Spatial Clustering of Citizen Science Data Improves Downstream Species Distribution Models

Nahian Ahmed[1], Mark Roth[1], Tyler A. Hallman[3], W. Douglas Robinson[2], Rebecca A. Hutchinson[1, 2]

[1]School of Electrical Engineering and Computer Science, Oregon State University, Corvallis, OR 97331, USA
[2]Department of Fisheries, Wildlife, and Conservation Sciences, Oregon State University, Corvallis, OR 97331, USA
[3]School of Natural Sciences, Bangor University, Bangor LL57 2DG, UK
ahmedna@oregonstate.edu, roth.markh@gmail.com, t.hallman@bangor.ac.uk,
{douglas.robinson, rah}@oregonstate.edu

Abstract

Citizen science biodiversity data present great opportunities for ecology and conservation across vast spatial and temporal scales. However, the opportunistic nature of these data lacks the sampling structure required by modeling methodologies that address a pervasive challenge in ecological data collection: imperfect detection, i.e., the likelihood of under-observing species on field surveys. Occupancy modeling is an example of an approach that accounts for imperfect detection by explicitly modeling the observation process separately from the biological process of habitat selection. This produces species distribution models that speak to the pattern of the species on a landscape after accounting for imperfect detection in the data, rather than the pattern of species observations corrupted by errors. To achieve this benefit, occupancy models require multiple surveys of a site across which the site's status (i.e., occupied or not) is assumed constant. Since citizen science data are not collected under the required repeated-visit protocol, observations may be grouped into sites *post hoc*. Existing approaches for constructing sites discard some observations and/or consider only geographic distance and not environmental similarity. In this study, we compare ten approaches for site construction in terms of their impact on downstream species distribution models for 31 bird species in Oregon, using observations recorded in the eBird database. We find that occupancy models built on sites constructed by spatial clustering algorithms perform better than existing alternatives.

Code — https://github.com/Hutchinson-Lab/Spatial-Clustering-for-SDM

Datasets — https://doi.org/10.5281/zenodo.14362178

Extended version — https://arxiv.org/abs/2412.15559

Introduction

Species distribution models (SDMs) combine species observations with environmental data to produce estimates of species patterns across landscapes (Elith and Leathwick 2009). SDMs are important tools for ecological science and natural resource management. Examples include analysis of avian population declines (Betts et al. 2022; Rosenberg et al. 2019), assessments of species' IUCN Red List status (Syfert et al. 2014), and decision support for species translocation programs under climate change (Barlow et al. 2021). These models and their applications operate at a variety of spatial scales, from global analyses of species ranges (Cole et al. 2023) to regional assessments that drive local-scale conservation action (Rugg, Jenkins, and Lesmeister 2023). This paper is particularly motivated by science and conservation questions at the local-to-regional scale, which often require inference about fine-scale habitat features and the corrections for observational error that we describe below.

A pervasive challenge in species distribution modeling stems from the inherent difficulty of observing all organisms present at a given location when completing a survey. Many species are secretive, camouflaged, and/or ephemeral, so species are often under-reported. This is known as the problem of *imperfect detection*, which is common to both expert- and volunteer-led surveys. A family of models and associated sampling schemes has been developed in ecology to address this issue. A key idea is to collect multiple observations of a location, or *site*, during a period when the species status remains constant; variation in observations across this period then speaks to the observational process itself. A foundational member of this family of approaches is the *occupancy model*, which links environmental features (e.g., elevation, land cover) to a binary latent variable representing the species occupancy at each site. Then the multiple observations at each site depend both on the true occupancy status and a set of detection-related features (e.g., time of day, ambient noise) to correct for imperfect detection. This framework has been extended beyond static, binary representations of occupancy to species dynamics and abundance (Bailey, MacKenzie, and Nichols 2014).

Citizen science (CS) programs engage volunteers to collect large-scale biodiversity datasets, providing exciting opportunities for machine learning where ecology meets 'big data' (Beery et al. 2021; Johnston, Matechou, and Dennis 2023). The eBird project gathers checklists of birds daily across the globe (Sullivan et al. 2014); a sample of these data are analyzed below. Other CS Programs include eButterfly, which collects butterfly observations with a structure similar to eBird (Prudic et al. 2017), and iNaturalist, which relies on photo-based observations of biodiversity (iNaturalist).

