Oregon State University

COLLEGE OF ENGINEERING

On the Role of Spatial Clustering Algorithms in Building Species Distribution Models from Community Science Data

Mark Roth[1], Dr. Tyler Hallman[2],

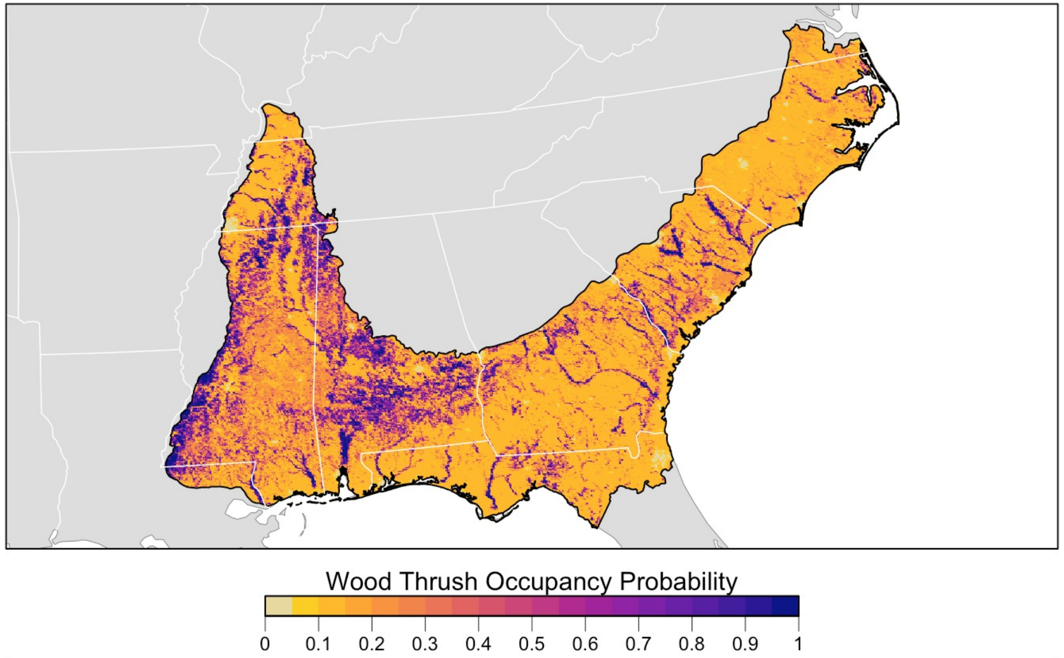
Dr. W. Douglas Robinson[3], Dr. Rebecca Hutchinson[1,3]




1: Department of Electrical Engineering & Computer Science, Oregon State University
2: Swiss Ornithological Institute, Sempach, Switzerland
3: Department of Fisheries, Wildlife, & Conservation Sciences, Oregon State University

Species Distribution Models (SDMs)

- Tools that predict the pattern of species activity
  - Integral in designing solutions to support threatened species



Data for SDMs

- Extent and accuracy of SDMs depend on the range and quality of the biodiversity dataset

- Community Science provides the data necessary to construct accurate, comprehensive SDMs !

Community Science (also known as citizen science)

- Voluntary crowdsourced data collection
- Low barriers to contribute
- Growing in size, quality, and importance
- New & existing challenges
  - Imperfect detection

