Novel machine learning methods for learning models of bird distribution and migration from citizen science data

Tom Dietterich

Oregon State University

In collaboration with Selina Chu, Rebecca Hutchinson, Dan Sheldon, Michael Shindler, Weng-Keen Wong, Liping Liu, and the Cornell Lab of Ornithology

NICTA/ANU May 2012

Bird Distribution and Migration

- Management:
  - Many bird populations are declining
  - Predicting aircraft-bird interactions
  - Siting wind farms
  - Night-time lighting of buildings (esp. skyscrapers)
  - How will climate change affect bird migration and survival?

- Science:
  - What is the migration decision making policy for each species
    - When to start migrating?
    - How far to fly each night?
    - When to stop over and for how long?
    - When to resume flying?
    - What route to take?

Why bird migration is poorly understood

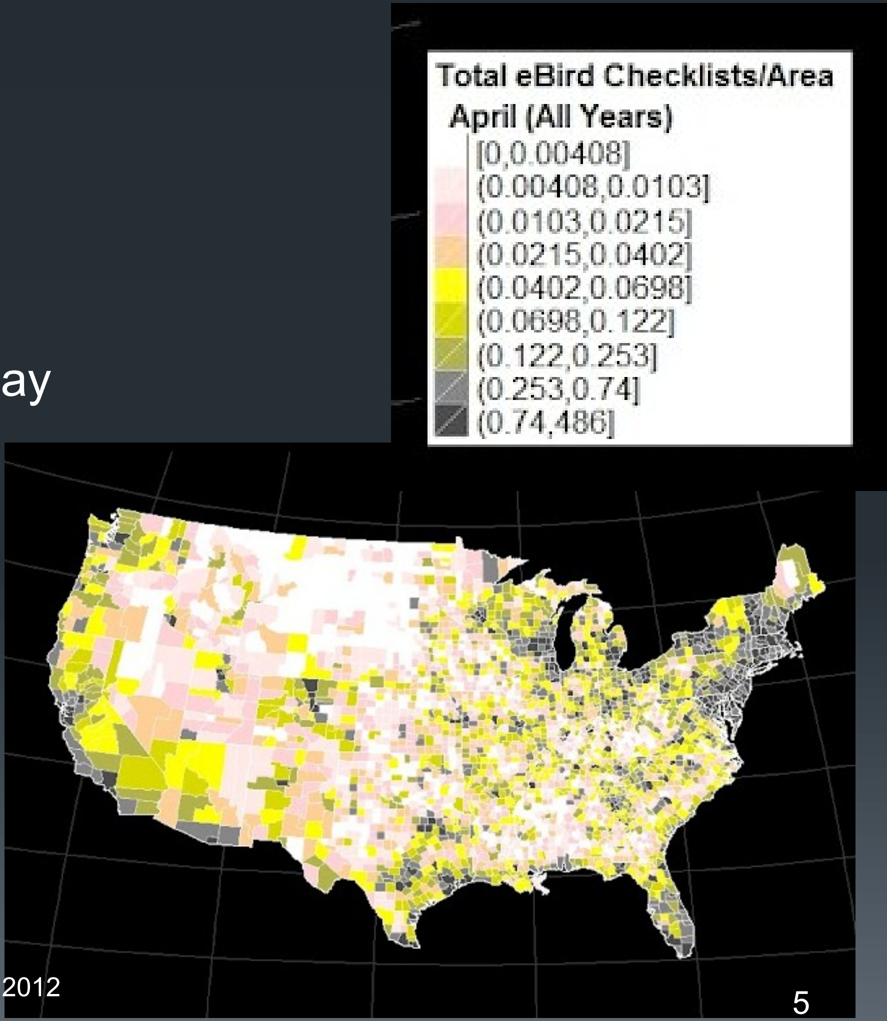
- It is difficult to observe
  - Takes place at continental scale (and beyond)
  - Impossible for the small number of professional ornithologists to collect enough observations
  - Very few birds have been individually tracked

What Data Are Available?

- Birdwatcher count data: eBird.org
- Doppler weather radar
- Night flight calls

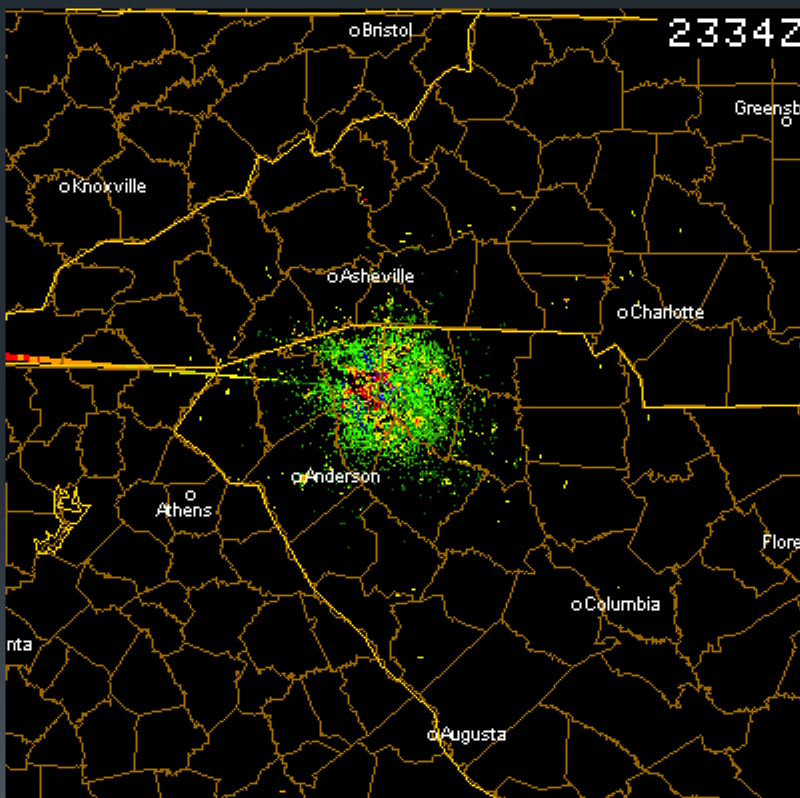
eBird Data

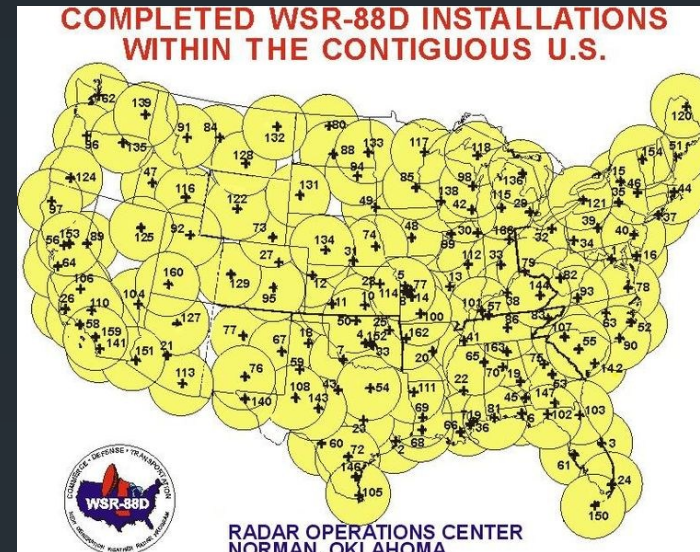
- Bird watchers record their observations in a database through eBird.org.
  - “Citizen Science”
- Dataset available for analysis
- Features
  - LOTS of data!
    - ~3 million observations reported last May
  - All bird species (~3,000)
  - Year-round
  - Continent-scale
- Challenges
  - Variable quality observations
  - No systematic sampling design



Doppler Weather Radar

- Weather radar detects migrating birds

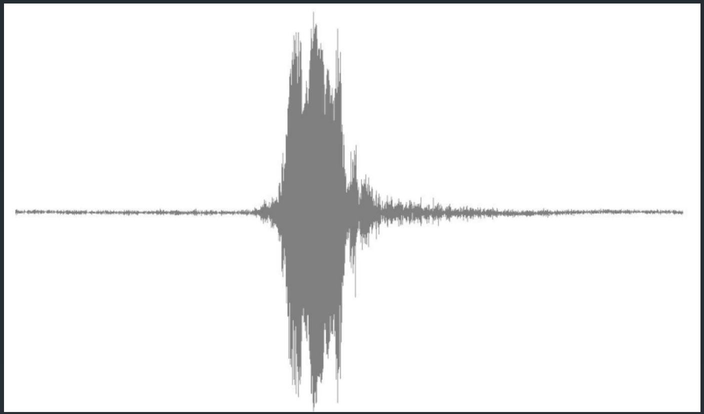




- Can estimate total biomass
- No species information
- Archived data available back to 1995

Night Flight Calls

- Many species of migrating birds emit flight calls that can be identified to species or species group
- New project at Cornell to roll out a large network of recording stations
- Automated detection and classification
- DTW kernel
  - Damoulas, et al, 2010
  - Results on 5 species
  - Clean recordings



Classifier	*Feature Extraction Method*	10×10CV %
J48	$\text{DTW}_{\text{global}}$	87.1 ± 1.14
Kstar	$\text{DTW}_{\text{global}}$	$\mathbf{96.6 \pm 0.65}$
BayesNet	$\text{DTW}_{\text{global}}$	93.2 ± 0.27
Simple Logistic	$\text{DTW}_{\text{global}}$	94.9 ± 0.55
Decision Table	$\text{DTW}_{\text{global}}$	72.8 ± 3.82
Random Forest	$\text{DTW}_{\text{global}}$	93.2 ± 0.84
Logit Boost	$\text{DTW}_{\text{global}}$	91.7 ± 1.64
Rotation Forest	$\text{DTW}_{\text{global}}$	94.5 ± 1.06
$\text{SVM}^{\text{multiclass}}$	$\text{DTW}_{\text{global}}$ Kernel	95 ± 0.43
VBpMKL	$\text{DTW}_{\text{global}}$ Kernel	$\mathbf{97.6 \pm 0.68}$

Prediction Tasks

- Species Distribution Models
  - Given site described by feature vector x
  - Predict whether a target species s will be present $y = 1$
    - At a particular point in time
    - At any time throughout the year
- Bird Migration Models
  - Given observations from ebird, radar, flight calls
    - Reconstruct migration behavior
  - Given observations + weather forecast
    - Predict migration behavior for next 24 hours, next 5 days