While citizen science datasets have great potential to in-

form science and policy, a challenge arises when building SDMs from these data: they are not collected with the multiple-observation protocol developed for models that account for imperfect detection. Ignoring the consequences of imperfect detection can negatively impact SDMs (Guillera-Arroita et al. 2014; Lahoz-Monfort, Guillera-Arroita, and Wintle 2014). To leverage the strengths of CS data while still accounting for imperfect detection, one can form the multiple-observation structure *post hoc* from opportunistic species reports; this is the *site clustering problem* (Roth et al. 2021). The sites created to solve this problem might be thought of as observational units, but ecologically, they may also have connections to ideas about species' home ranges or territory sizes. A variety of approaches for site clustering exist (Johnston et al. 2021; von Hirschheydt, Stofer, and Kéry 2023; Hochachka, Ruiz-Gutierrez, and Johnston 2023). Some potential disadvantages of existing approaches include overly stringent constraints on what may constitute a site, the need to discard some data points that do not fit into the site definitions, and the inability to consider environmental features as well as geographic information. Spatial clustering techniques from machine learning (ML) have the potential to improve upon existing methods by incorporating both environmental and geographic similarity measures.

This paper offers an empirical study of ten approaches to the site clustering problem, drawn from both the ecology and machine learning literature. We show that occupancy models are sensitive to the choice of site clustering and that the ML approaches perform well. Our specific contributions are:

- We provide an empirical analysis with open-source data and code to compare solutions to the site clustering problem.
- We find evidence in support of approaches that (1) keep all data points rather than discarding some and (2) incorporate environmental features.
- We investigate a method for automatically tuning parameters to reduce the modeling burden for practitioners.

Background

Occupancy Modeling. Motivated by the need to account for imperfect detection of organisms on surveys, occupancy models simultaneously represent the species occurrence pattern (and its relationship to environmental features) along with the species observation pattern (and its relationship to detection-related features) (MacKenzie et al. 2002; Bailey, MacKenzie, and Nichols 2014). Occupancy models define a binary latent variable Z_i for each site $i = 1, ..., M$ that represents whether or not the species occurs there, and this is linked to occupancy features X_i which encode environmental habitat information in the style of a logistic regression: $p(Z_i = 1) = \psi_i = \sigma(\beta^T X_i)$, where $\sigma()$ denotes the logistic function. Fig. 1 shows the graphical view of the latent variable model. Detection probabilities p_{it} for repeated observations $t = 1, ..., T_i$ of each site are linked similarly to observation-related features W_{it} (e.g., time of day, weather, observer expertise): $p_{it} = \sigma(\gamma^T W_{it})$. The observations Y_{it} link the occupancy and detection components of the model such that $p(Y_{it} = 1) = Z_i p_{it}$. The parameters $\{\beta, \gamma\}$ of the model can be fit by maximum likelihood estimation.

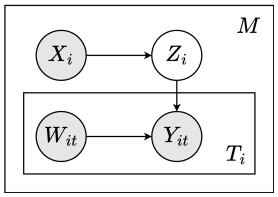

Figure 1: Graphical representation of occupancy model. Latent variable $Z_i \in \{0, 1\}$ represents occupancy at site $i = 1, ..., M$ and $Y_{it} = \{0, 1\}$ represents the observation during $t = 1, ..., T_i$. X_i represent site features and W_{it} represent survey features.

Implicit in these equations are key assumptions of occupancy models. In particular, each site has a single value for its occupancy status that remains constant across repeated (imperfect) observations t; this is the *closure* assumption. The closure assumption is of special importance to the current paper, since our task of interest is to group opportunistically-collected biodiversity reports into sites that are suitable for occupancy modeling *post hoc*. Violation of the closure assumption can lead to biased estimates of occupancy probability (Rota et al. 2009). We note that closure naturally has a temporal dimension as well; we follow common practice of identifying a time period of minimal distributional change for the species of interest (e.g., the breeding season for birds) and focus on the spatial aspects of the problem here. In addition to the closure assumption, occupancy models also assume no false positive observations; if Z_i is 0, $p(Y_{it} = 1)$ is also 0 regardless of the value of p_{it}. Extensions to the occupancy modeling framework that relax the assumption of no false positives exist (Royle and Link 2006; Miller et al. 2011; Hutchinson, He, and Emerson 2017), but they are beyond the scope of this paper.