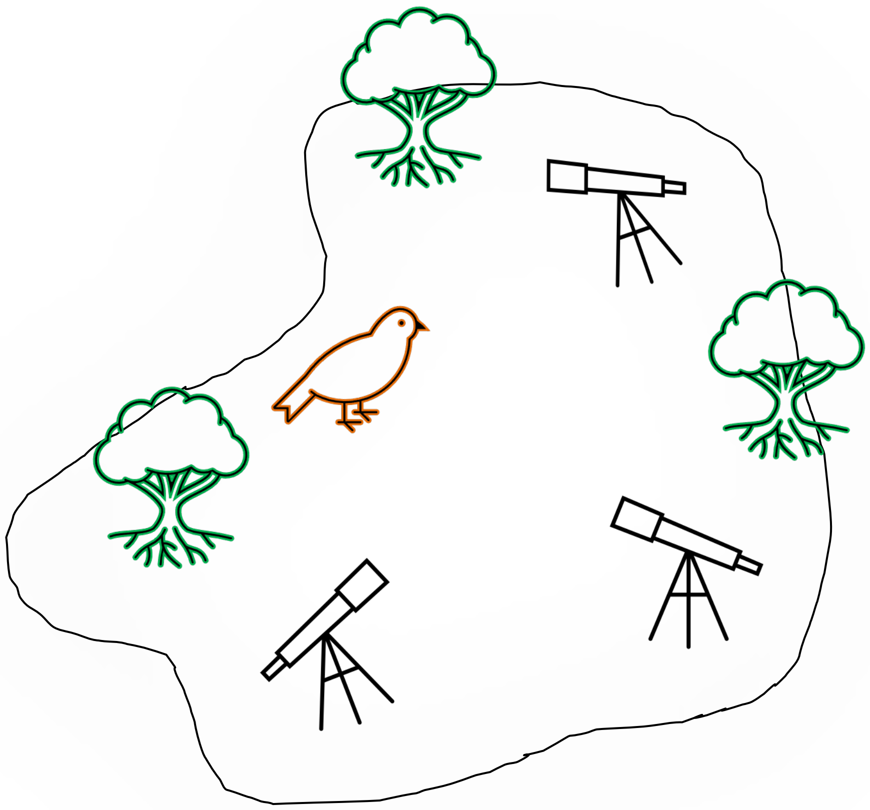
Imperfect Detection

- Probability of detecting a species given that it is present is less than 1
- Ignoring imperfect detection can lead to biased estimates of occupancy (Guillera-Arroita et al., 2014)
- Occupancy Models!

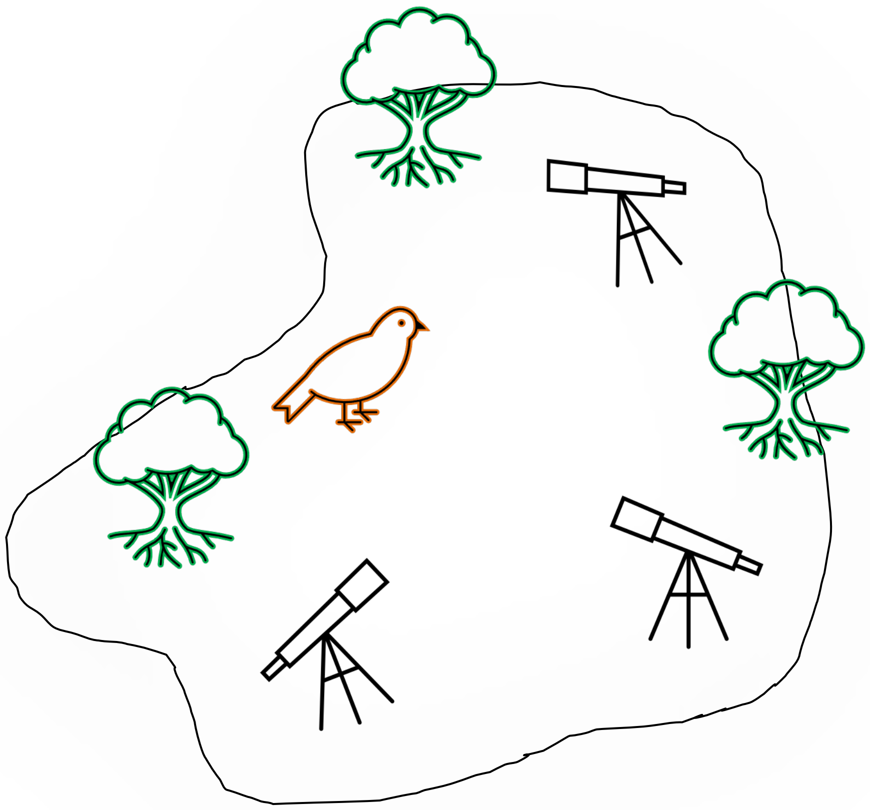
Occupancy Models

- Rely on a few key assumptions to account for imperfect detection:
  1. No false positives
  2. N observations are organized into a set of $\leq$N *sites*
  3. At each site, we assume *closure*: the occupancy status remains unchanging across all observations