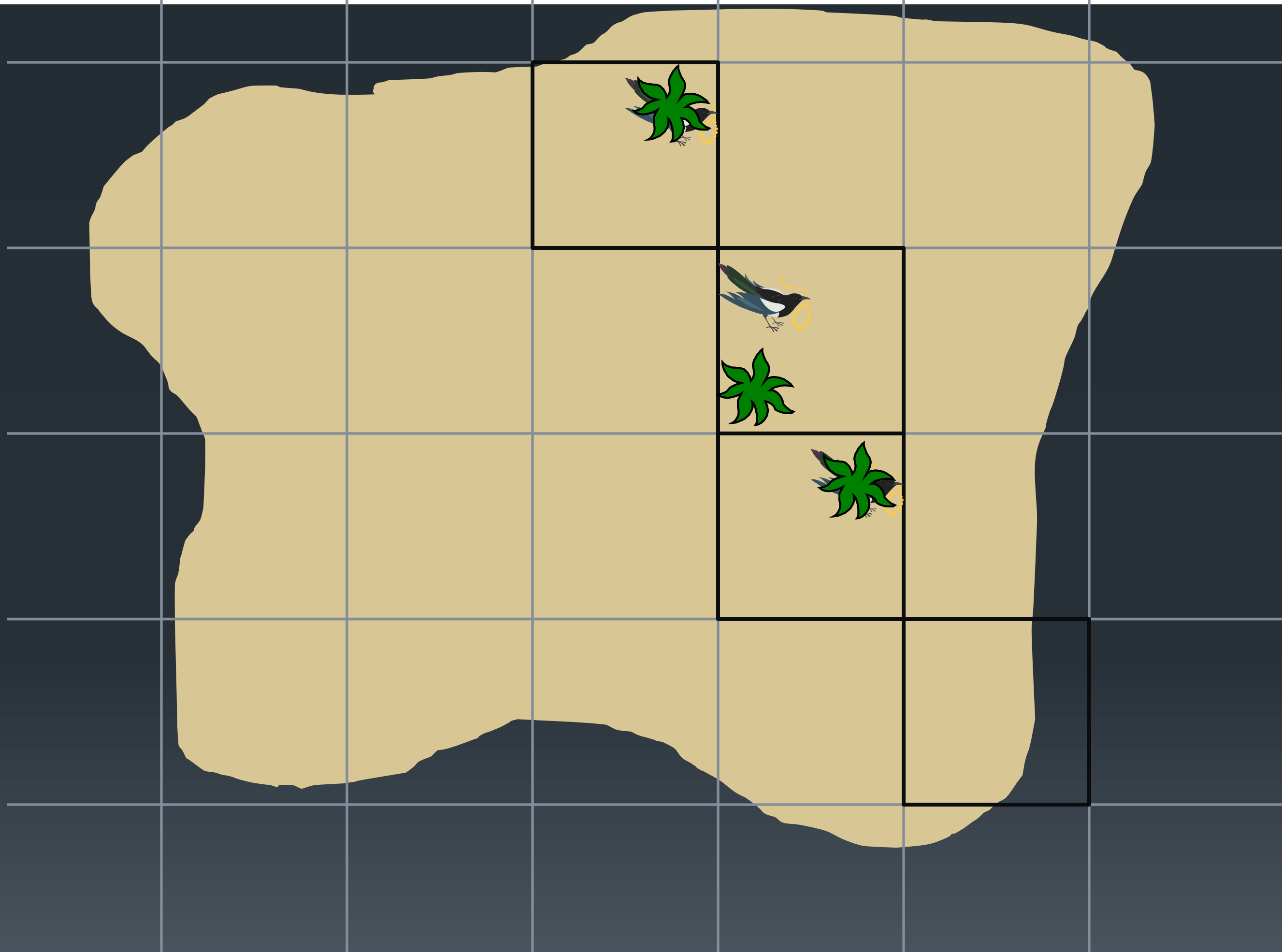
Species Distribution Model Challenges

1. Partial Detection
   - Observer may not detect the species even though it is present
2. Observer Expertise
   - Observer may not recognize the species even though it is detected
3. Sampling Bias
   - Birders choose where and when to observe
4. Population Size Effects
   - Bird population may be too small to occupy all suitable habitat
     - Unoccupied and occupied sites may be identical
5. Spatial Dynamics
   - In order to occupy habitat, the birds must discover it, so it needs to be accessible
6. Spatial and Temporal Dynamics of other species
   - Food: insect and plant species
   - Competitors/Predators

1. Imperfect Detection

Problem: Some birds are hidden

Pa nt birds hide on different visits

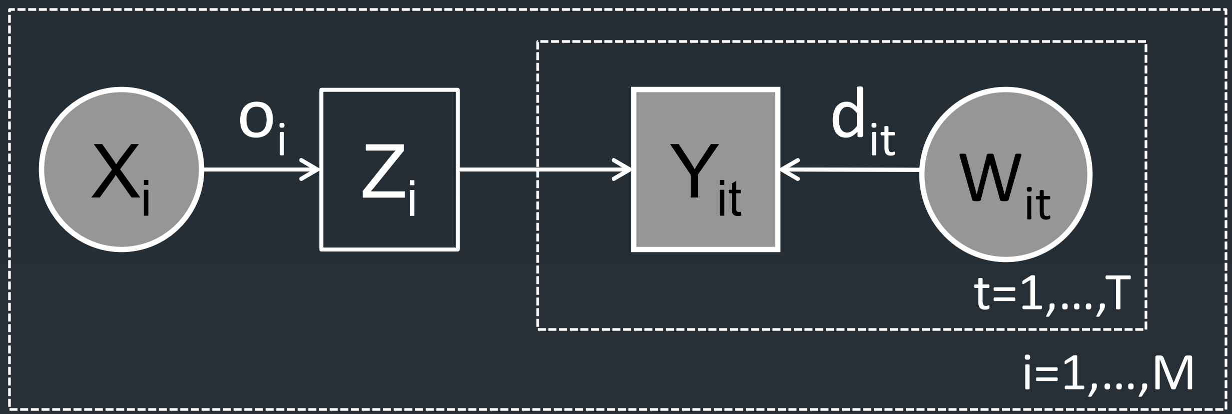


Multiple Visit Data

		Detection History		
Site	*True occupancy (latent)*	Visit 1 (rainy day, 12pm)	Visit 2 (clear day, 6am)	Visit 3 (clear day, 9am)
A (forest, elev=400m)	*1*	0	1	1
B (forest, elev=500m)	*1*	0	1	0
C (forest, elev=300m)	*1*	0	0	0
D (grassland, elev=200m)	*0*	0	0	0

Occupancy-Detection Model

MacKenzie, et al, 2006



$z_i \sim P(z_i|x_i)$: Species Distribution Model

$P(z_i = 1|x_i) = o_i = F(x_i)$ "occupancy probability"

$y_{it} \sim P(y_{it}|z_i, w_{it})$: Observation model

$P(y_{it} = 1|z_i, w_{it}) = z_i d_{it}$

$d_{it} = G(w_{it})$ "detection probability"

The Power of Probabilistic Graphical Models

- Probabilistic graphical models have many advantages
  - Excellent language for representing models
  - Learning and reasoning via probabilistic inference
  - Support hidden (latent) variables

- However, they have disadvantages
  - Designer must choose the parametric form of each probability distribution
  - Must decide on the number and form of interactions
  - Data must be scaled and transformed to match model assumptions
  - Somewhat difficult to adapt the complexity of the model to the amount and complexity of the data

Important Contribution of Machine Learning: Flexible Models

- Classification and Regression Trees
  - Require no model design
  - Require no data preprocessing or transformation
  - Automatically discover interactions as needed
  - Achieve high accuracy via ensembles

- Support Vector Machines
  - Still require data preprocessing and transformation
  - Powerful methods for tuning model complexity automatically

Goal: Combine Probabilistic Graphical Models with Flexible Models

- Major open problem in machine learning

- Current efforts:
  - Kernel (SVM) methods for computing with probability distributions
  - Bayesian Non-Parametric Models: Dirichlet process mixture models

- Our approach: **Boosted regression trees**
  - Represent F and G using weighted sums of regression trees
  - Learn them via boosting
  - This can be done using functional gradient descent (Mason & Bartlett, 1999; Friedman, 2000; Dietterich, et al, 2008; Hutchinson & Dietterich, 2011)

L2-Tree Boosting Algorithm
(Friedman 2000)

- Let $F^0(X) = f^0(X) = 0$ be the zero function
- For $\ell = 1, \ldots, L$ do
  - Construct a training set $S^\ell = \{(X^i, \tilde{Y}^i)\}_{i=1}^N$
    - where $\tilde{Y}$ is computed as
    - $\tilde{Y}^i = \left.\frac{\partial LL(F)}{\partial F}\right|_{F=F^{\ell-1}(X^i)}$ how we wish F would change at X^i
  - Let f^ℓ = regression tree fit to S^ℓ
  - $F^\ell := F^{\ell-1} + \eta_\ell f^\ell$
- The step sizes η_ℓ are the weights computed in boosting
- This provides a general recipe for learning a conditional probability distribution for a Bernoulli or multinomial random variable

Alternating Functional Gradient Descent

- Loss function $L(F, G, y)$
- $F^0 = G^0 = f^0 = g^0 = 0$
- For $\ell = 1, \dots, L$
  - For each site i compute
    $$\tilde{z}_i = \partial L(F^{\ell-1}(x_i), G^{\ell-1}, y_i) / \partial F^{\ell-1}(x_i)$$
  - Fit regression tree f^ℓ to $\{\langle x_i, \tilde{z}_i \rangle\}_{i=1}^{M}$
  - Let $F^\ell = F^{\ell-1} + \rho_\ell f^\ell$
  - For each visit t to site i, compute
    $$\tilde{y}_{it} = \partial L\left(F^\ell(x_i), G^{\ell-1}(w_{it}), y_{it}\right) / \partial G^{\ell-1}(w_{it})$$
  - Fit regression tree $g^\ell$to $\{\langle w_{it}, \tilde{y}_{it} \rangle\}_{i=1,t=1}^{M,T_i}$
  - Let $G^\ell = G^{\ell-1} + \eta_\ell g^\ell$

Hutchinson, Liu, Dietterich, AAAI 2011

Experiment

- Algorithms:
  - Supervised methods:
    - S-LR: logistic regression from $(x_i, w_{it}) \rightarrow y_{it}$
    - S-BRT: boosted regression trees $(x_i, w_{it}) \rightarrow y_{it}$
  - Occupancy-Detection methods:
    - OD-LR: F and G logistic regressions
    - OD-BRT: F and G boosted regression trees
- Data:
  - 12 bird species
  - 3 synthetic species
  - 3124 observations from New York State, May-July 2006-2008
  - All predictors rescaled to zero mean, unit variance

Synthetic Species

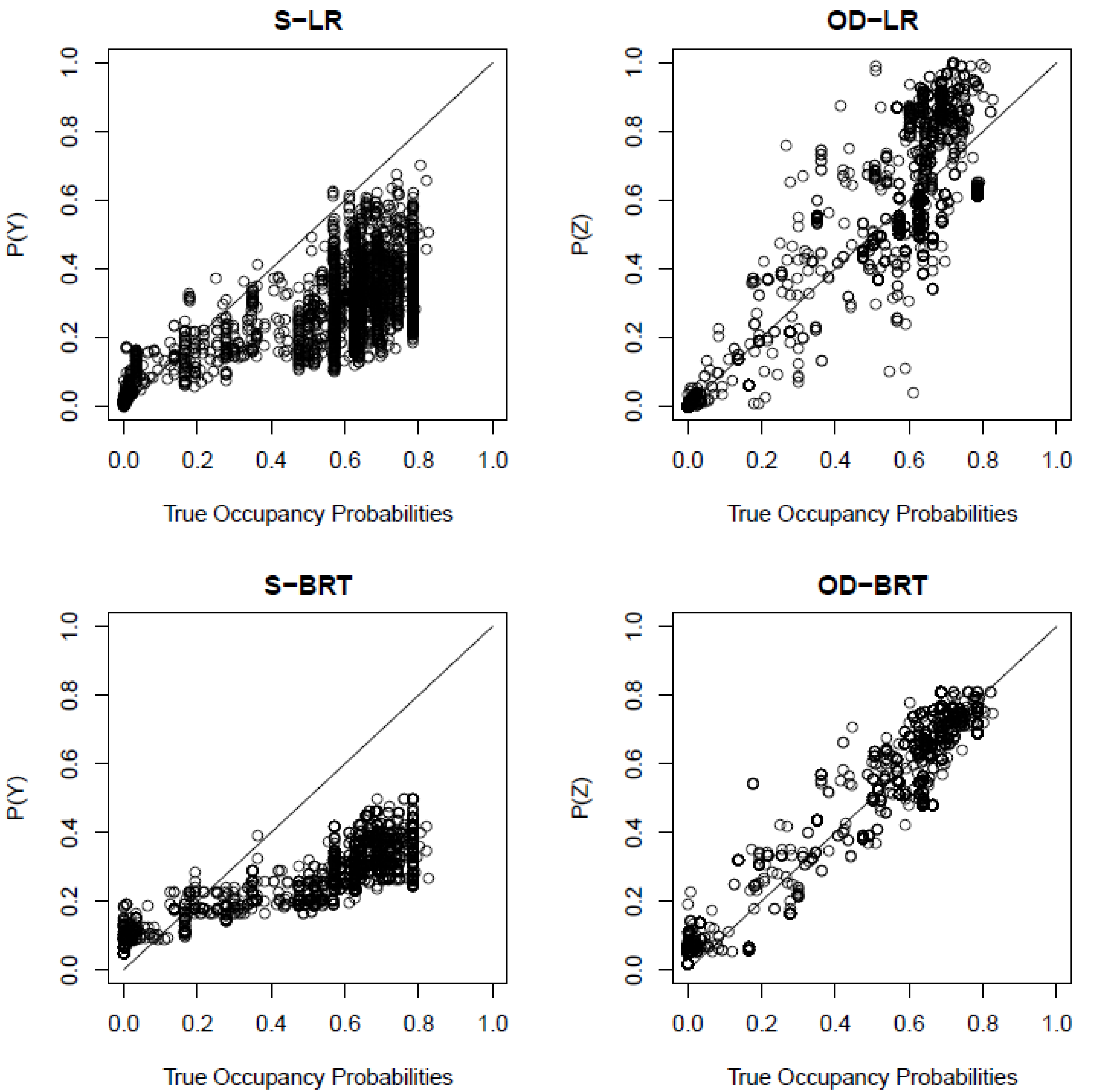
- Synthetic Species 2: F and G nonlinear

$$o_i \propto \exp\left(-2\left[x_i^{(1)}\right]^2 + 3\left[x_i^{(2)}\right]^2 - 2x_i^{(3)}\right)$$

$$d_{it} \propto \exp\left(\exp\left(-0.5w_{it}^{(4)}\right) + \sin\left(1.25w_{it}^{(1)} + 5\right)\right)$$

