\}$	8	789.03	12.75
$\{\psi(.)\ p(.)\}$	2	793.14	16.86
$\{\psi(.)\ p(t \times \text{survey type})\}$	13	794.03	17.75
$\{\psi(.)\ p(\text{stage})\}$	3	794.86	18.58
$\{\psi(.)\ p(t + \text{stage})\}$	8	797.70	21.42
$\{\psi(.)\ p(t)\}$	7	799.72	23.44

Wiens et al 2011. Barred Owl Occupancy Surveys Within the Range of the Northern Spotted Owl. Journal of Wildlife Management

alabama cooperative fish and wildlife research unit

U.S. FISH & WILDLIFE SERVICE

Why estimate occupancy?

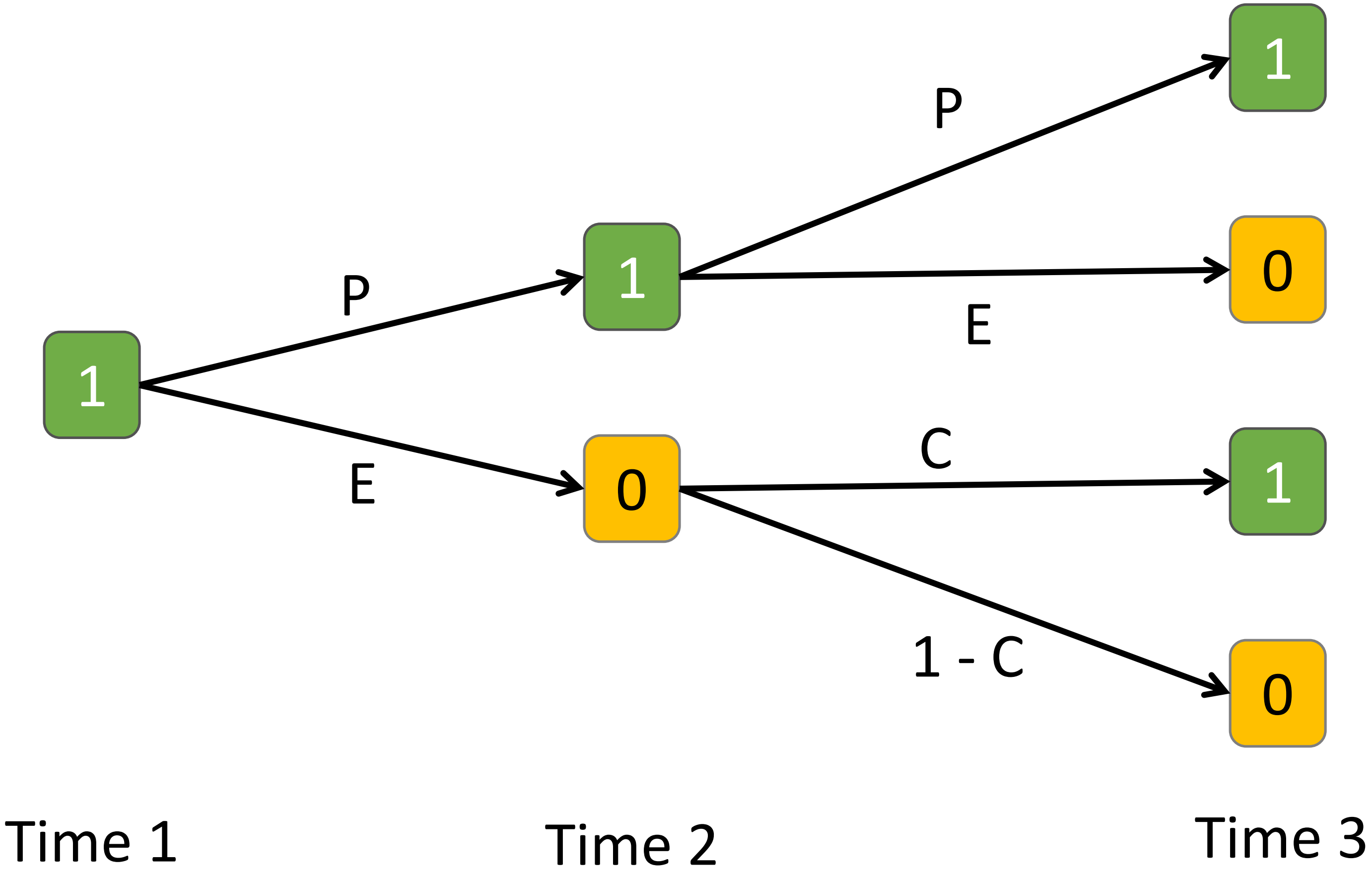
- Abundance can tell you more about species status, site use, and ecology
- More difficult to collect abundance data (time, money) and therefore often limited in spatial and/or temporal scope
- Presence/absence data is typically:
    - Less intensive to collect
    - Cheaper
    - Covers larger area/time scale
    - Can be more practical depending on objectives