Occupancy Model – Intuition

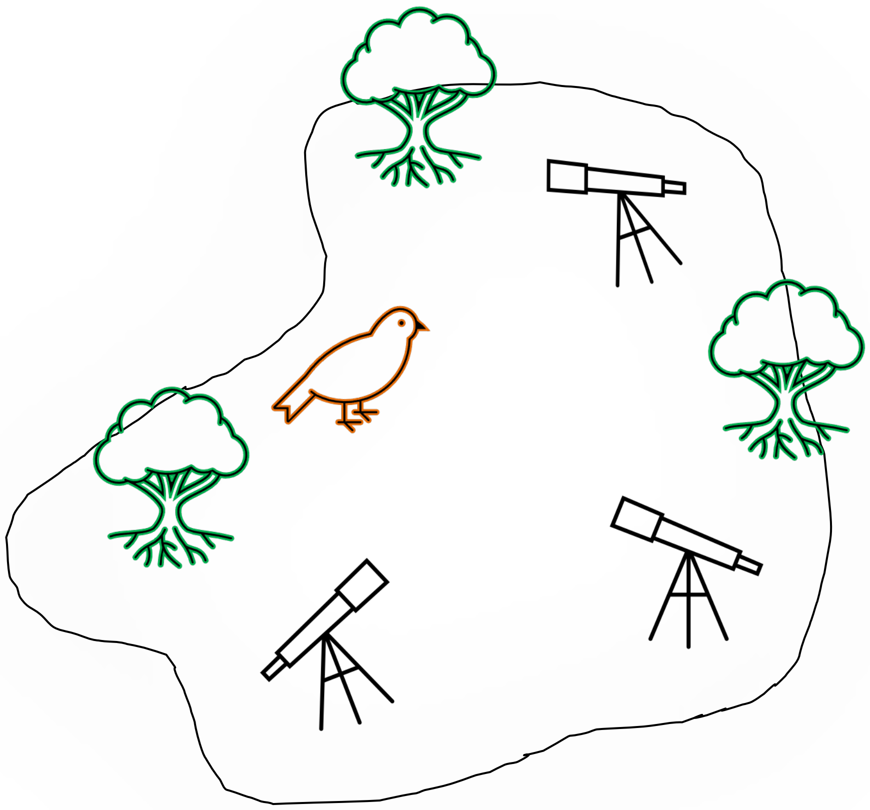


Occupancy Model – Intuition



Observations: [0, 0, 1]

Occupancy Model – Intuition



Observations: [0, 0, 1]

Detection probability = 1/3

Oregon State University
College of Engineering

Occupancy Model MLE for a single site

$$L(\psi, \mathbf{p}) = \left[\psi^{n.} \prod_{t=1}^{T} p_t^{n_t} (1-p_t)^{n. - n_t}\right] \times \left[\psi \prod_{t=1}^{T} (1-p_t) + (1-\psi)\right]^{N-n}$$

ψ: *occupancy probability*
p_t: *detection probability at time t*
N: *total number of sites*
T: *number of distinct sampling occasions*
n_t: *number of sites where the species was detected at time t*
$n.$: *number of sites at which a species was detected*

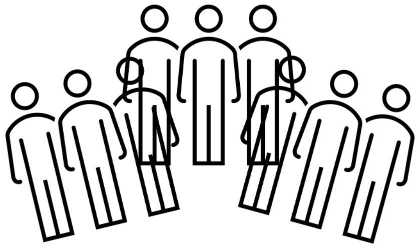
MacKenzie et al., 2002

Occupancy Models

- Rely on a few key assumptions to account for imperfect detection:
  1. No false positives
  2. N observations are organized into a set of $\leq$N *sites*
  3. At each site, we assume *closure*: the occupancy status remains unchanging across all observations

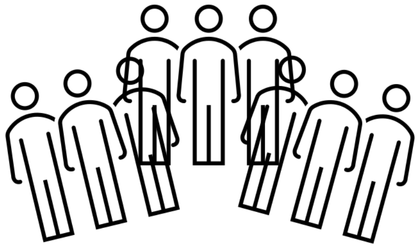
Scientists design sites prior to sampling to ensure closure, but this is not the case with community science!

Pathway to climate change mitigation



Unstructured, crowdsourced biodiversity datasets

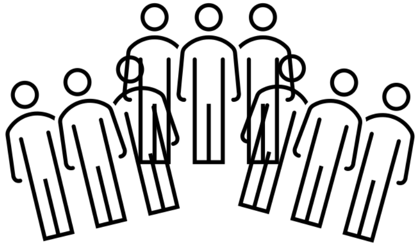
Pathway to climate change mitigation



Unstructured, crowdsourced biodiversity datasets

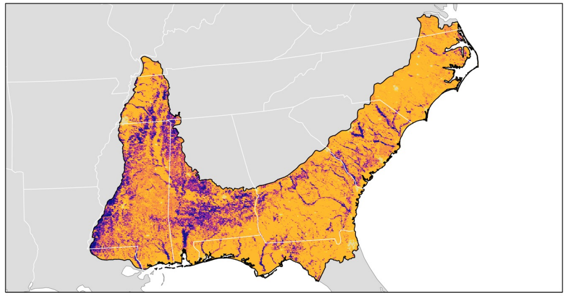
→ OMs

Pathway to climate change mitigation



Unstructured, crowdsourced biodiversity datasets

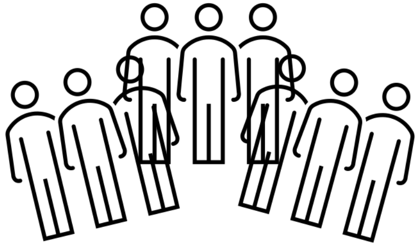
→ OMs →



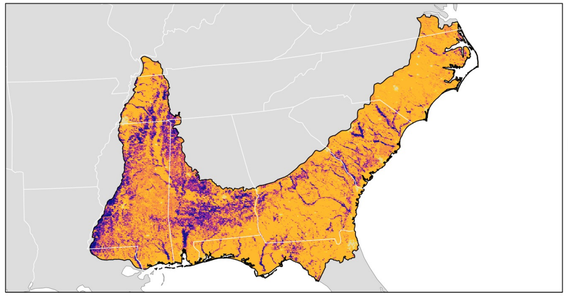
SDMs

Pathway to climate change mitigation

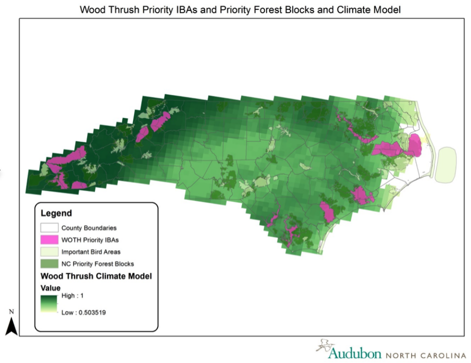
Oregon State University
College of Engineering