Single-Visit Approaches. While occupancy models are designed for repeated observations to sites, some work has investigated the idea of applying occupancy models to individual observations, which we refer to here as Single Visit (SV) models. The appeal of the SV approach is that the closure assumption is satisfied automatically, since there are no repeated visits to consider. The concern that arises with the SV approach is parameter identifiability; in the classical occupancy model, repeated visits are necessary to identify the occupancy and detection probabilities separately. Lele et al. (2012) argued that occupancy models could be applied in the SV setting under certain conditions on the occupancy and detection features that essentially require that the two feature sets be sufficiently different. Knape et al. (2015) expressed concern that the assumptions underpinning that work were unrealistic, and Sólymos et al. (2016) responded by clarifying the assumptions and reiterating the case for the potential of SV approaches. More recently, Stoudt et al. (2023) provided another argument against identifability in SV approaches based on ideas from econometrics. In our experiments below, we include the SV approach for completeness, but these concerns suggest that practitioners should take caution with this method.

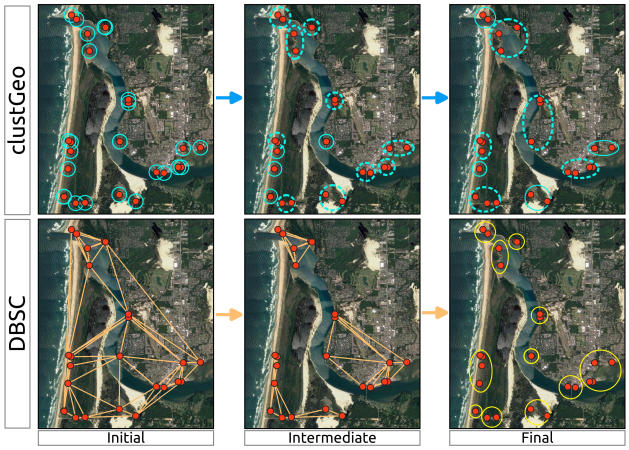

Figure 2: Simulated example of site formation using clustGeo and DBSC clustering algorithms. eBird observation locations from southwest Oregon, United States, are shown as red dots overlaid on satellite imagery from the corresponding region. clustGeo aggregates points iteraritvely and stops when the desired number of clusters is reached. Newly created clusters at each step are shown using bold dashed circles and ellipses. DBSC constructs a Delaunay Triangulation (shown using orange triangles) and then splits it based on spatial constraints and feature similarity.

Site Clustering Problem. The concerns about trivially satisfying the closure assumption with SV approaches and the known problems with violations of the closure assumption underpin the importance of the *site clustering problem*, introduced by Roth et al. (2021). A key difference between this problem and typical clustering settings is that the quality of the clustering cannot simply be formulated as a mathematical objective. Instead of measuring quality via similarity metrics among points within clusters or dissimilarity between points across clusters, we are interested in how the clustering influences performance on downstream tasks like occupancy modeling. Given a set of observations at geolocated points, the objective of the site clustering problem is to construct a set of clusters $\mathcal{C}$ which optimize performance evaluation metrics for a downstream model. For example, in this paper, we seek a clustering that optimizes the area under the receiver operating characteristic curve (AUC) for predictions of held-out observations made by occupancy models.

Spatial Clustering Approaches. A variety of clustering methods for spatial data exist in the machine learning literature (Ng and Han 1994). Here, we outline two spatial clustering approaches that have potential application to the site clustering problem.

First, clustGeo is a hierarchical, agglomerative spatial clustering method (Chavent et al. 2018). It calculates the distance d between two objects, i_1 and i_2, using a weighted combination of two Euclidean distance metrics: $d(i_1, i_2) = \alpha d_1(i_2, i_2) + (1-\alpha) d_2(i_1, i_2)$. Here, d_1 measures geospatial distance, d_2 measures environmental distance, and $\alpha \in [0, 1]$ is a parameter weighting the relative importance of these two. The algorithm iteratively merges the most similar objects (top row of subfigures in Fig. 2) and stops when a specified number of clusters, controlled by the parameter λ, is reached. For instance, $\lambda = 80$ sets the number of clusters created to approximately 80% of the number of unique locations.

