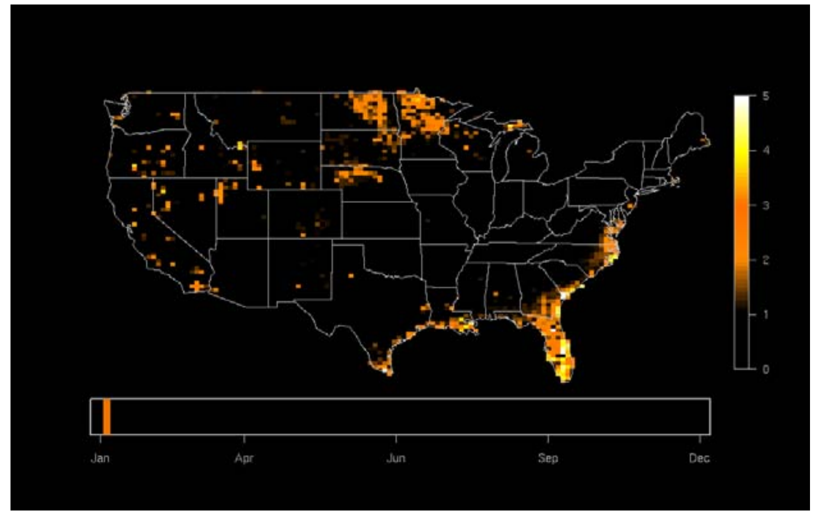
Modeling Experts and Novices in Citizen Science Data

Jun Yu, Weng-Keen Wong, Rebecca Hutchinson
{yuju,wong,rah}@eecs.oregonstate.edu

Introduction

Species Distribution Modeling important for:

- Understanding species-habitat relationships
- Conservation and reserve design
- Predicting effects of climate / land use change



Predicted distribution of tree swallows across North America (from D. Fink)

Many research questions require data to be collected at broad spatial and temporal scales

Introduction

Citizen science: scientific research in which volunteers from the community participate as field assistants [Cohn 2008]

Pros:	Cons
• Inexpensive • Can collect data over large spatial areas and long time periods	• Reliability of data

Introduction

- One of the largest citizen science programs
- Online checklist database developed by Cornell Lab of Ornithology and National Audubon Society
- Birders submit checklists of birds observed (> 1.5 million checklists in Jan 2010)

Introduction

Can we use eBird data for accurate SDM?

- Main issue: birders have different levels of expertise

Novice Expert

- How reliable is the data?
  - Data reviewed through a verification process
  - But biases still exist