Unstructured, crowdsourced biodiversity datasets → OMs →

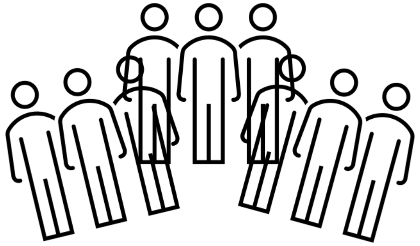


SDMs →



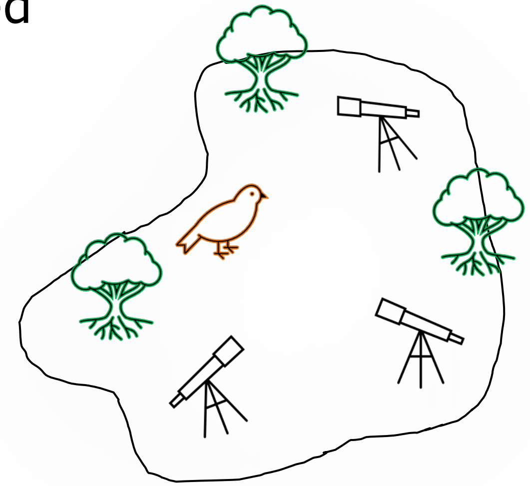
Natural Resource Management

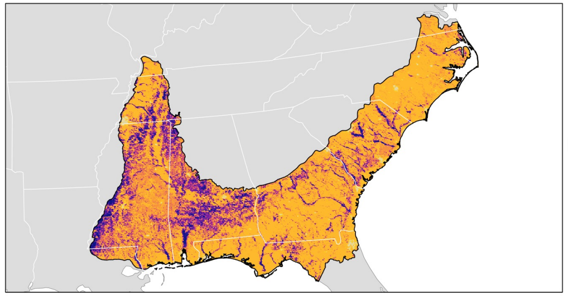
Pathway to climate change mitigation



Unstructured, crowdsourced biodiversity datasets

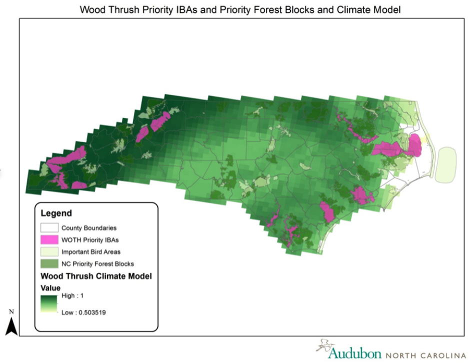
→ OMs →





SDMs

→

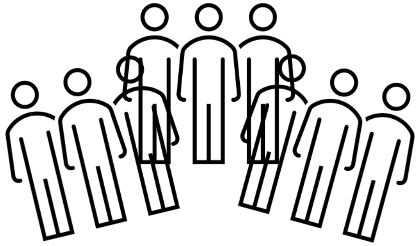


Natural Resource Management

Pathway to climate change mitigation

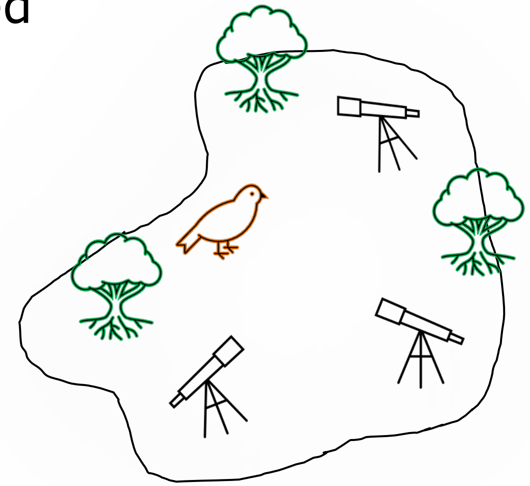
Oregon State University
College of Engineering

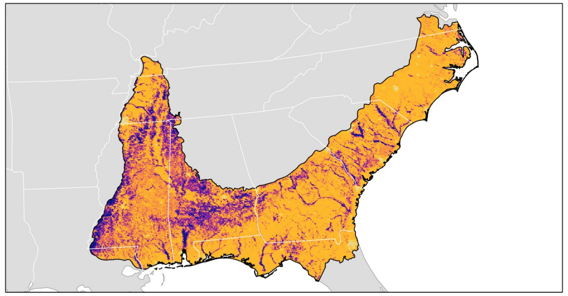
Group independent observations into sites while maintaining closure