Common analysis approaches

- Logistic regression
- Site-occupancy models
- Dynamic occupancy models

Dynamic occupancy models

- Estimate change in occupancy over time (colonization and extinction)



Dynamic occupancy models

	closed			*closed*			*closed*		
	Year 1			**Year 2**			**Year 3**		
Site	**Visit 1**	**Visit 2**	**Visit 3**	**Visit 1**	**Visit 2**	**Visit 3**	**Visit 1**	**Visit 2**	**Visit 3**
A	1	0	1	0	0	0	0	1	0
B	1	1	0	0	1	1	1	0	1
C	0	1	1	1	0	1	0	0	0
D	0	0	0	0	1	0	1	1	0

alabama cooperative fish and wildlife research unit

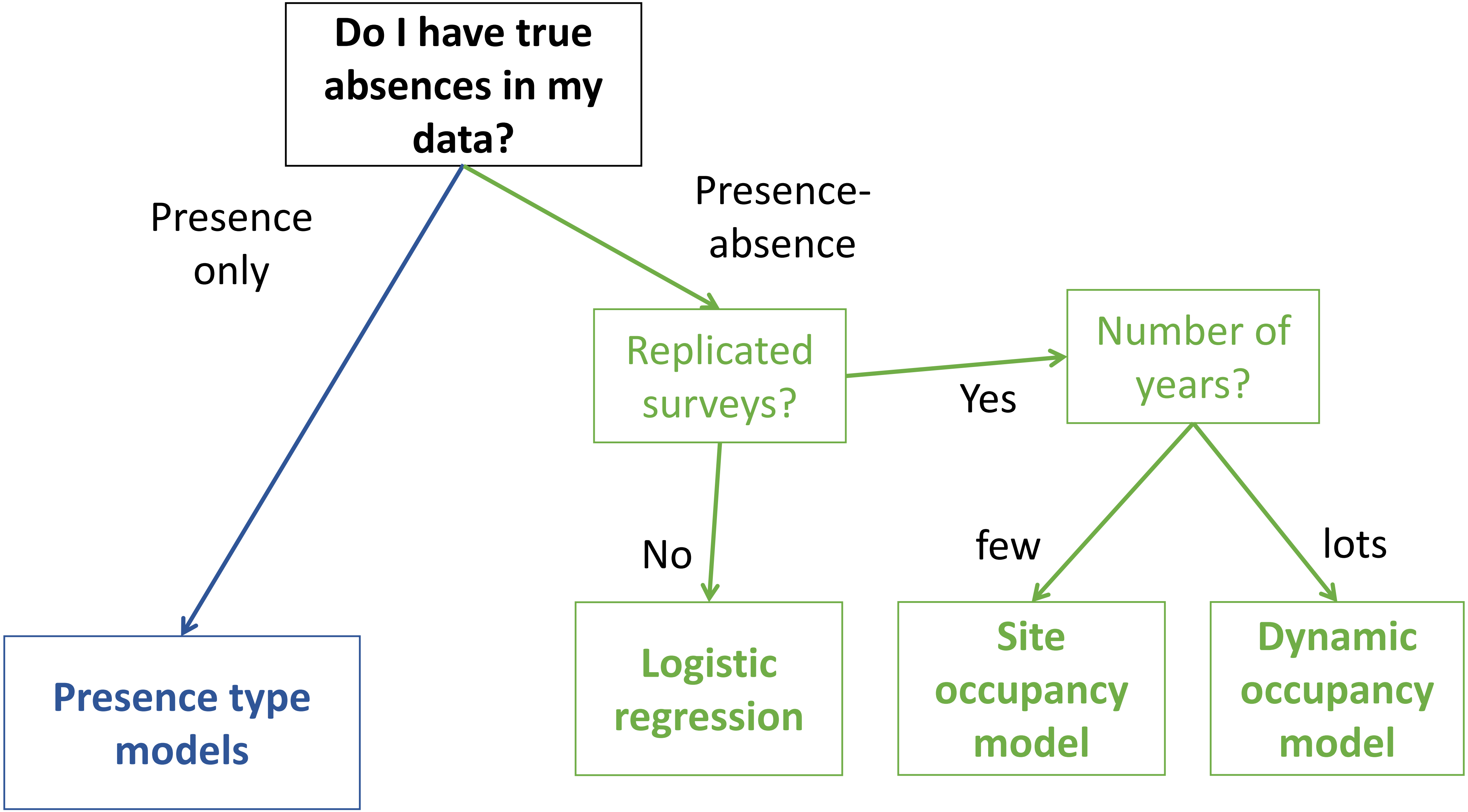
U.S. FISH & WILDLIFE SERVICE

Common analysis approaches

- Logistic regression – *lacking spatial and/or temporal replication*
- Site-occupancy models – *single year, several sites*
- Dynamic occupancy models – *several years, several sites*

Account for imperfect detection

Assume detection is perfect



Always check specific model assumptions!