Second, density based spatial clustering (DBSC) uses a two-step, divisive clustering approach and is able to discover clusters of arbitrary shapes and sizes (Liu et al. 2012). The first step imposes spatial constraints on the clustering process. A Delaunay Triangulation (DT) graph of the points is constructed (bottom row of subfigures in Fig. 2). A DT graph consists of triangles where the minimum angle between points is maximized. Long edges, which have lengths exceeding a threshold based on the average length of edges in the DT graph, are removed to form spatially disjoint DT subgraphs. These subgraphs are split again in a similar manner, but this time the long edges are defined based on the localized characteristics of edges in the DT subgraphs. The final partitions are used to construct the clustering. Points in the same partitions are candidates for being in the same cluster. The second step clusters the points in the partitions based on their feature similarity while enforcing the spatial constraints from the first step.

Bayesian Optimization. The primary objective of Bayesian Optimization routines is to optimize black-box functions (Snoek, Larochelle, and Adams 2012; Garnett 2023). We introduce this technique for tuning parameters of the clustering algorithm (for clustGeo in particular), to avoid requiring users to add another step to their modeling workflow. The optimization routine has two main components. The first component, the *acquisition function*, has the task of acquiring potential solutions over which fitness is to be evaluated. The second component, the *fitness function*, decides how fit the potential solutions are to optimizing our objective. The routine iterates between using the acquisition function to find the next potential solution to evaluate and the fitness function to gauge the effectiveness of the potential solution.

Candidate Clustering Approaches

In this study, we implemented and compared ten methods to address the site clustering problem. Similarities and differences among the methods are summarized in Table 1.

1. *SVS*: Single Visit Sites. Trivially, every data point is treated as a site with a single observation (i.e., a cluster of size 1). When points have identical coordinates, they are still treated as different sites.
2. *1/UL*: One per Unique Location. Every unique location is treated as a site. If there are multiple points with identical coordinates, one is chosen randomly to keep and the rest are discarded.
3. *lat-long*: Latitude-longitude. Points with the same latitude-longitude coordinates are assigned to the same site. Sites can have any number of observations (i.e., cluster size can range from 1 to any number of co-located points).

No.	Site-clustering approach	Can cluster size be > 1?	Might some points be excluded?	Can clusters have points w/ diff. geospatial coords.?	Is similarity in feature space considered?
1	SVS	No	No	No	No
2	1/UL	No	Yes	No	No
3	lat-long	Yes	No	No	No
4	2to10	Yes	Yes	No	No
5	2to10-sameObs	Yes	Yes	No	No
6	rounded-4	Yes	No	Yes	No
7	1-kmSq	Yes	No	Yes	No
8	best-clustGeo	Yes	No	Yes	Yes
9	BayesOptClustGeo	Yes	No	Yes	Yes
10	DBSC	Yes	No	Yes	Yes

Table 1: Properties of the ten candidate approaches to the site-clustering problem.

4. *2to10*: Approach with cluster size in 2-10. Constructs sites based on analytical guidelines for eBird data (Johnston et al. 2021). Points with identical coordinates form sites, but the number of observations per site is constrained to be within $[2, 10]$. Singletons and observations beyond the limit of 10 are discarded.
5. *2to10-sameObs*: Approach with cluster size in 2-10 and all records from the same observer. This is the same approach as *2to10* with the added requirement of having all points being recorded by the same observer.
6. *rounded-4*: Lat-long rounded to 4 decimal places. Points with the same coordinates after rounding latitude and longitude to the fourth decimal place are assigned to the same site.
7. *1-kmSq*: 1 square kilometer grid. This method overlays a grid with one square kilometer cells on the study area, and points falling within grid cells are assigned to the same site.
8. *best-clustGeo*: clustGeo with the best tuning parameters selected *post hoc*. Sites are clustered by the clustGeo algorithm. We set parameters using all possible combinations of parameters $\alpha = \{0.25, 0.5, 0.75\}$ and $\lambda = \{60, 70, 80, 90\}$. The final parameters reported for this method are the values that produced the best results at test time; this method essentially uses an oracle to determine the best that the clustGeo approach could perform.
9. *BayesOptClustGeo*: clustGeo with Bayesian optimization of parameter tuning. Sites are clustered by clustGeo algorithm, and the parameters α and λ are tuned via Bayesian optimization, requiring no manual input from the user nor additional cross-validation.
10. *DBSC*: Density-Based Spatial Clustering. Sites are built from clusters defined by the DBSC algorithm. Unlike clustGeo, this method has no externally tunable parameters.