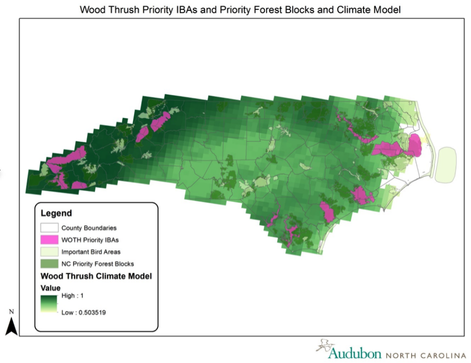
Unstructured, crowdsourced biodiversity datasets

OMs





SDMs



Natural Resource Management

Oregon State University
College of Engineering

Group independent observations into sites while maintaining closure

Site Clustering Problem

Group independent observations into sites while maintaining closure

Site Clustering Problem

Group independent observations into sites while maintaining closure

1. Discover the optimal number of sites automatically

Site Clustering Problem

Oregon State University
College of Engineering

Group independent observations into sites while maintaining closure

1. Discover the optimal number of sites automatically
2. Respect geospatial & temporal constraints imposed by species behavior

Site Clustering Problem

Oregon State University
College of Engineering

Group independent observations into sites while maintaining closure

1. Discover the optimal number of sites automatically
2. Respect geospatial & temporal constraints imposed by species behavior
3. Consider similarity in geospatial & feature space

Site Clustering Problem

Oregon State University
College of Engineering

Group independent observations into sites while maintaining closure

1. Discover the optimal number of sites automatically
2. Respect geospatial & temporal constraints imposed by species behavior
3. Consider similarity in geospatial & feature space
4. Run efficiently on large datasets

Site Clustering Problem

Oregon State University
College of Engineering

Group independent observations into sites while maintaining closure

Site Clustering Problem

Group independent observations into sites while maintaining closure

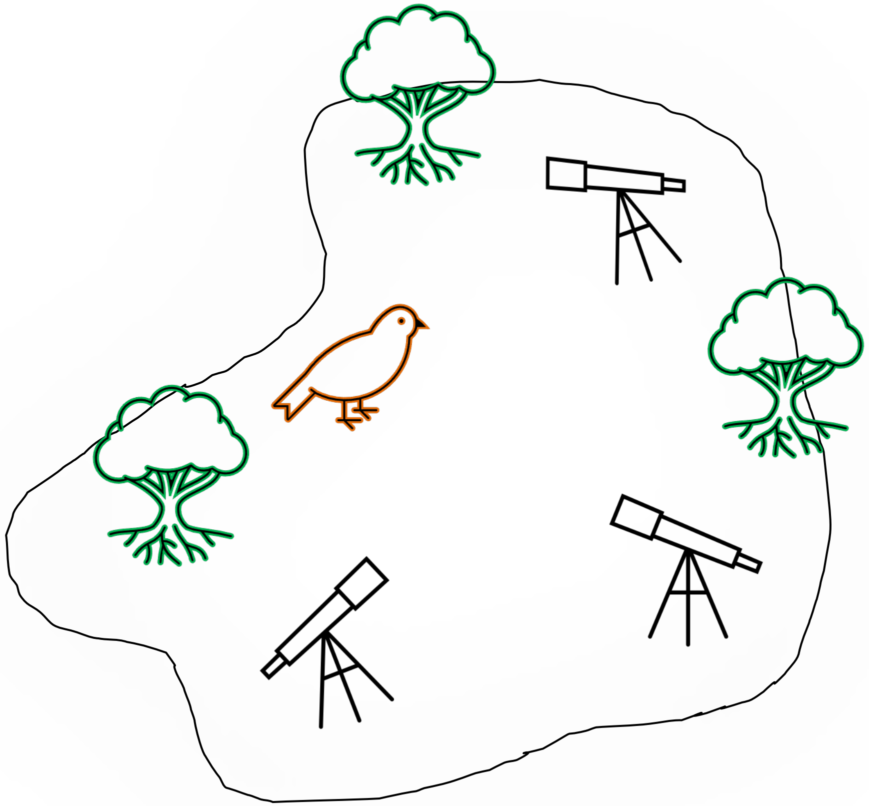
- Our proposal focuses on the eBird dataset

Site Clustering Problem

Group independent observations into sites while maintaining closure

- Our proposal focuses on the eBird dataset
- Observers submit *checklists* that list the birds they saw and the time and location of observation