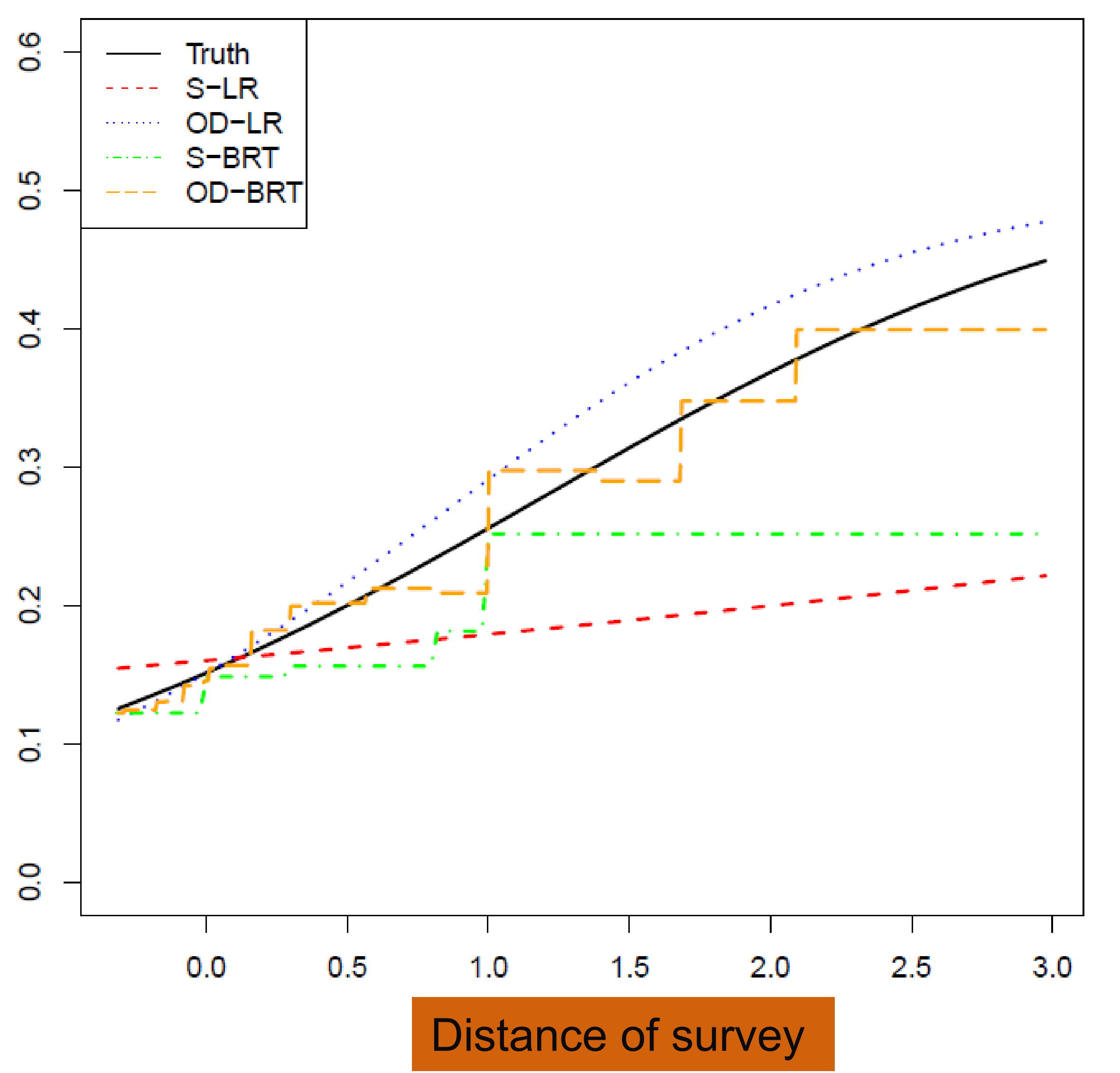
Predicting Occupancy

Synthetic Species 2



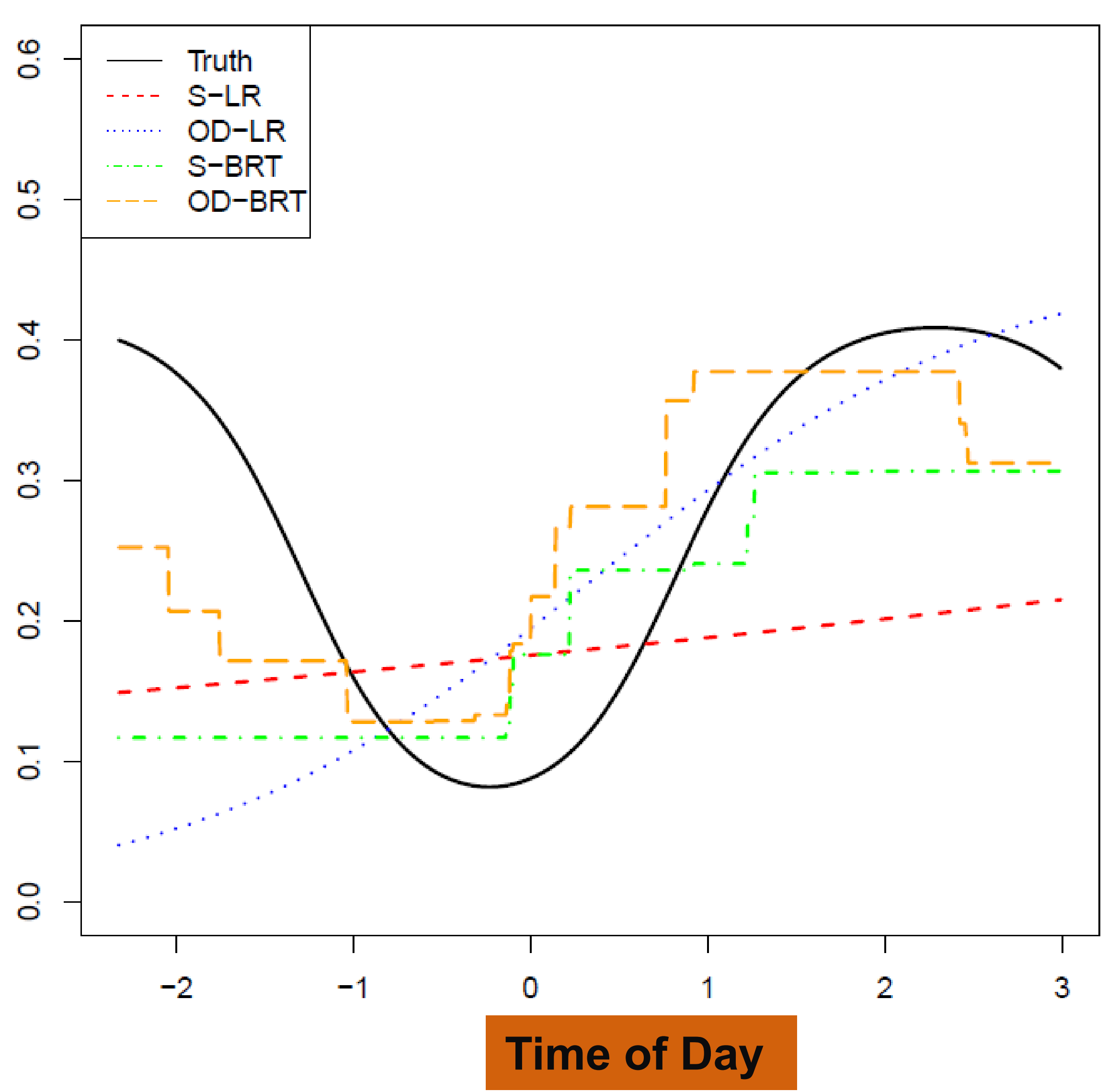
Partial Dependence Plot
Synthetic Species 1