Experimental Design

Data Selection and Pre-processing

Our study comprises data from the eBird basic dataset in a region of southwestern Oregon, USA collected in 2017 and 2018 during May 15th - July 9th of each year. This window corresponds to the breeding season of many bird species, which is a common focus of ecological analyses and a common choice of temporal window for meeting the occupancy modeling closure assumption. We obtained all complete checklists from the region, resulting in 2,497 checklists in 2017 and 3,490 in 2018 (Fig. 3). We linked the checklist data reporting detections vs. non-detections for each species to two sets of features. For modeling occupancy probabilities, we used five habitat features: elevation derived from a Digital Elevation Model (DEM), and Tasseled Cap Brightness, Tasseled Cap Greenness, Tasseled Cap Wetness, and Tasseled Cap Angle derived from Landsat data (Baig et al. 2014). For modeling detection probabilities, we used five detection features provided with the eBird dataset: day_of_year, time_observations_started, duration_minutes, effort_distance_km, and number_observers.

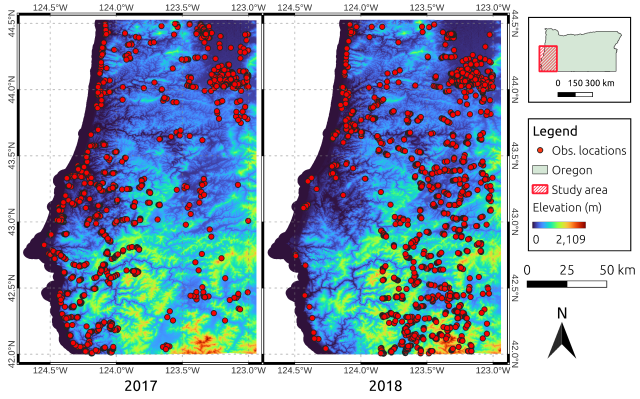

Figure 3: Observation locations from eBird checklists in 2017 and 2018 recorded over southwest Oregon, United States, are shown as red dots. Of these, there were 2,497 checklists at 1,314 unique locations in 2017, and 3,490 checklists at 1,519 unique locations in 2018.

We selected 31 species for analysis to represent a range of prevalences, degrees of conspicuousness, home range sizes, and habitat and diet breadths. Species were categorized as primarily inhabiting forested or non-forested habitats, such as grasslands and early seral habitats. They were further

classified as specialists or generalists based on habitat and diet diversity, with specialists occupying fewer habitat types and having a narrower diet (e.g., insectivores), and generalists occupying various habitats and consuming a wider range of foods (e.g., omnivores). Information on habitat specialization, diet, and home range sizes was sourced from published species life history reviews (Billerman et al. 2020). Species were grouped into three categories based on home range size (small, less than 2.5 ha; medium, 2.5 to 39 ha; large, 40 ha and greater) and three prevalence levels (low, less than 2.84%; medium, 2.84% to 13.12%; high, 14.42% and higher). Details on species names, taxonomic abbreviations, prevalence, and traits are provided in Table S1.

We followed the recommendations of Johnston et al. (2021) for eBird data filtering, pre-processing, and occupancy modeling. Specifically, we excluded checklists where the distance traveled exceeded 0.25 km and those from eBird "hotspots" to ensure location accuracy, as bird watchers might report the hotspot location rather than the precise nearby location.

Model Fitting

The only clustering approach with parameter tuning was *BayesOptClustGeo*, where parameters α and λ were selected via Bayesian optimization. We used the upper confidence bound (UCB) as the acquisition function and defined a custom fitness function for our problem. Specifically, we measured the Silhouette width averaged over all points used for constructing the clusterings and used that as unsupervised feedback for the Bayesian optimization routine. Silhouette width measures how similar points are to the clusters they are assigned to with respect to other clusters (Rousseeuw 1987); a higher value indicates a clustering where points are similar to the clusters they are in and dissimilar to the other clusters. We defined the similarity measure using a uniformly weighted Euclidean distance computed from the geospatial features (latitude and longitude) and environmental habitat features. We used the average Silhouette width of the clusterings formed by clustGeo based on the specified parameter combination as our fitness function. We ran 30 iterations of parameter acquisition followed by fitness evaluation, using real values in the ranges $\alpha = [0.01, 0.99]$ and $\lambda = [10, 90]$. This allowed for a more granular search over clustering parameter combinations compared to manually experimenting over a uniform grid of parameter values.