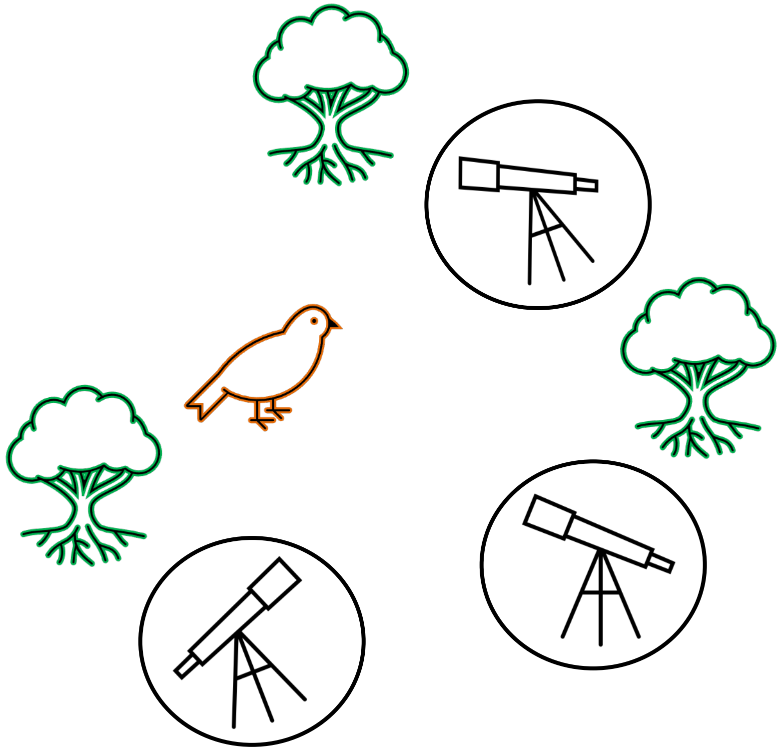
Existing Methods



Oregon State University College of Engineering

Existing Methods

1. eBird Best Practices
   - Same observer, same latitude-longitude coordinate, > 1 visit and at most 10 visits



Existing Methods

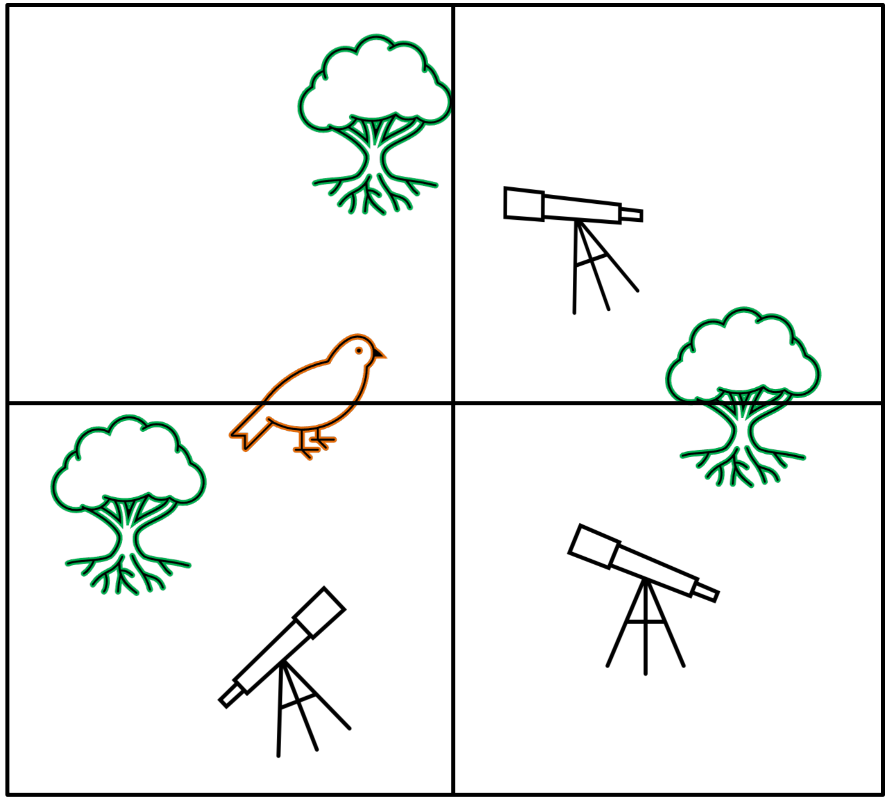
1. eBird Best Practices

- Same observer, same latitude-longitude coordinate, > 1 visit and at most 10 visits

Retains less than 25% of available data!

2. Grid

- Most commonly, 1x1km



Our Proposal

- Can we improve upon the existing methods by framing the **Site Clustering Problem** as a spatial clustering problem?

Oregon State University
College of Engineering

Existing Spatial Clustering Algorithms

- k-means (Llyod, 1982)
- CLARANS (Ng & Han, 2002)

Partitioning

- DBSCAN (Ester et al., 1996)
- DBRS (Wang & Hamilton, 2003)

Density Based

- SKATER (Assunção et al., 2006)
- REDCAP (Guo, 2008)