Methodology

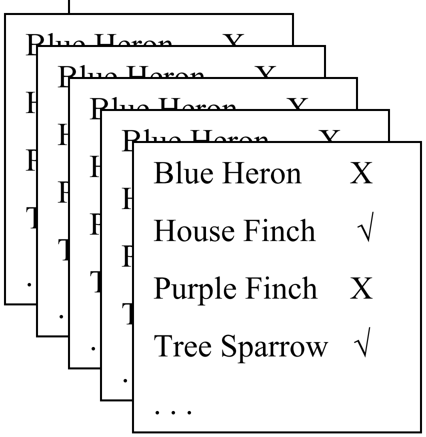
Labeled Training Set

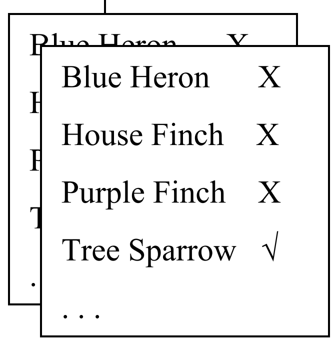
Birder ID: 42

Expertise: Expert

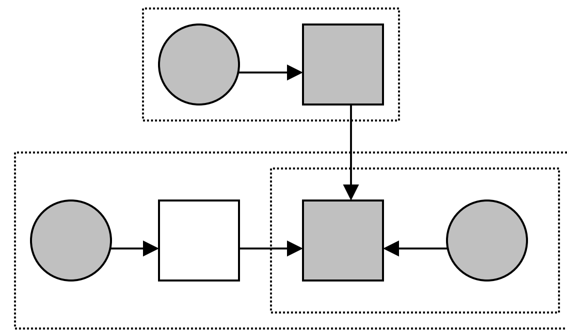
Birder ID: 56

Expertise: Novice

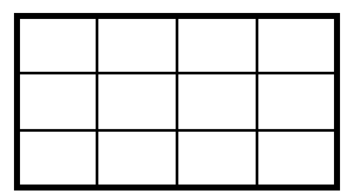




Train model



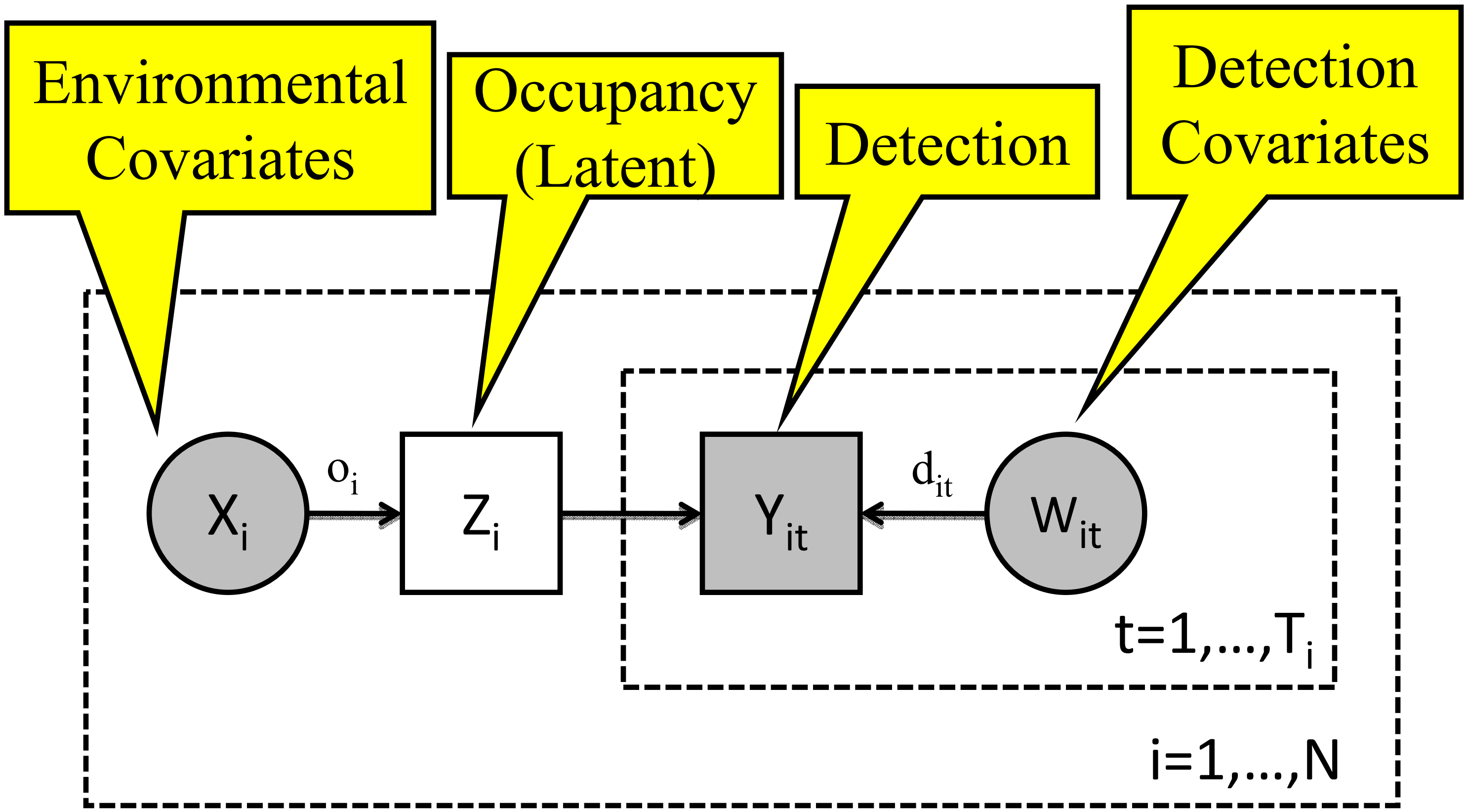
Use model



32 experts (2532 checklists)

88 novices (2107 checklists)