- OD-BRT has the least bias

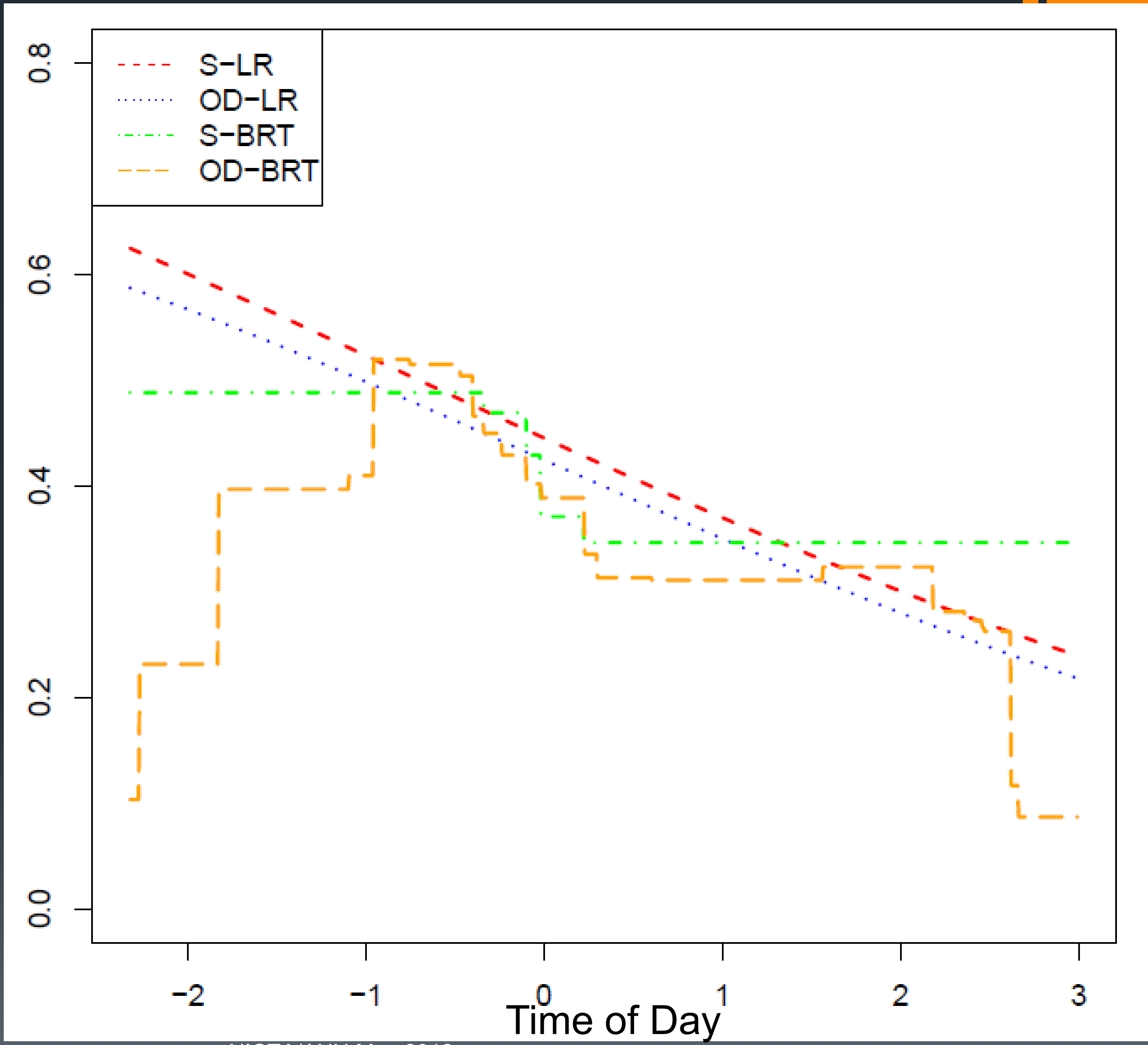


Partial Dependence Plot
Synthetic Species 3

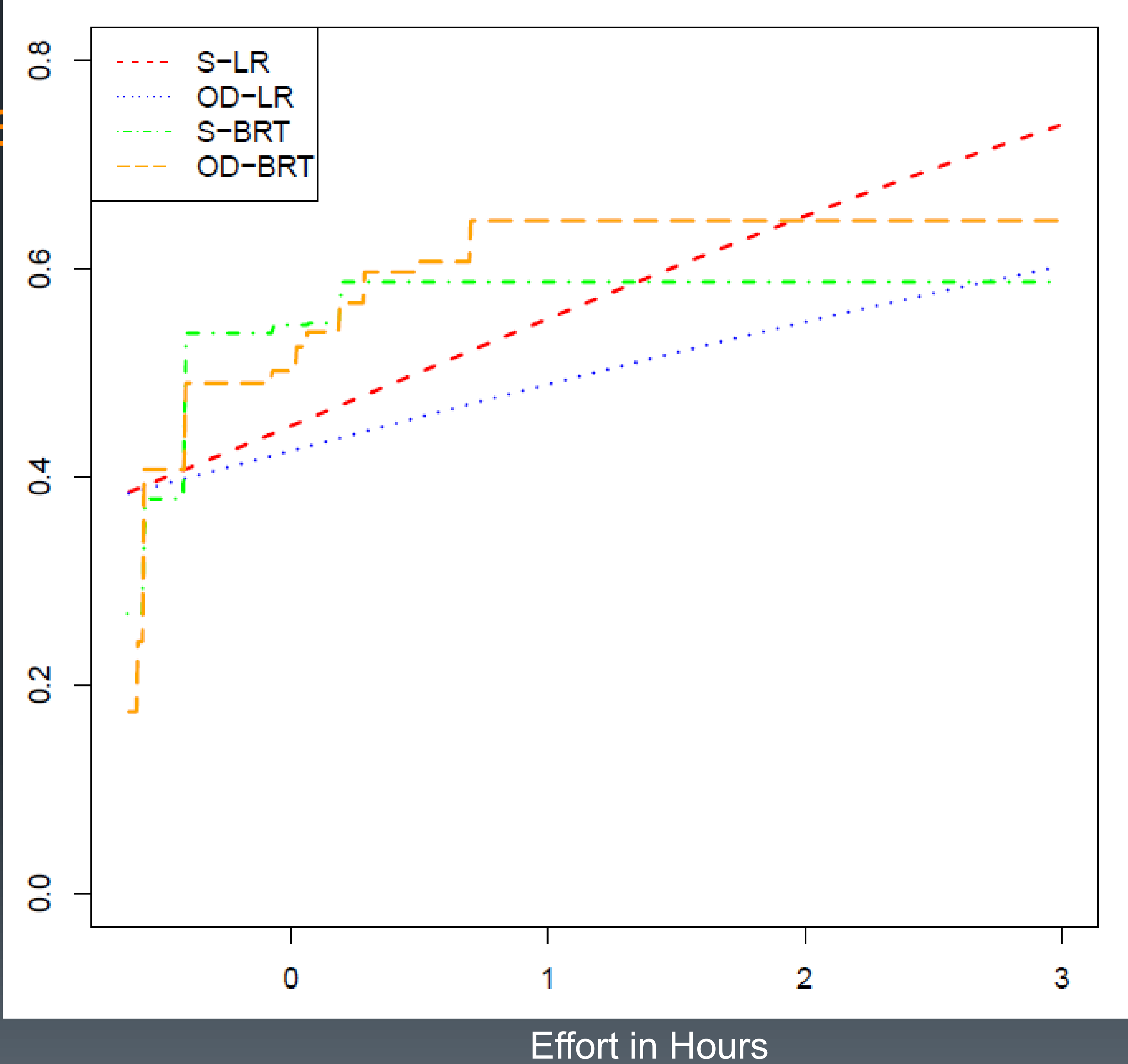
- OD-BRT has the least bias and correctly captures the bi-modal detection probability



Partial Dependence Plot Blue Jay vs. Time of Day



Partial Dependence Plot Blue Jay vs. Duration of Observation

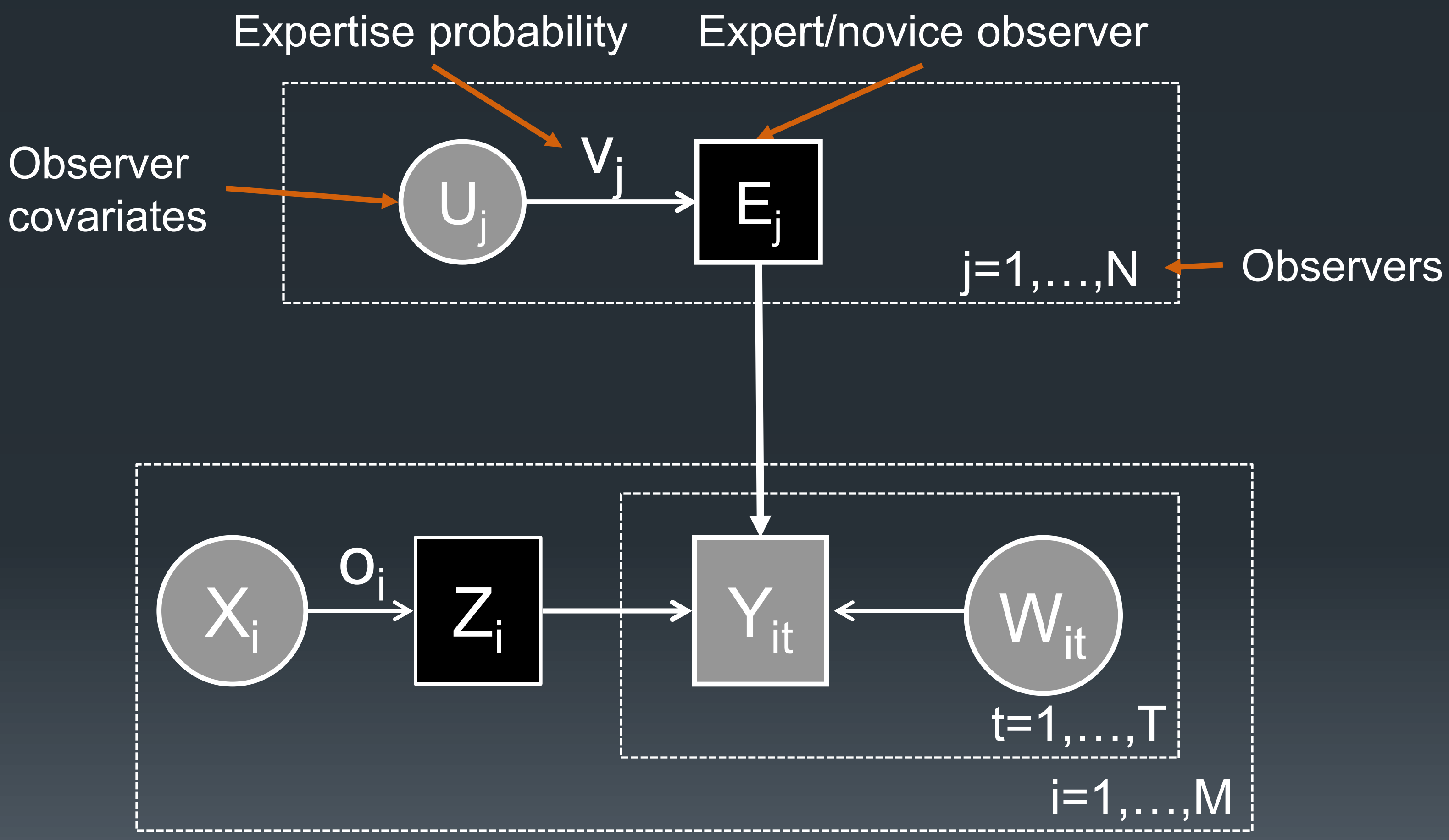


2. Variable Expertise

- Problem: expert and novice observers contributing observations to citizen science data generate different mistakes/biases

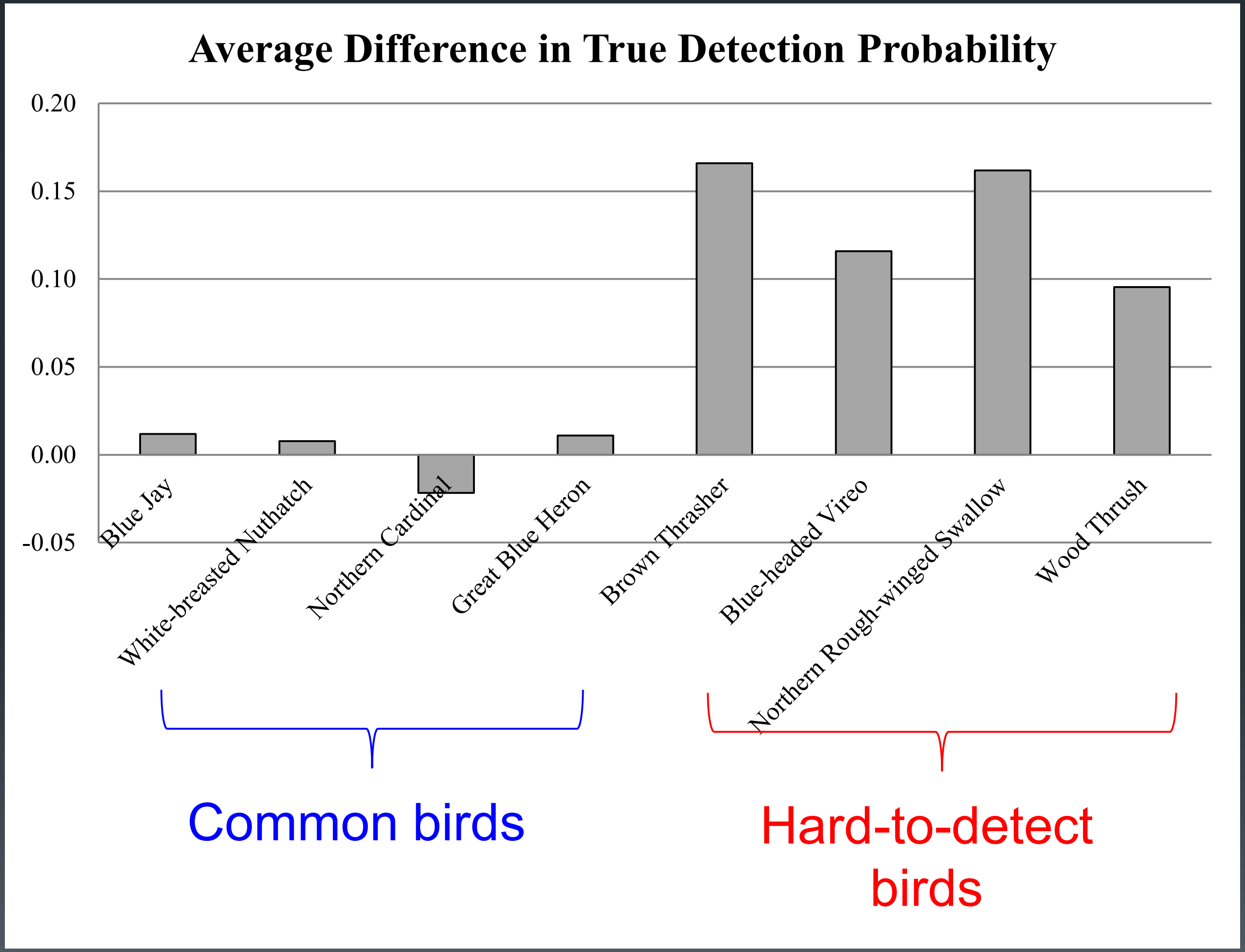
- Solution: extend occupancy models so that observer expertise affects the detection model

Occupancy-Detection-Expertise (ODE) model



Yu, et al, 2010

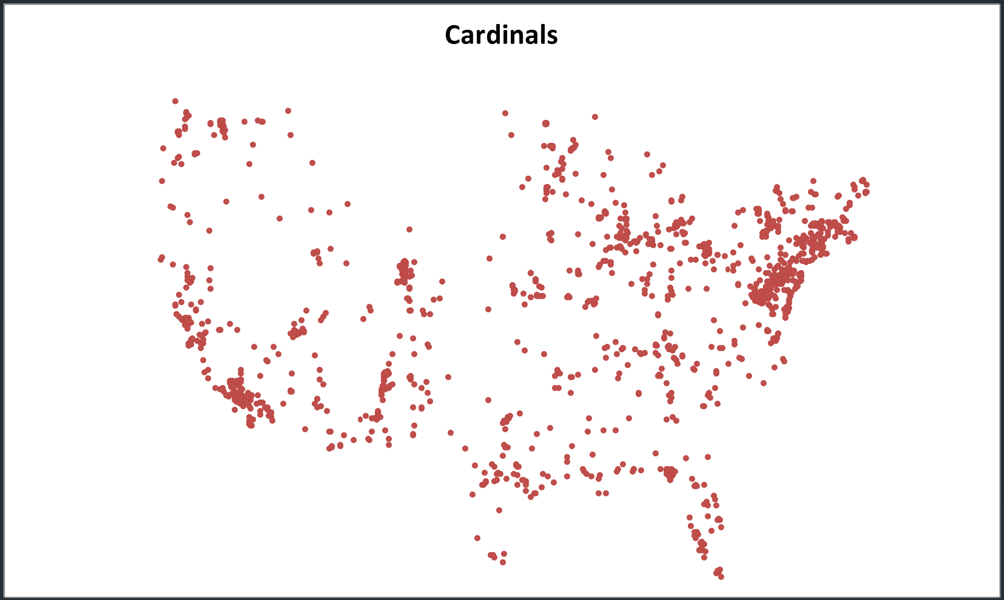
Expert vs Novice Differences



Yu, et al, 2010

3. Sample Selection Bias

- Citizen scientists tend to stay close to home
- How can we make good predictions across the whole US?