Regionalization

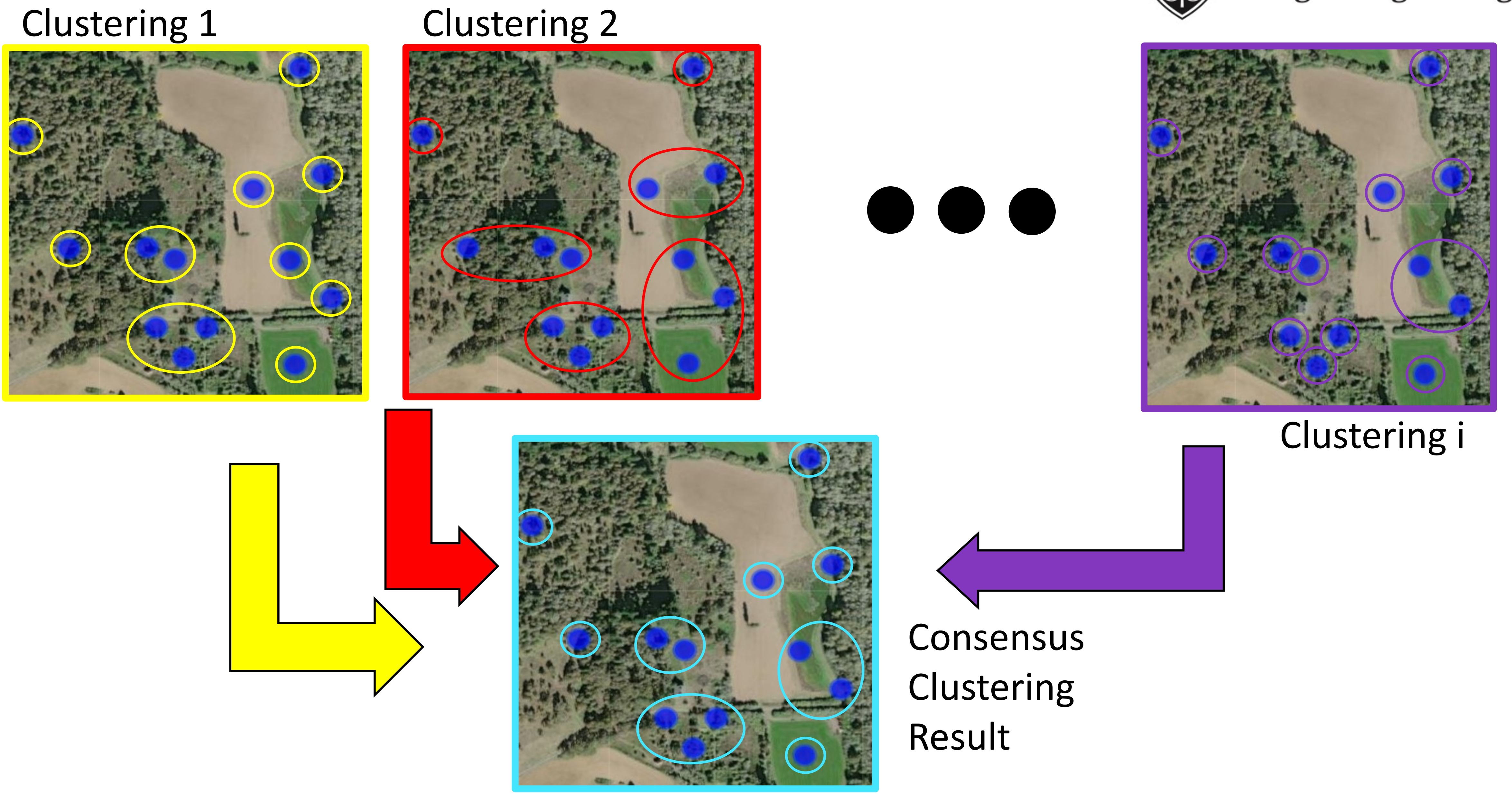
- For a more complete review see Liu et al.

Algorithms in this proposal

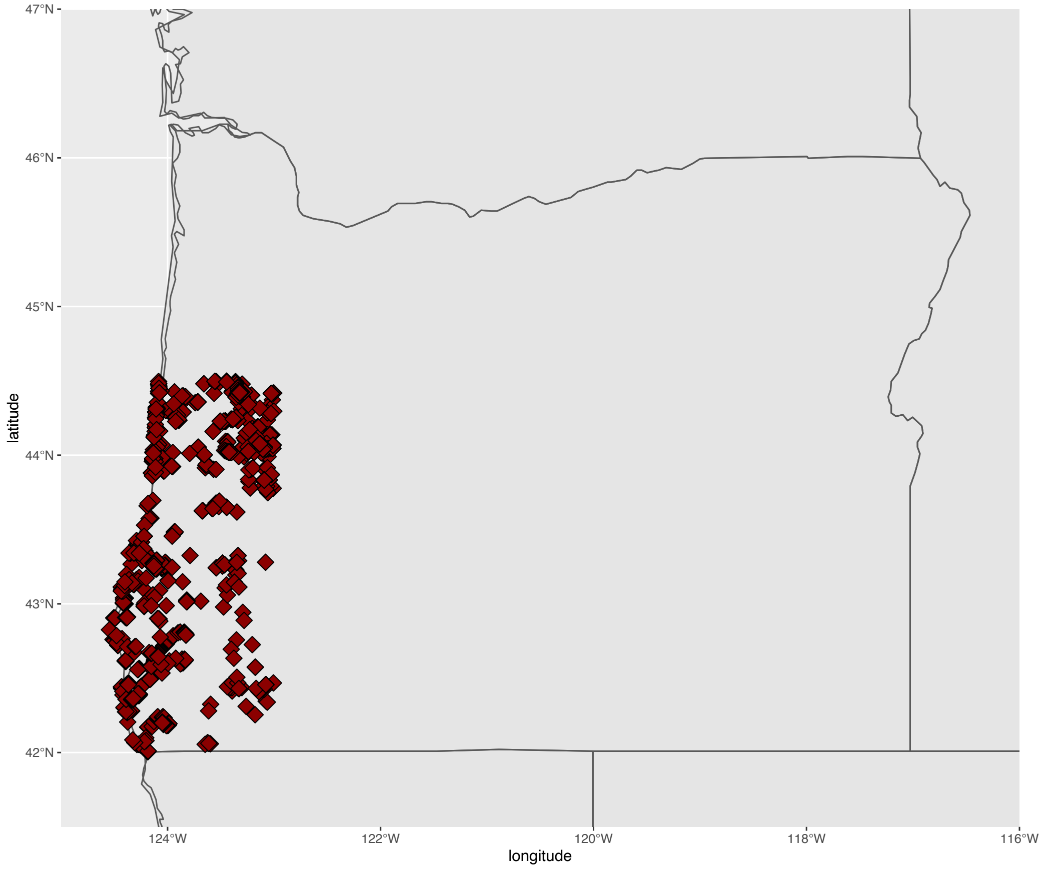
- lat-long
- rounded-4
- Density-based spatially-constrained (DBSC) (Liu et al., 2012)
- clustGeo (Chavent et al., 2018)
- Consensus Clustering
  - Agglomerative & Balls Implementations (Gionis et al., 2007)