Methodology



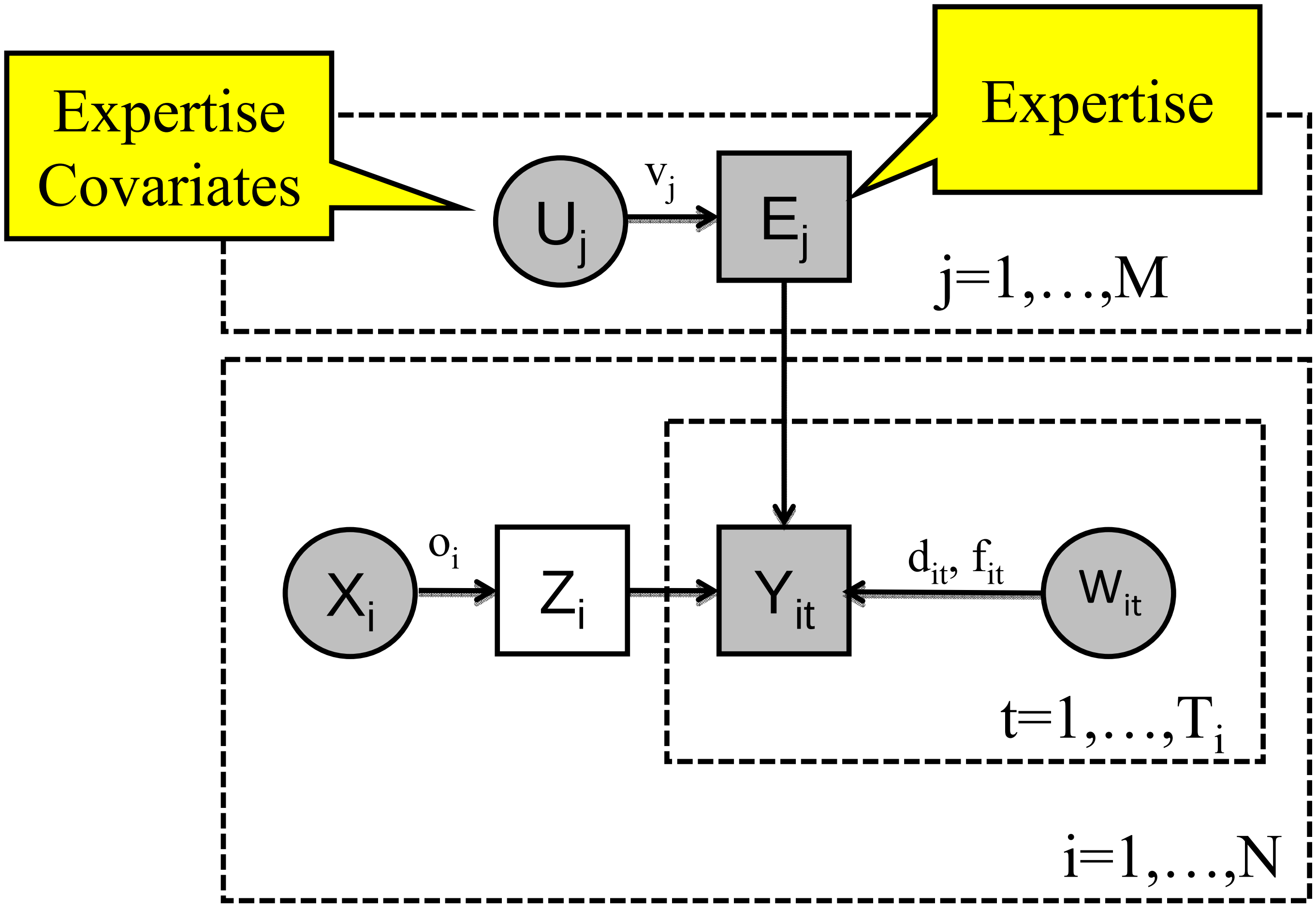
Start with Occupancy-Detection (OD) model [Mackenzie et al. 2006]

Methodology

Assumptions on OD model

- Site closure assumption: species occupancy status stays the same over the site visits
- No false detections: can't detect a bird if it doesn't occupy the site