We fit occupancy models to the site structures produced by each clustering algorithm on the training data. Model parameters were fit via maximum likelihood estimation with the `unmarked` package (Fiske et al. 2015) in R version 4.4.2. Since test splits vary across repeats, we trained once and repeated the testing process 25 times per species, following Johnston et al. (2021).

Model Assessment

We used a temporally independent test set to measure the performance of occupancy models fit with different site structures. We trained all models on the 2017 checklists and used the 2018 checklists for testing. To form the testing dataset, we again followed the recommendations of Johnston et al. (2021). First, we split the 2018 checklists into detections and non-detections. Then we placed an equal area hexagonal grid with centers separated by distance of 5 km over our study region using the `dggridR` R package (Barnes and Sahr 2017), and spatially subsampled by keeping no more than two checklists from each hexagon (up to one detection and one non-detection). Trained models were evaluated 25 times on the spatially subsampled test set.

We compared the outputs of occupancy models fitted with different checklist clusterings based on their ability to predict held-out observations of detection vs. non-detection. We multiplied the occupancy and detection probabilities together to estimate observation probability, which we compared with the species observations from the test set to measure performance. We measured the area under the receiver operating characteristic curve (AUC) for each set of predictions. The results were summarized by calculating percentage AUC improvement over lat-long. For each species and for each test split, algorithm a has percentage AUC improvement, $\delta_a = ((AUC_a - AUC_{lat-long})/AUC_{lat-long}) \times 100$. We did a parallel assessment with area under the precision-recall curve (AUPRC), which can be an appropriate metric especially for more rare species.

We also analyzed the relationship between species traits and performance of the clustering algorithms by building linear mixed-effects models (Kuznetsova, Brockhoff, and Christensen 2017). These models treat species as a random effect and treat the interactions between species traits and clustering algorithms as the fixed effects. The non-intercept coefficients of these linear mixed effect models inform us of how general combinations of algorithms and species traits (interaction groups) affect model performance in terms of percentage AUC improvement, while factoring in species specific variance of performance. We built four such models to study the relationships between algorithm and percentage AUC improvement based on (i) prevalence level, (ii) home range size, (iii) habitat type, and (iv) whether the species is a generalist or a specialist. We built a fifth mixed effect model on the relationship between algorithm choice and percentage AUC improvement in general. This analysis aims to understand whether different clustering approaches might be preferred for different types of species.

Finally, we assessed the effects of the different clustering approaches qualitatively by examining predictive maps of occupancy probability across the region. We note that the metrics described above focus on predictive performance of the occupancy models, i.e., their ability to predict held-out observations of the species. However, while this is the metric available from the existing data, it is not the output of scientific interest from the model. The model of the latent occupancy process speaks to the actual biological process of interest, but the quality of this model is hard to evaluate because it is only observed through the lens of imperfect detection. Given the scientific importance of this aspect of the modeling, we constructed maps of the occupancy patterns predicted for each species from each clustering approach for visual inspection and qualitative evaluation.

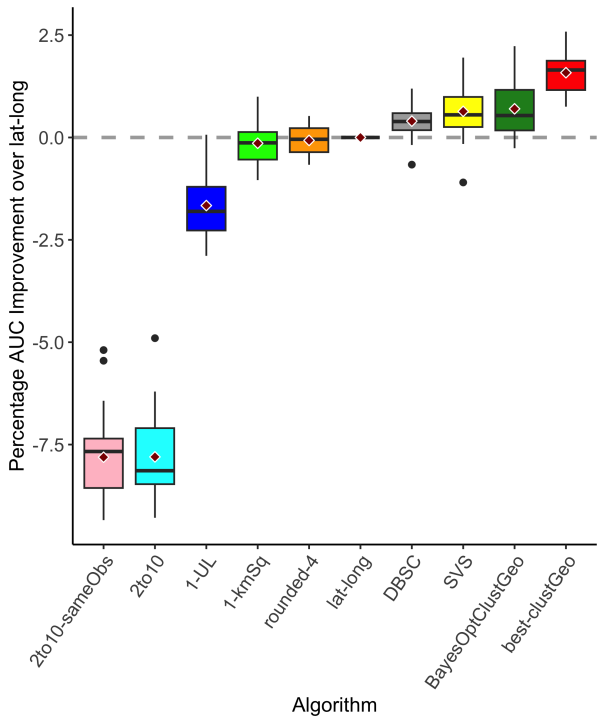

Figure 4: Boxplots show the percentage AUC improvement of each method over lat-long. Larger positive values indicate better performance than lat-long; negative values indicate worse performance than lat-long.