Distribution of check lists mentioning explicit presence or absence of Cardinal

Covariate Shift Reweighting

- Distribution of training data: $P_{train}(x)$
- Target test distribution $P_{test}(x)$ is uniform
- Reweight training examples according to

$$r(x) = \frac{P_{test}(x)}{P_{train}(x)}$$

- Fit classifier to weighted training data

Density Estimation

- Assume $x \in \mathbb{R}^d$ d-dimensional Euclidean space
- Let v be a volume of $\mathbb{R}^d$
- $P_{train}(x|v) = \frac{N_v}{N|v|}$
- Volume is a tricky concept
  - Effective dimension of the data may be much less than d
  - Sample complexity of scales with the dimension

Direct Density Ratio Estimation
(Sugiyama et al., 2011, 2012)

- Direct density ratio estimation

$$r(x) = \frac{P_{test}(x)}{P_{train}(x)} = \frac{N_{test}(v)}{N_{test}|v|} \cdot \frac{N_{train}|v|}{N_{train}(v)} = \frac{N_{test}(v)}{N_{test}} \cdot \frac{N_{train}}{N_{train}(v)}$$

- The volumes cancel

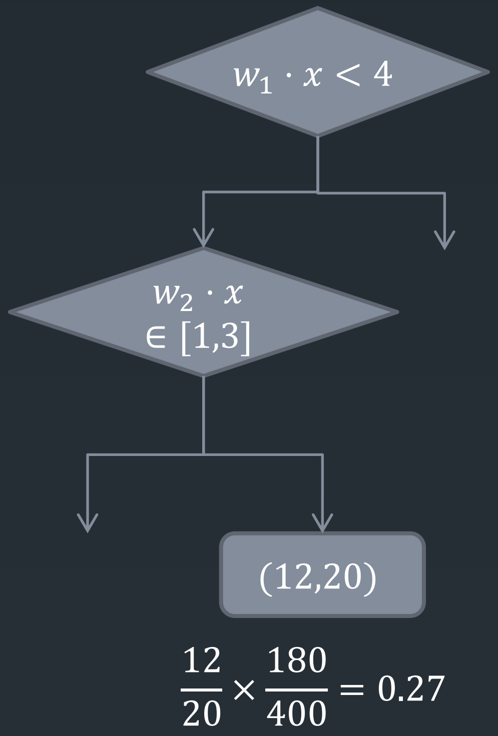
Random Projection Trees for Direct Density Estimation

- RP-Trees (Dasgupta & Freund, STOC 2008)
  - Project training data onto random vector
  - Two kinds of splits:
    - Split by perpendicular bisector randomized near the median of the data
    - Split by an interval centered on the median (tails to the left, center to the right)
  - Guarantees that the tree “follows” the data
  - Scales with the true dimensionality of the data, rather than the apparent dimensionality d

Algorithm Idea

- Given
  - N_{train} points sampled from P_{train}
  - N_{test} points sampled from P_{test}

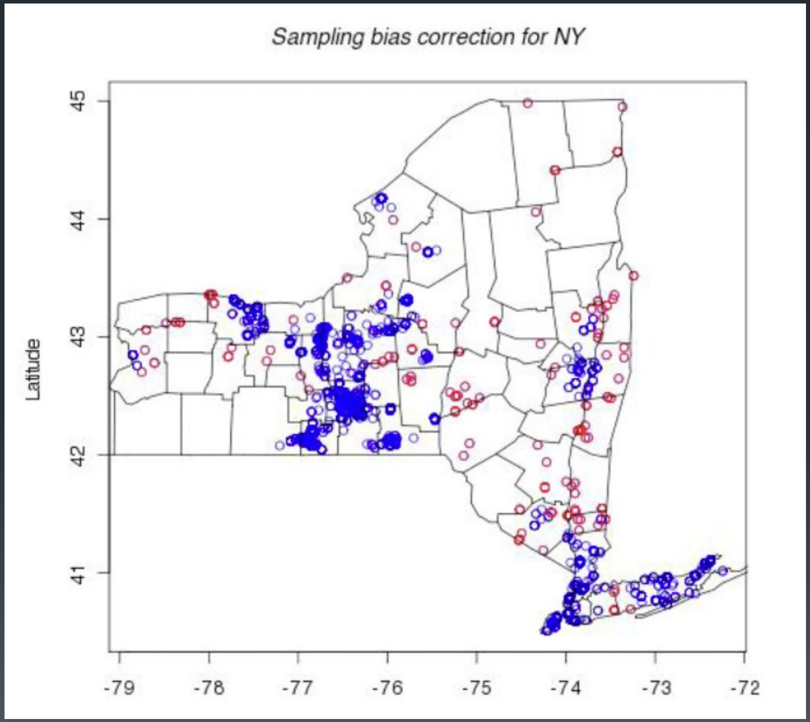
- Build an RP tree using the N_{train} data points
- Drop the N_{test} data points through the tree
- Prune the tree so that each leaf ℓ contains at least
  - N_{min} data points, and
  - $r_{min} \leq \frac{N_{test}(\ell) N_{train}}{N_{train}(\ell) N_{test}} \leq \frac{1}{r_{min}}$
- Combine in large ensemble
- Conjectures
  - Consistent: $\hat{r}(x) \rightarrow r(x)$ as sample sizes $\rightarrow \infty$
  - Generalization bounds on $||\hat{r}(x) - r(x)||^2$ that depend only on true dimension of data



Results: None Yet

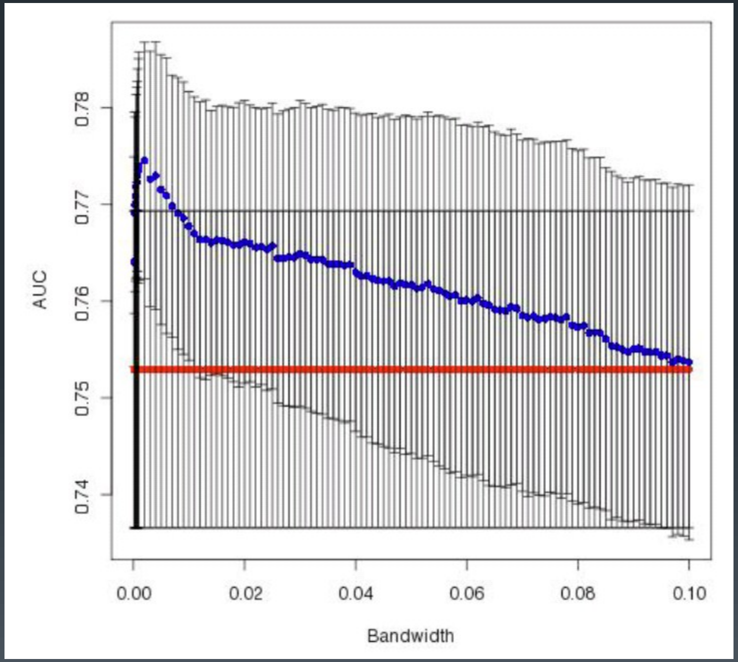
- Results of previous study (Damoulas & Dilkina) that employed kernel density estimates of P_{train}

Computed weights



Red: w(x) > 1

Results

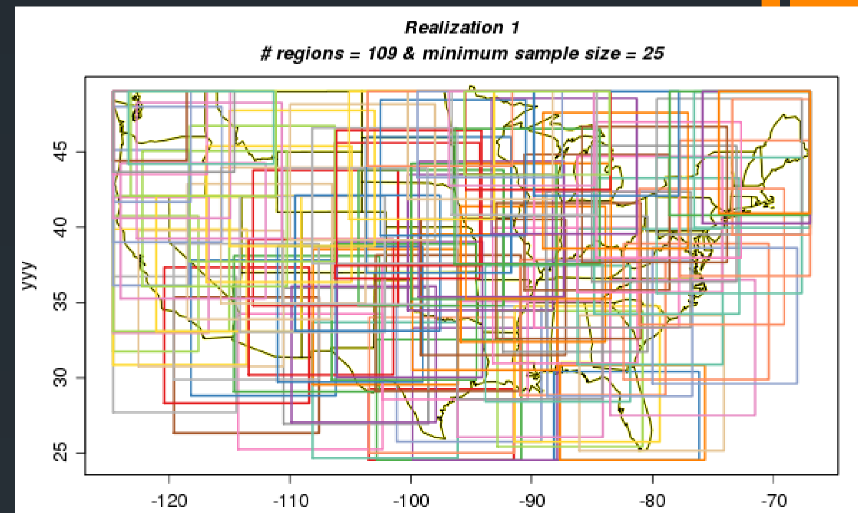


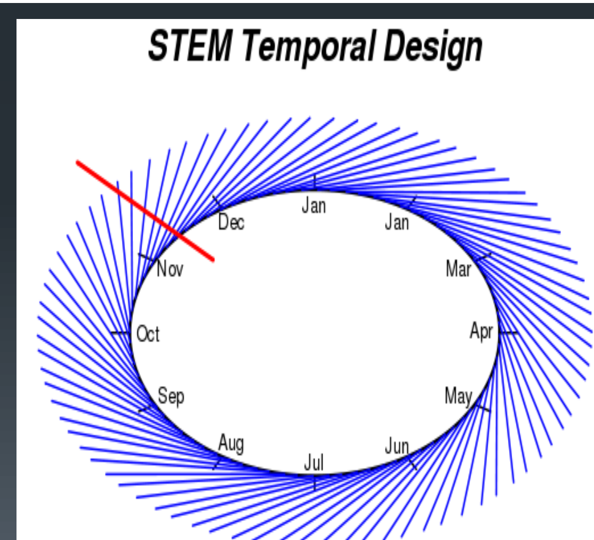
Red: unweighted data
Blue: covariate shift correction

Current State of the Art: STEM
(Fink et al., 2010)

- Idea:
  - Slice space and time into hyperrectangles: lat x long x time
  - Train a decision tree on the data inside each hyperrectangle
  - To predict at a new point x, vote the predictions of all trees whose hyperrectangle contains x

- Hyperrectangles:
  - Space: random rectangles of fixed size
  - Time: 40-day overlapping intervals spaced evenly throughout the year
  - Discard hyperrectangles that contain fewer than 25 training locations





Fink, et al, 2010

Indigo Bunting: Animation from static SDM predictions

Indigo Bunting

slide courtesy of Daniel Fink

Open Problems

4. Population Size Effects
   - Bird population may be too small to occupy all suitable habitat
     - Unoccupied and occupied sites may be identical
5. Spatial Dynamics. Occupied habitat can depend on
   - Discovery – it can be found by existing bird population
   - Accessibility – it can be reached by existing bird population (migration distance)
6. Spatial and Temporal Dynamics of other species
   - Food: insect and plant species
   - Competitors/Predators

Modeling Bird Migration

- Migration most naturally described at level of individual behavior, but we can only observe population-level statistics
  - We need a modeling technique to link the two

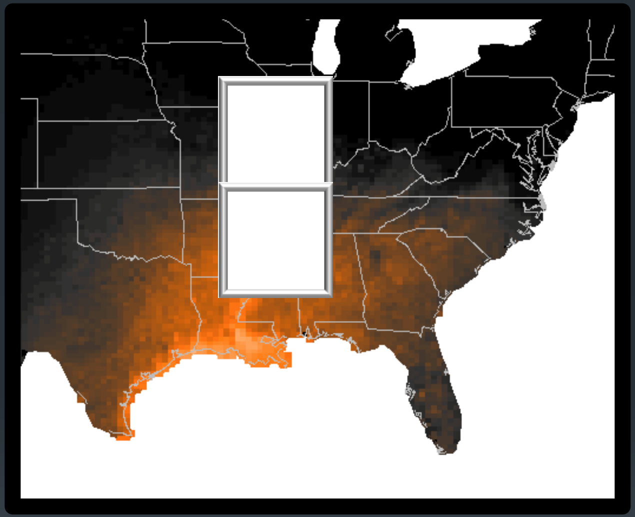
- Our Approach: Collective Graphical Models

Modeling Approach

- Place a grid of cells over North America
- State of a bird at time t = cell it occupies at time t