Consensus Clustering

Oregon State University
College of Engineering



Experimental Setup



- 2,146 eBird checklists
  - Collected between May and July 2017
  - Remotely sensed environmental variables at each checklist
- Manually constructed a ground truth clustering
- Simulated occupancy and detection probabilities for each checklist

Experimental Setup

Oregon State University
College of Engineering

$$occ\,prob = .75 * var_1 - 1.25 * var_2 + .1 * var_3$$

Evaluation

- Predictive Accuracy
  - Mean squared error (MSE) of occupancy probability
- External Validation
  - Similarity to ground truth clustering
    - Adjusted Rand Index (ARI), Adjusted Mutual Information (AMI), Normalized Information Distance (NID) (Vinh et al. 2010)

Results

Oregon State University College of Engineering

	ARI	AMI	NID	occ MSE
ground truth	1.0	1.0	0	.0389 ± .015
eBird-BP	-	-	-	.1177 ± .041
1-kmSq	.9948	.9401	.0599	.1065 ± .027
lat-long	.9992	.9825	.0175	.0422 ± .017
rounded-4	.9992	.9826	.0174	.0424 ± .017
density-based	.9806	.9566	.0434	.1193 ± .031
clustGeo	.9994	.9909	.0091	.0460 ± .019
CC-agglom	.9992	.9835	.0166	.0421 ± .017
CC-balls	.9992	.9834	.0165	.0422 ± .017

* inputs for both CC algorithms were *lat-long*, *density-based*, *rounded-4*

References

Oregon State University College of Engineering

Assunção, R. M., Neves, M. C., Câmara, G., and Freitas, C. D. C. Efficient regionalization techniques for socioeconomic geographical units using minimum spanning trees. International Journal of Geographical Information Science, 20(7):797–811, 2006.

Chavent, M., Kuentz-Simonet, V., Labenne, A., and Saracco, J. Clustgeo: an r package for hierarchical clustering with spatial constraints. Computational Statistics, 33(4): 1799–1822, Jan 2018.

Ester, M., Kriegel, H.-P., Sander, J., and Xu, X. A densitybased algorithm for discovering clusters in large spatial databases with noise. pp. 226–231, 1996.

Gionis, A., Mannila, H., and Tsaparas, P. Clustering aggregation. ACM Trans. Knowl. Discov. Data, 1(1):4–es, March 2007. ISSN 1556-4681.

Guillera-Arroita, G., Lahoz-Monfort, J., MacKenzie, D. I., Wintle, B. A., and McCarthy, M. A. Ignoring imperfect detection in biological surveys is dangerous: a response to ’fitting and interpreting occupancy models’. PloS one, 9(7):e99571–e99571, 07 2014.

Guo, D. Regionalization with dynamically constrained agglomerative clustering and partitioning (redcap). International Journal of Geographical Information Science, 22 (7):801–823, 2008.

Liu, Q., Deng, M., Shi, Y., and Wang, J. A density-based spatial clustering algorithm considering both spatial proximity and attribute similarity. Computers Geosciences, 46:296–309, 2012.

Lloyd, Stuart P. "Least squares quantization in PCM." Information Theory, IEEE Transactions on 28.2 (1982): 129-137.

Ng, R. and Han, J. Clarans: a method for clustering objects for spatial data mining. IEEE Transactions on Knowledge and Data Engineering, 14(5):1003–1016, 2002.

Vinh, N. X., Epps, J., and Bailey, J. Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance. J. Mach. Learn. Res., 11:2837–2854, December 2010. ISSN 1532-4435.

Thank You!

- rothmark@oregonstate.edu
- @rothm_osu on Twitter