Methodology



Occupancy-Detection-Expertise (ODE) model

Methodology

ODE model details

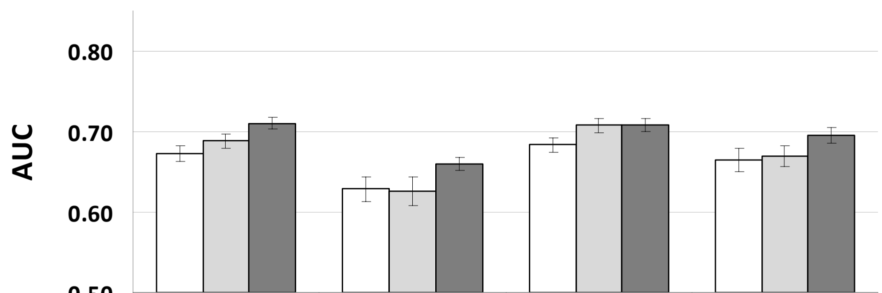
- Allow for false detections. Results in four sets of parameters:
  - True detection and false detection parameters for experts
  - True detection and false detection parameters for novices
- Introduces an identifiability problem
  - Add constraint during training
- Train using Expectation-Maximization

Results

1. **Want to predict occupancy (Z_i) but ground truth not available. Instead, predicting observation (Y_{it})**
   – eBird data from NY, breeding season (2006-2008)
   – Expertise nodes observed in training data, unobserved in test data
   – Evaluating spatial data is challenging: use checkerboarding
   – Compare with Logistic Regression and OD model

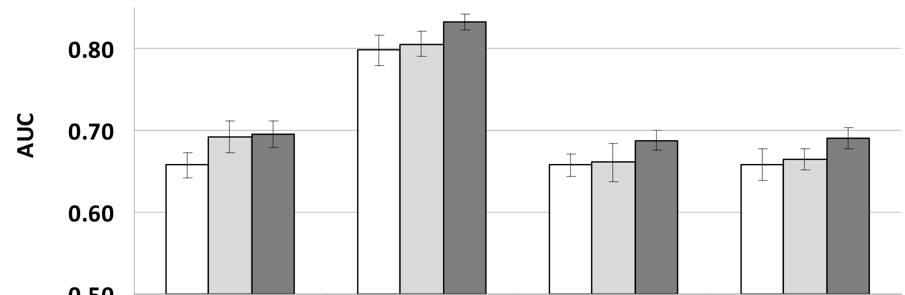
Results

Average AUC on four common bird species



	Blue Jay	White-breasted Nuthatch	Northern Cardinal	Great Blue Heron
LR	0.6726	0.6283	0.6831	0.6641
OD	0.6881	0.6262	0.7073	0.6691
ODE	0.7104	0.6600	0.7085	0.6959

Average AUC on four hard-to-detect bird species



	Brown Thrasher	Blue-headed Vireo	Northern Rough-winged Swallow	Wood Thrush
LR	0.6576	0.7976	0.6575	0.6579
OD	0.6920	0.8055	0.6609	0.6643
ODE	0.6954	0.8325	0.6872	0.6903

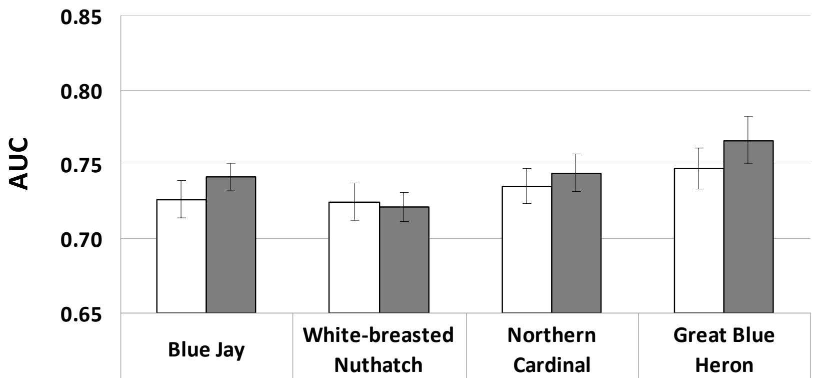
Results

2. Predict Expertise (E_j) of birder given checklist history

- Site occupancy (Z_i) is unobserved in both training and testing
- Two-fold cross-validation on birders
- Repeat 20 times and report average AUC
- Compare against Logistic Regression