Time

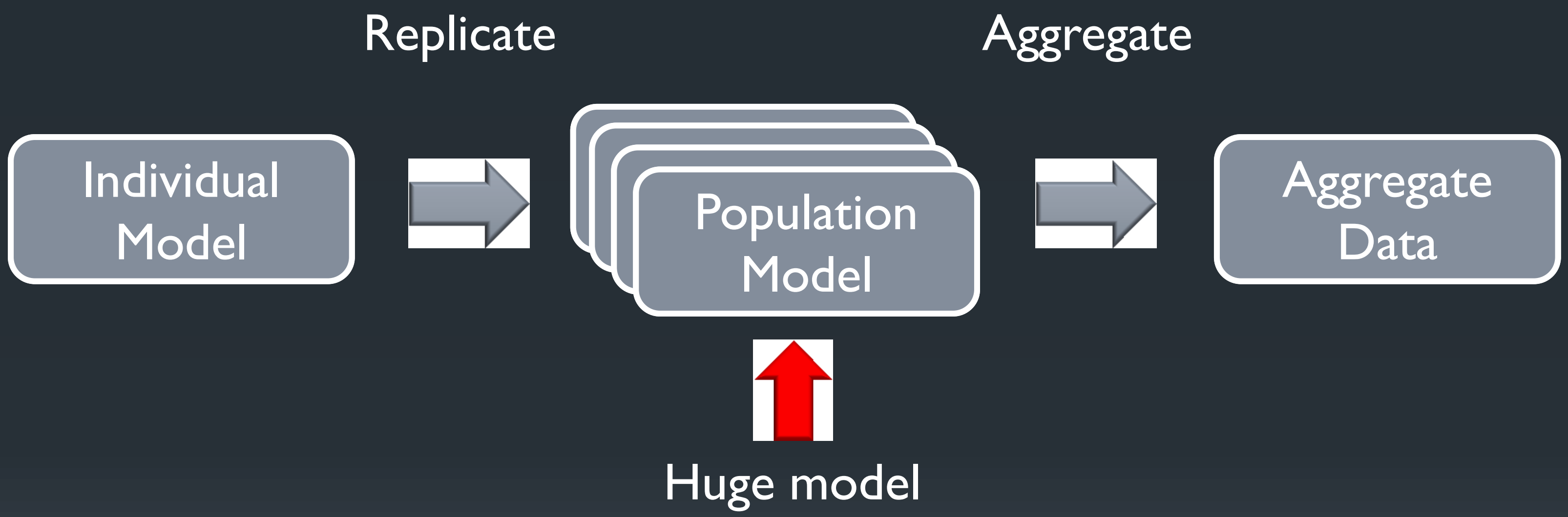
Cell		1	2	3
	A	87	61	22
	B	13	39	78



- Aggregate data: does not track individual birds

Key Modeling Idea

- Build a model for aggregate data starting with a model of individual behavior



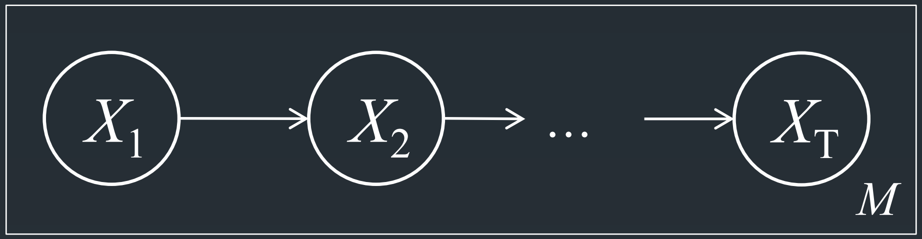
- Goals
  - Infer unobserved quantities about population
  - Learn parameters of individual model

Step 1: Individual Model

$$X_1 \rightarrow X_2 \rightarrow \cdots \rightarrow X_T$$

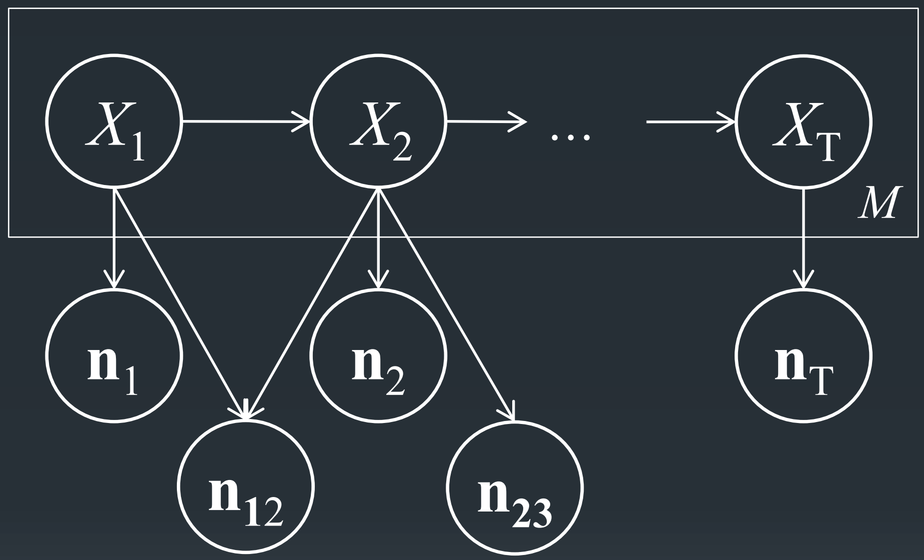
Individual model:
Markov chain

Step 2: iid Population Model



Population model:
iid copies of Markov chain

Step 3: Derive aggregate state variables



Population model:
iid copies of Markov chain

Location counts

Transition counts

Step 4: Marginalize out the Individuals

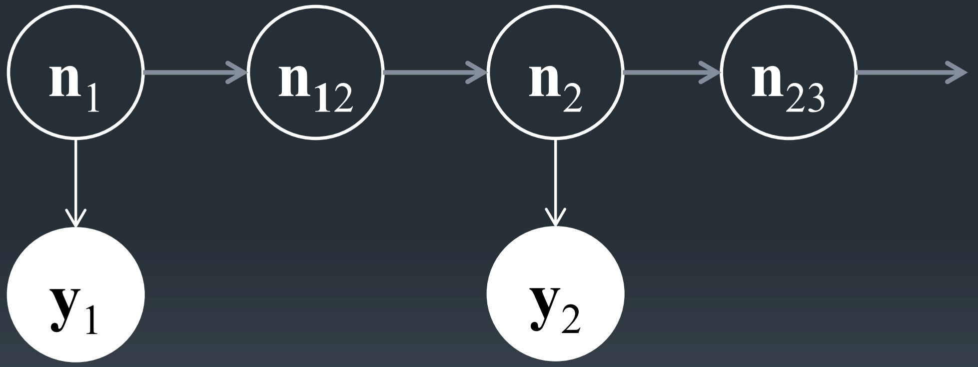
Theorem (Lauritzen, 1996): Count model will have the same dependency structure as the population model

$$\mathbf{n}_1 \rightarrow \mathbf{n}_{12} \rightarrow \mathbf{n}_2 \rightarrow \mathbf{n}_{23} \rightarrow$$

Location counts and transitions

Note that point estimates of these counts give the sufficient statistics for the individual model

Step 5: Attach Observations



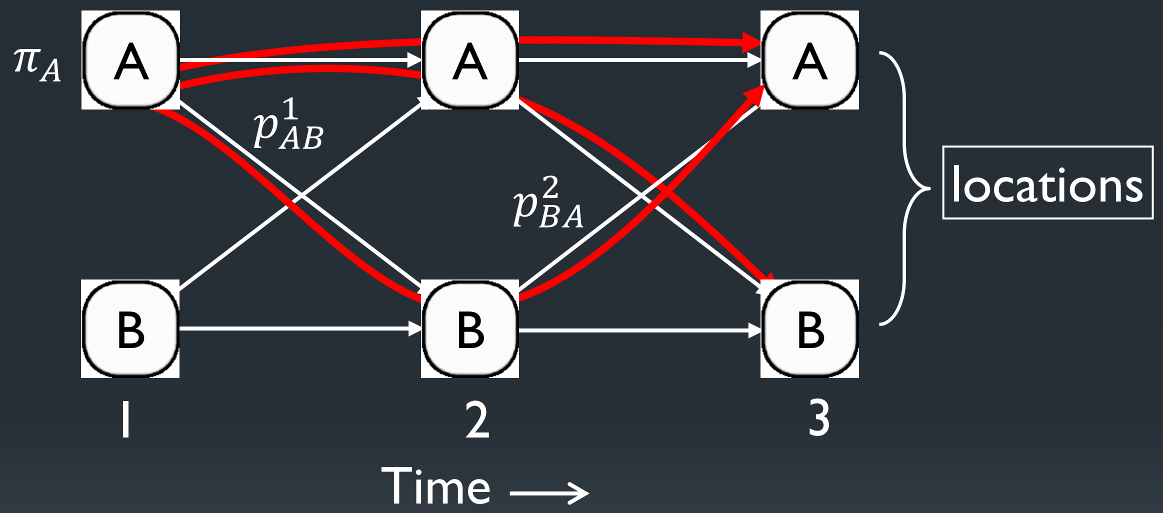
Location counts and transitions

Noisy counts

Posterior inference over n_1, n_{12}, n_2, n_{23}, … gives sufficient statistics for the individual model

Learning in CGMs

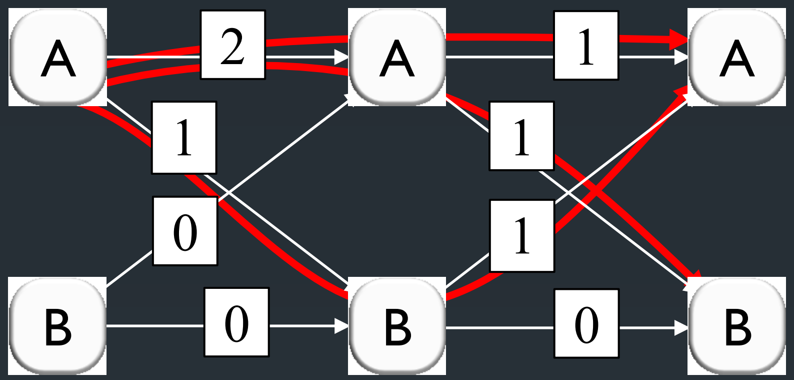
- Migration routes → paths through *trellis graph*



- Parameters: $\theta = \{\pi_i, p_{ij}^{t}\}$
- If we could observe the paths, we could infer θ

Network Flow

- Key observation: collection of M paths → M-unit flow
- To learn θ it is enough to know the flows on each edge

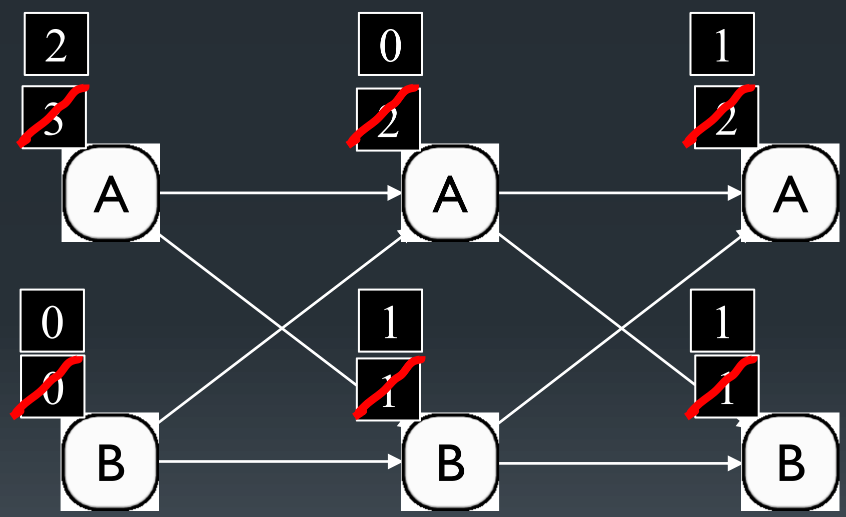


[Sheldon, Kozen, Elmohamed, NIPS 2007]

Learning in CGMs

- Given: Noisy aggregate observations of the # of birds in each cell at each time step
- Find: The parameters θ that maximize $P(observations|\theta)$

$$\theta = \{\pi_i, p_{ij}^t\}$$



Learning the Model is Hard

$$P(\text{observations}|\theta) = \sum_{\text{flows } f} P(f|\theta)P(\text{observations}|f,\theta)$$

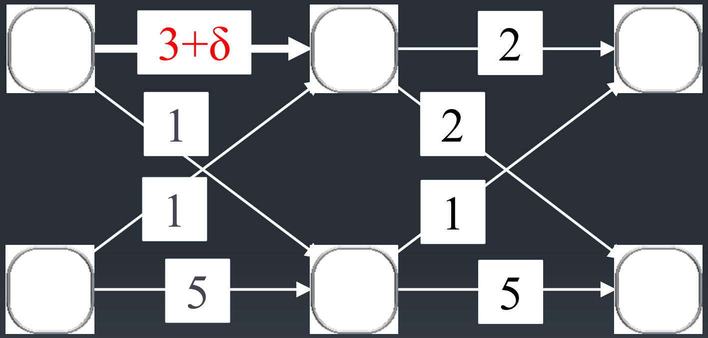
- Solution: Gibbs sampling of the flows

EM/Gibbs

- Expectation Maximization (EM)
  - E-step: Compute $\mathbb{E}[\mathrm{flow}|\mathrm{observations}, \theta]$
  - M-step: Update estimates of the model parameters
- Gibbs sampler for the E-step
  - Sample from $P(\mathrm{flow}|\mathrm{observations}, \theta)$