Results and Discussion

Predictive Performance. When measuring the performance of the clustering algorithms with AUC on held-out observations, the highest-performing approach varies across species, but some general trends point to the promise of the spatial clustering approaches. Fig. 4 presents performance of the site clustering approaches relative to the performance of lat-long. As expected, best-clustGeo has the highest overall mean and lowest variance since it is tuned to each species based on test performance; it is not directly comparable to the other methods and represents an upper limit on how well clustGeo might perform with optimal tuning. BayesOptClustGeo has the next highest mean performance, with substantial variation. Comparison between best-clustGeo and BayesClustGeo reveals the performance gap that results (at least partially) from the parameter tuning process, both in terms of the slightly lower mean performance and the higher variance. This suggests that further work on the parameter tuning procedure, ideally without requiring extra effort from modelers, may be fruitful. The other spatial clustering method, DBSC, also performs well, though slightly behind the clustGeo approaches. In contrast to best-clustGeo and BayesOptClustGeo, DBSC does not require prior knowledge about the number of sites to generate or how to weight distance metrics. The relatively good performance, combined with this ease of use, gives DBSC a potential advantage over clustGeo and BayesOptClustGeo since DBSC does not require parameter tuning, which may be challenging to incorporate into the modeling workflow. We see similar trends when we measure the percentage AUPRC improvement over lat-long (Fig. S6, S7).

The site clustering approaches that produce sites with single observations, SVS and 1/UL, show mixed results. In terms of this AUC-based metric, the SVS approach is a close competitor to the top performing clustering algorithm BayesOptClustGeo. However, as discussed above, literature suggests that this approach incurs substantial risk of non-identifiability of parameters. These identifiability problems may compromise scientific insight into the model without being detectable when performance is measured solely with predictive metrics. Recall that the 1/UL method is a special case of SVS that discards all but one data point at each location; its lower performance is likely attributable to smaller data set sizes. Despite the potential concerns surrounding these trivial solutions to the site clustering problem, we have included them for completeness.

The lat-long, rounded-4, and 1-kmSq methods make use of all data points and rely solely on geographic information to form sites. These approaches perform similarly to each other and form the 'middle of the pack' across the set of clustering algorithms. That these methods trail the spatial clustering algorithms suggests that there is benefit to be gained from considering environmental space as well as geographic space.

The site clustering approaches based on the eBird recommendations have negative values in Fig. 4, indicating weaker performance than lat-long. The requirements for defining sites in these methods may imply discarding too much data for the consequent models to remain competitive. 1/UL is the only other method that discards data, and these three methods rank last in predictive performance. Overall, our results indicate that methods which make use of all available observation data outperform methods which do not.

Recent work has similarly noted the utility of 'mixed' occupancy designs, meaning site structures that include some SV sites and some sites with multiple visits (or observations; MV). In particular, instead of discarding all SV sites and only keeping MV sites, including SV sites can increase the precision of occupancy estimates (von Hirschheydt, Stofer, and Kéry 2023; Hochachka, Ruiz-Gutierrez, and Johnston 2023). In our comparison, lat-long, 1-kmSq, rounded-4, DBSC, best-clustGeo, and BayesOptClustGeo all allow the creation of such mixed occupancy designs.