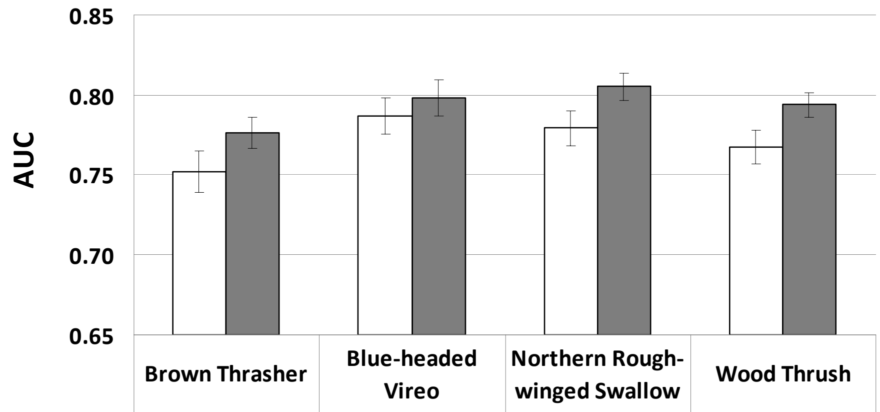
Results

Average AUC on four common bird species



	Blue Jay	White-breasted Nuthatch	Northern Cardinal	Great Blue Heron
☐ LR	0.7265	0.7249	0.7352	0.7472
■ ODE	0.7417	0.7212	0.7442	0.7661

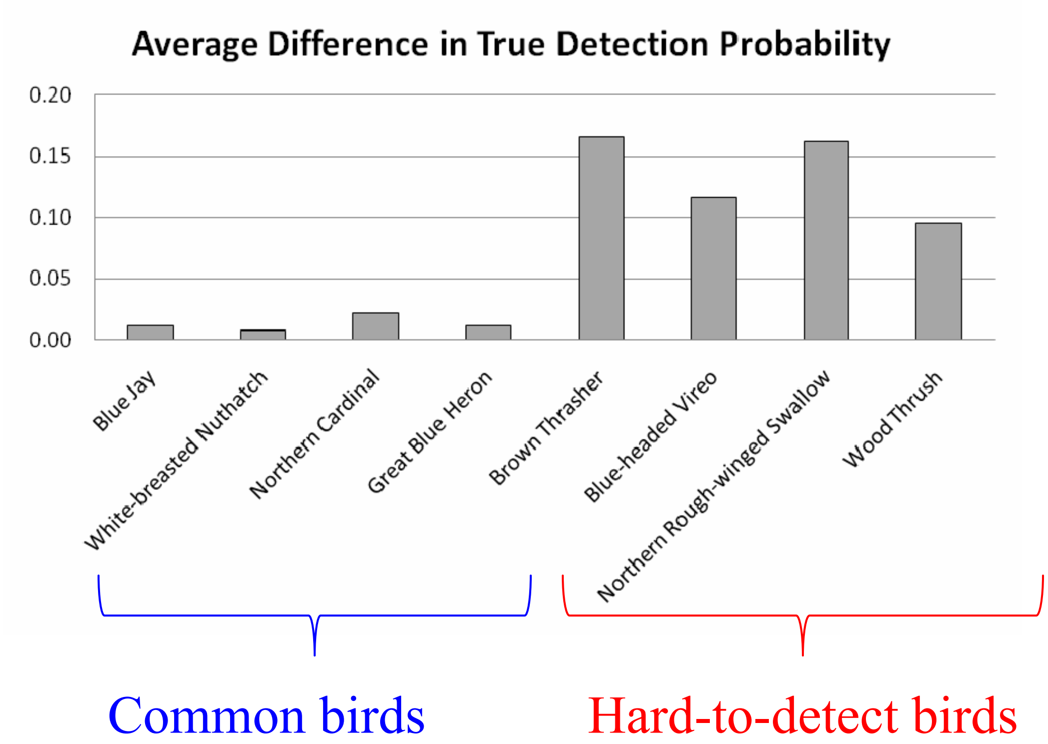
Average AUC on four hard-to-detect bird species



	Brown Thrasher	Blue-headed Vireo	Northern Rough-winged Swallow	Wood Thrush
☐ LR	0.7523	0.7869	0.7792	0.7675
■ ODE	0.7761	0.7981	0.8052	0.7937

Results

3. Discovering differences between experts and novices



Future work

- Discover sources of novice bias
- Improve accuracy of species distribution models by adjusting for this novice bias
- Incorporate tree-models in occupancy and detection components
- Semi-supervised version of ODE model

Acknowledgements

- Cornell Lab of Ornithology:
  - Marshall Iliff
  - Brian Sullivan
  - Chris Wood
  - Steve Kelling
- This project supported by NSF grant CCF 0832804