Gibbs Sampler

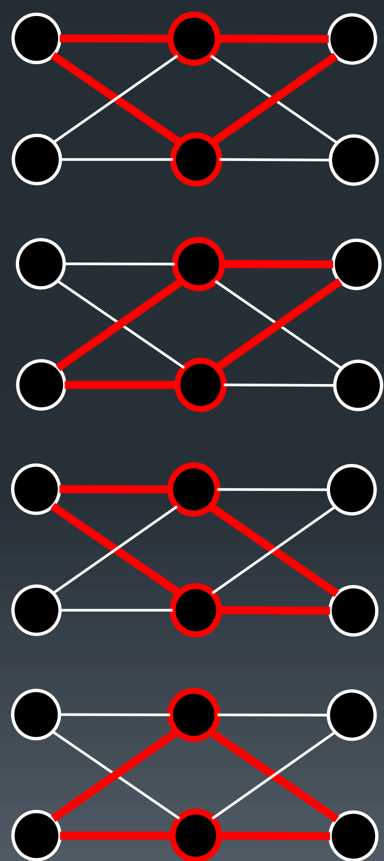
- Initialize flow arbitrarily, then iteratively update by making random “moves”
- Traditionally: update a single variable according to $n^t_{A,A} \sim P\left(n^t_{A,A} \middle| observations, \boldsymbol{n}^t_{-(A,A)}\right)$

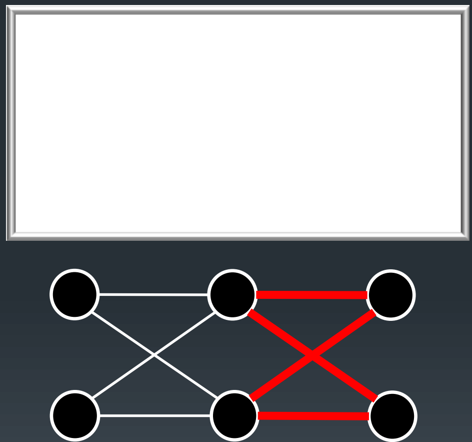


This violates conservation of flow

Make Moves Based on Cycles

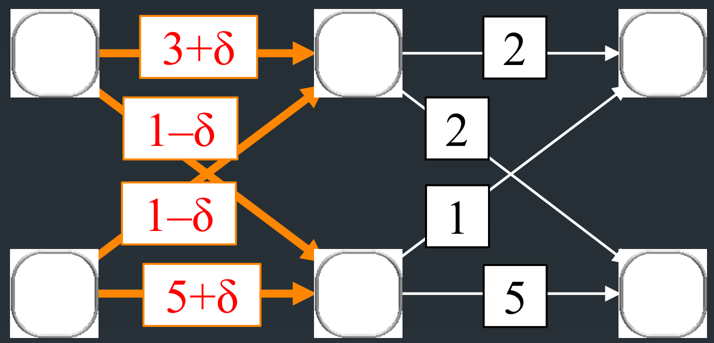
- First, select a 4-cycle in trellis uniformly at random

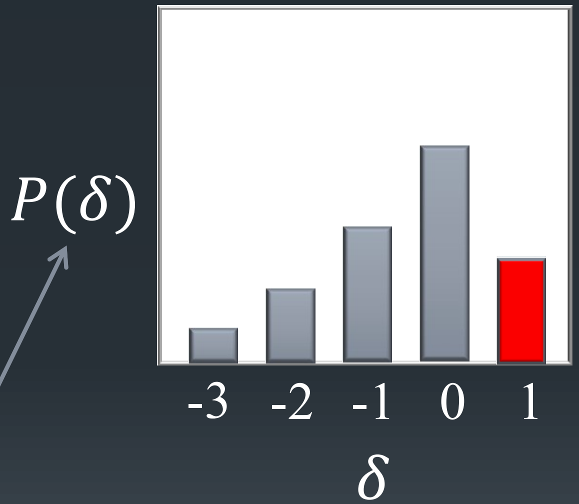




Update

- Send δ units of flow "around the cycle"

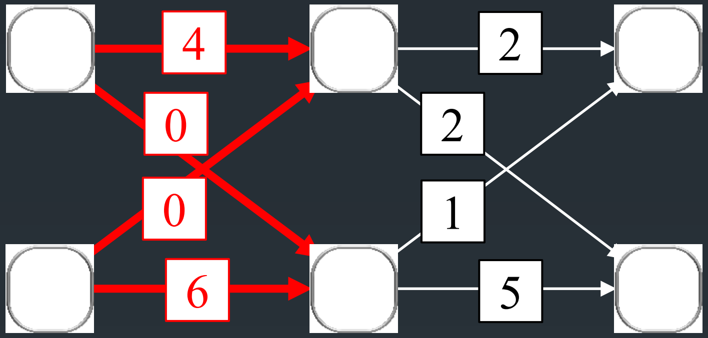




Gibbs update rule: select each value of δ with probability proportional to $P(\text{new flow} \mid \text{observations}, \theta)$

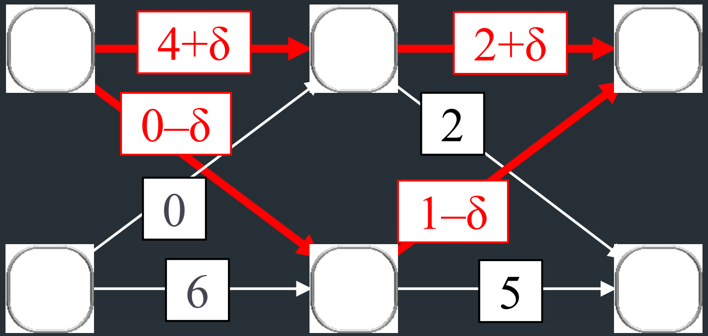
Flow Update Step

- Make the update



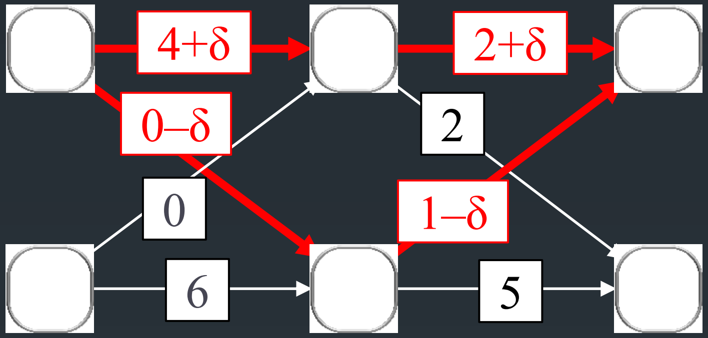
Repeat

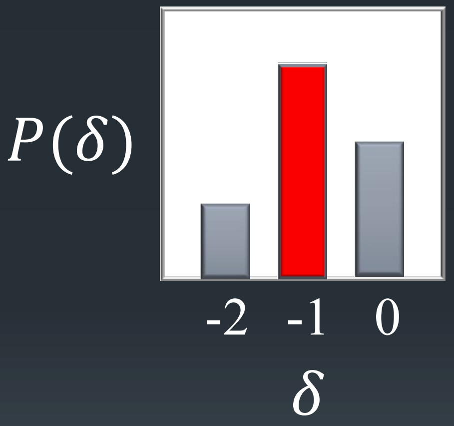
- Select a new random 4-cycle



Repeat

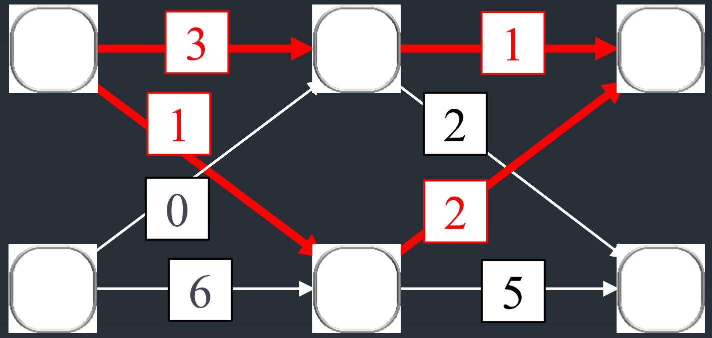
- Choose δ





Repeat

- Make the update



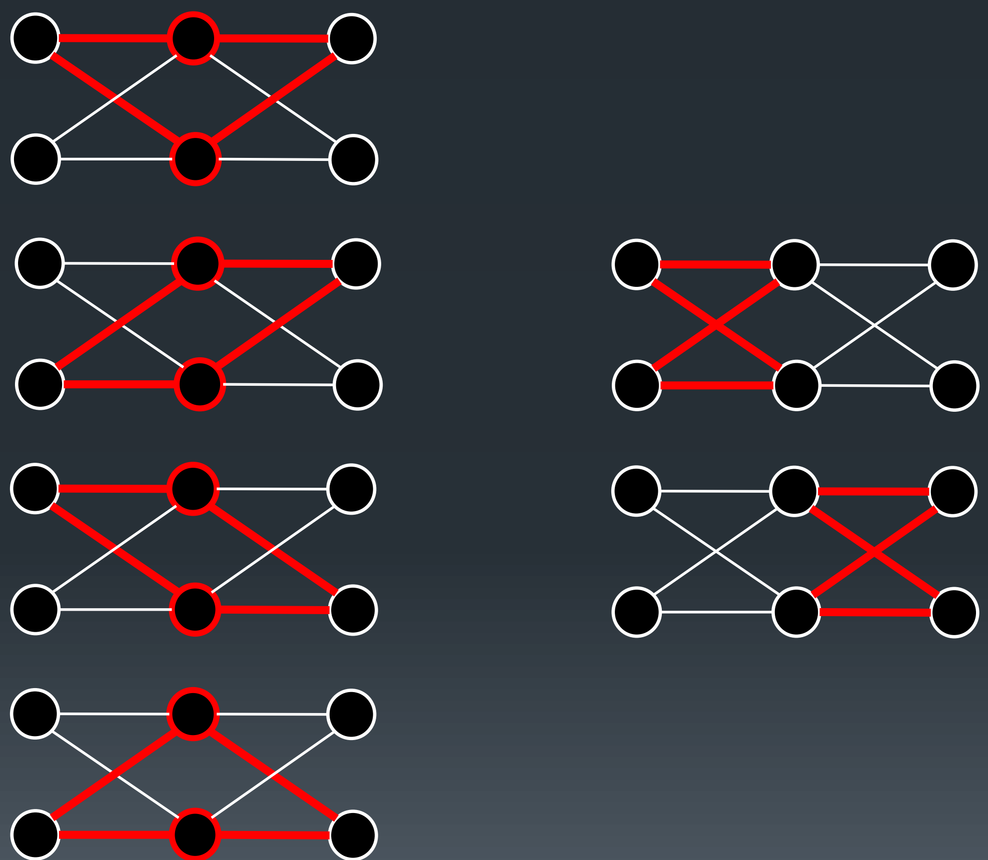
Requirement

- Must be able to move between any two valid flows using this set of moves

… a Markov Basis [Diaconis and Sturmfels, 1998]

Markov Basis

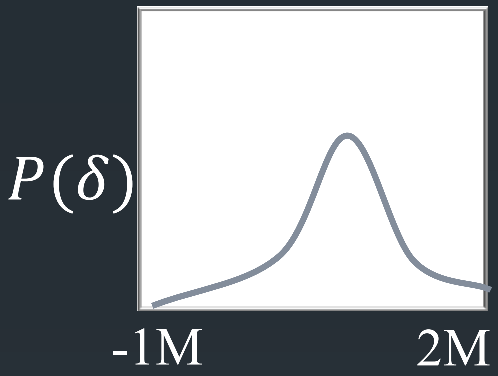
- Theorem: cycles of length four form a Markov basis



Fast Sampling

- How to sample δ quickly when there are many possible values?

