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

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

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

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

<table>
<thead>
<tr>
<th>Pros</th>
<th>Cons</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Inexpensive</td>
<td>• Reliability of data</td>
</tr>
<tr>
<td>• Can collect data over large spatial areas and long time periods</td>
<td></td>
</tr>
</tbody>
</table>
Introduction

**eBird**

- 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
  - Blue Heron: X
  - House Finch: √
  - Purple Finch: X
  - Tree Sparrow: √

- Birder ID: 56
  - Expertise: Novice
  - Blue Heron: X
  - House Finch: X
  - Purple Finch: X
  - Tree Sparrow: √

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

Expertise Covariates

\[ U_j \xrightarrow{v_j} E_j \]

\[ j = 1, \ldots, M \]

\[ X_i \xrightarrow{o_i} Z_i \xrightarrow{d_{it}, f_{it}} Y_{it} \xrightarrow{w_{it}} \]

\[ t = 1, \ldots, T_i \]

\[ i = 1, \ldots, N \]

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

<table>
<thead>
<tr>
<th></th>
<th>Blue Jay</th>
<th>White-breasted Nuthatch</th>
<th>Northern Cardinal</th>
<th>Great Blue Heron</th>
</tr>
</thead>
<tbody>
<tr>
<td>LR</td>
<td>0.6726</td>
<td>0.6283</td>
<td>0.6831</td>
<td>0.6641</td>
</tr>
<tr>
<td>OD</td>
<td>0.6881</td>
<td>0.6262</td>
<td>0.7073</td>
<td>0.6691</td>
</tr>
<tr>
<td>ODE</td>
<td>0.7104</td>
<td>0.6600</td>
<td>0.7085</td>
<td>0.6959</td>
</tr>
</tbody>
</table>

Average AUC on four hard-to-detect bird species

<table>
<thead>
<tr>
<th></th>
<th>Brown Thrasher</th>
<th>Blue-headed Vireo</th>
<th>Northern Rough-winged Swallow</th>
<th>Wood Thrush</th>
</tr>
</thead>
<tbody>
<tr>
<td>LR</td>
<td>0.6576</td>
<td>0.7976</td>
<td>0.6575</td>
<td>0.6579</td>
</tr>
<tr>
<td>OD</td>
<td>0.6920</td>
<td>0.8055</td>
<td>0.6609</td>
<td>0.6643</td>
</tr>
<tr>
<td>ODE</td>
<td>0.6954</td>
<td>0.8325</td>
<td>0.6872</td>
<td>0.6903</td>
</tr>
</tbody>
</table>
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: 0.7265
- White-breasted Nuthatch: 0.7249
- Northern Cardinal: 0.7352
- Great Blue Heron: 0.7472

Average AUC on four hard-to-detect bird species

- Brown Thrasher: 0.7523
- Blue-headed Vireo: 0.7869
- Northern Rough-winged Swallow: 0.7792
- Wood Thrush: 0.7675

LR 0.7265 0.7249 0.7352 0.7472
ODE 0.7417 0.7212 0.7442 0.7661
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

![Average Difference in True Detection Probability Graph]

- **Common birds**: Blue Jay, White-breasted Nuthatch, Northern Cardinal, Great Blue Heron, Brown Thrasher
- **Hard-to-detect birds**: Blue-headed Vireo, Northern Rough-winged Swallow, Wood Thrush
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