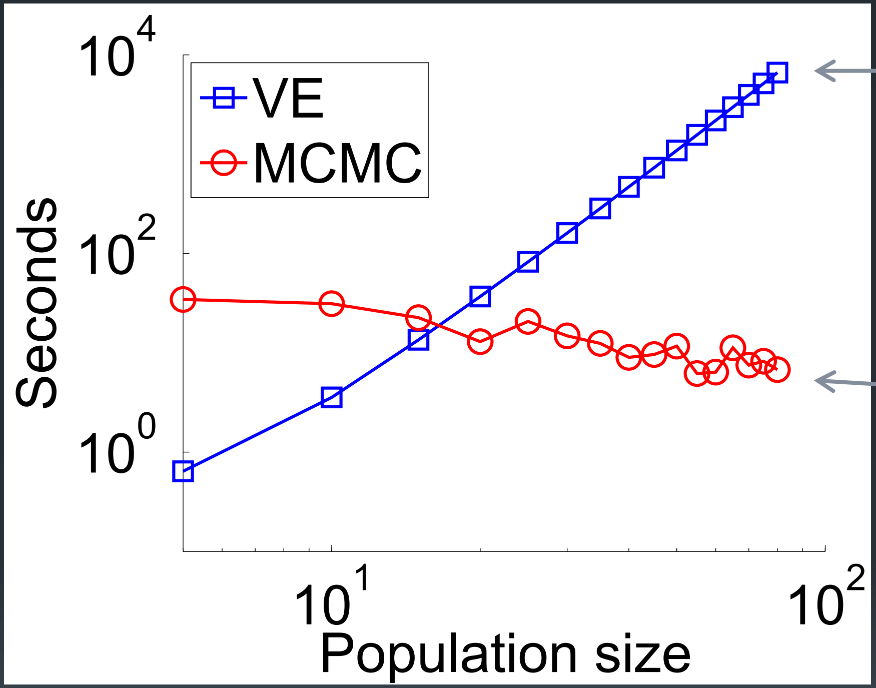
Large Population



- Theorem: $P(\delta)$ is log-concave
  - → Can sample in constant expected running time by rejection sampling [Devroye 1986]
  - → Running time of Gibbs move is *independent of population size*

Result [Sheldon & Dietterich, NIPS 2011]

- Running time on EM task



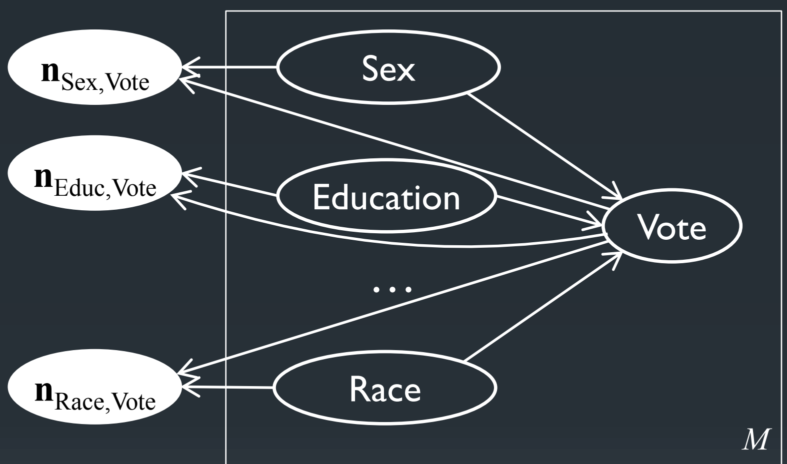
Best exact method
(cubic in M)

Our method
(to 2% relative error)

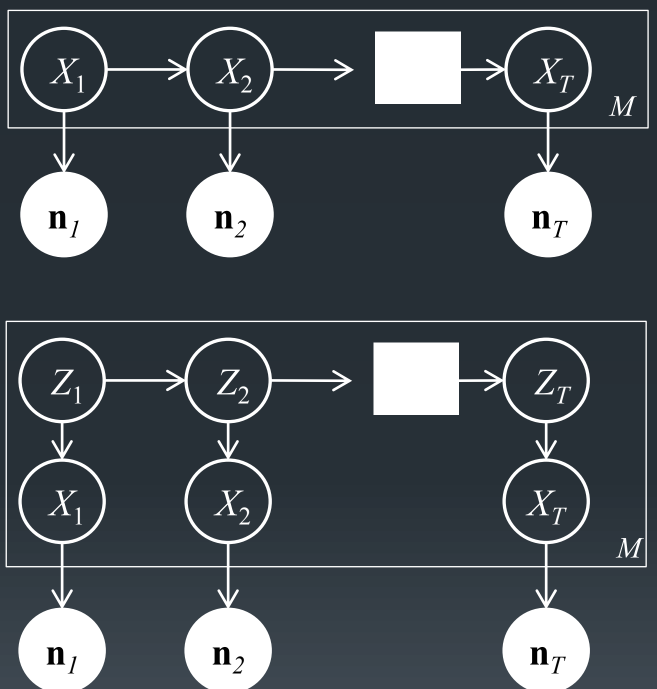
- Running time *independent of population size*
  - Previous best: exponential

Can Generalize to Many Other Settings

- Common situation: only have aggregate data, but want to model individual behavior



US Census
(privacy)

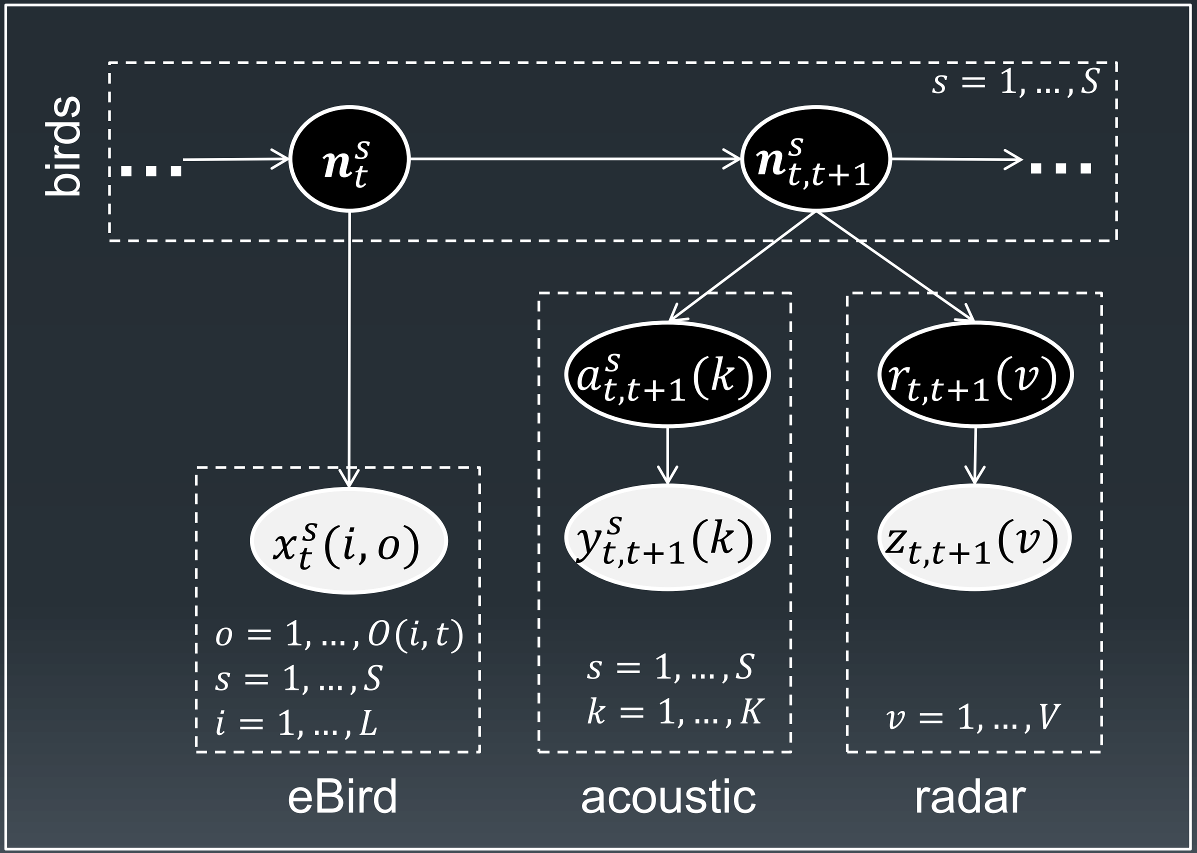


Multiple target tracking
Fish migration

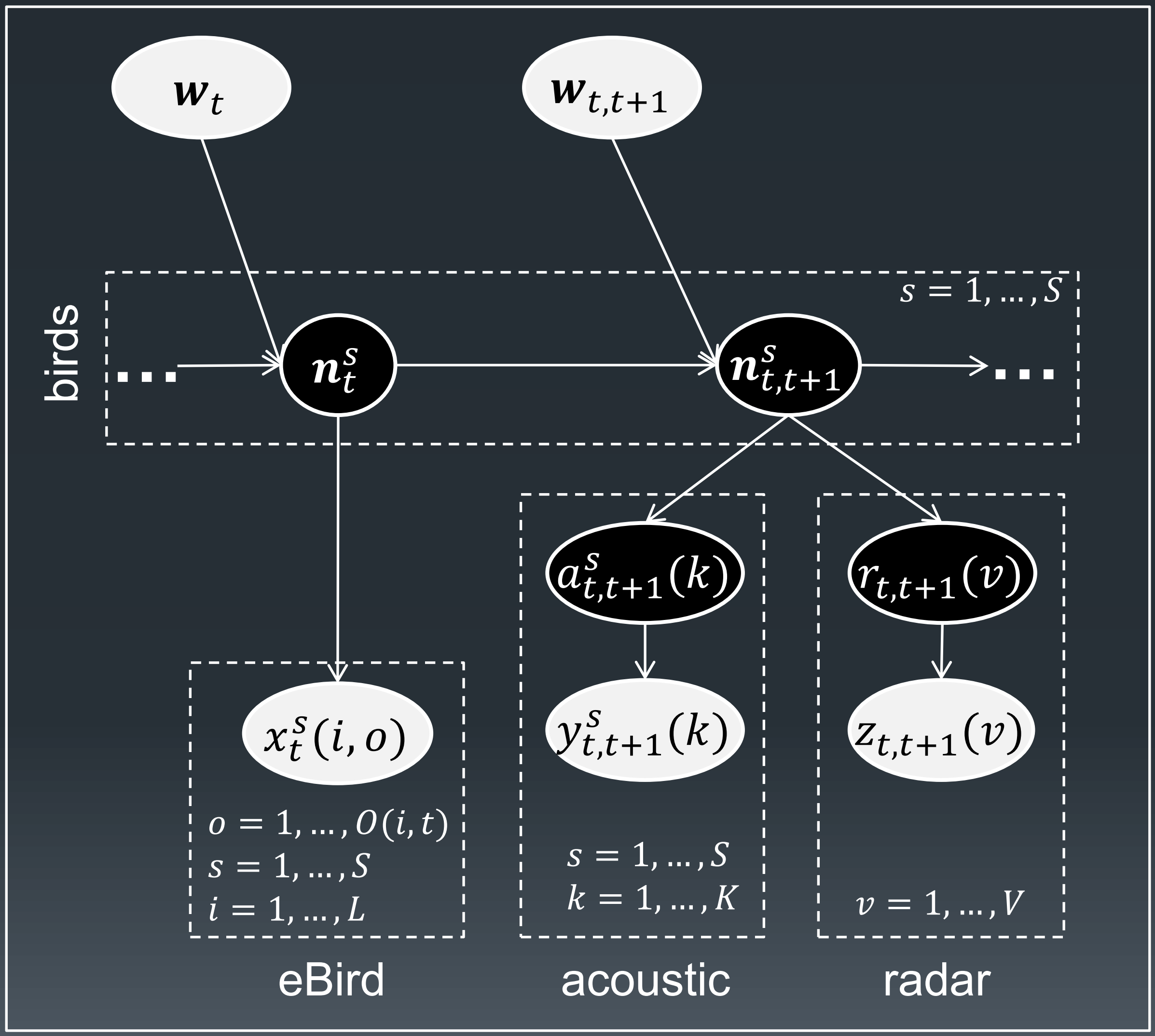
CGM to fuse eBird, radar, and acoustic data

- Species s
- Observers o
- Sites i
- Acoustic stations k
- Radar sites v

- Observation model for eBird (detection, expertise, etc.)
- Observation model for night flight calls (distance to ground, ambient noise)
- Observation model for radar (signal cone, weather, radar "plankton")



Adding Covariates



Summary

- Fitting Species Distribution Models to Citizen Science Data
  - Imperfect Detection
  - Observer Expertise
  - Sampling Bias
- Fitting Dynamical Models to Multiple Data Sources
  - eBird + radar + night flight calls
  - Collective Graphical Models: General Methodology
  - Fast Gibbs sampler for CGMs (independent of population size)

Acknowledgements

- NSF Expeditions in Computing
  - Carla Gomes, PI (Cornell)
- NSF CDI BirdCast grant
  - Steve Kelling, PI (Cornell Lab of Ornithology)
- NSF Bioinformatics Postdoc
  - Dan Sheldon
- NSF/CCC CI Fellows Postdoc
  - Selina Chu
- DARPA Anomaly Detection at Multiple Scales (ADAMS) program
  - Michael Shindler Postdoc
- BRT work: Rebecca Hutchinson (postdoc), Liping Liu (PhD student)
- Density ratio estimation: Selina Chu (postdoc), Michael Shindler (postdoc)
- CGMs: Dan Sheldon (postdoc → UMass Amherst)