While most of the candidate clustering approaches here produce the same site clusterings for all species, analysis of best-clustGeo (the 'oracle' method), which does provide species-specific clusterings, suggests directions for further improvements. In this study, the training data points have the same geospatial coordinates and environmental habitat features for all 31 species. However, the species observations vary across those points, producing problem instances with different prevalence rates, or class balances. The only method that uses information about the species observations is best-clustGeo, where the parameter settings are chosen based on test set performance; best-clustGeo is the only method with species-specific clustering. This provides clues

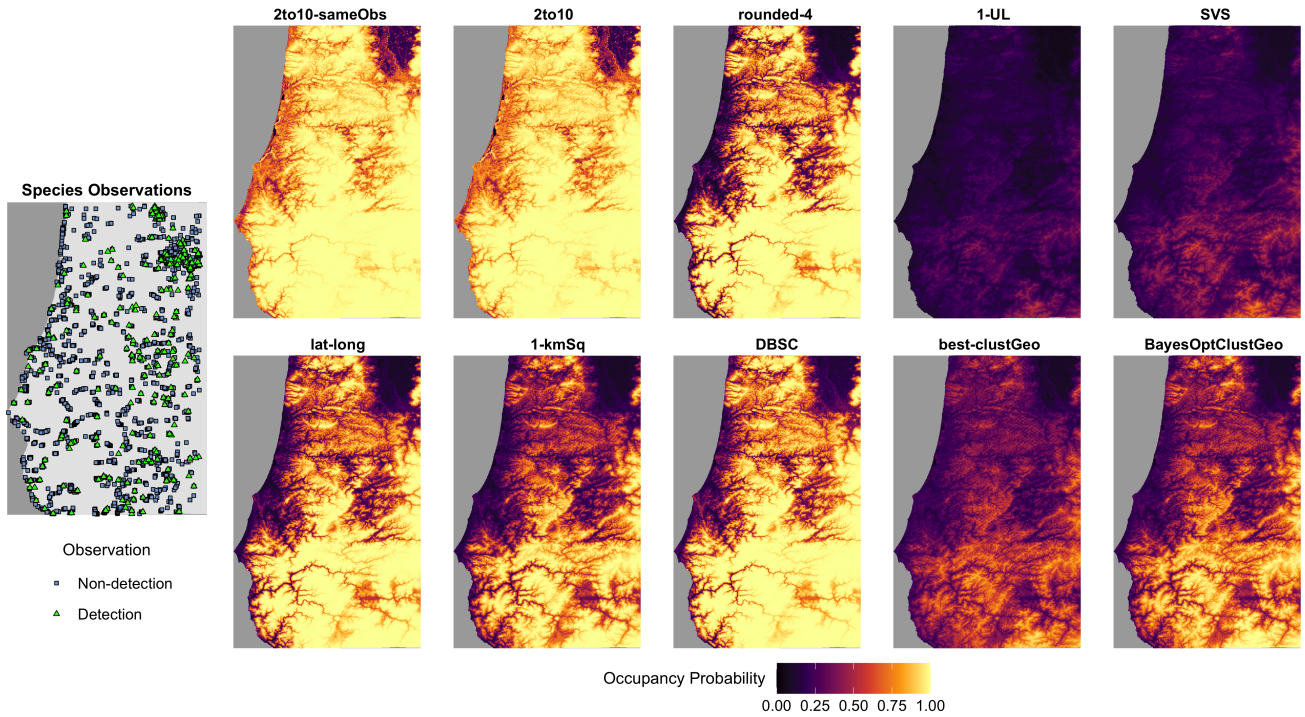

Figure 5: Occupancy probability of Northern Flicker (*Colaptes auratus*) over southwestern Oregon, United States predicted by species distribution models built from sites produced by ten clustering algorithms.

into potential directions for future work. We found that λ had a bigger influence than α and that the performance of clustGeo parameterizations on different species was not uniform; i.e., the optimal parameter values varied across species (Fig. S1, S2; Table S2). Thus, a potential direction for future work is to fold species-specific information into the Bayesian optimization routine.

Effects of Species Traits. The mixed-effects models provided preliminary insights into the interactions between species traits and site clustering approaches, with additional support for spatial clustering methods. Detailed results are in the supplemental material (Fig. S9); here, we summarize general trends. Clustering algorithms performed better on species that had low prevalence rates, large home range sizes, lived in forested habitats, and were specialists. best-clustGeo and BayesOptClustGeo are parts of interaction groups with the highest effects on percentage AUC improvement over lat-long. 2to10 and 2to10-sameObs are frequently parts of interactions groups which negatively impact AUC. The ordering of algorithms in Fig. 4 is mirrored by the coefficients of the mixed effect model linking algorithm choice and raw AUC (Fig. S8).

Qualitative Results. While the results above judge performance based on predictions of held-out observations, recall that the scientific interest in occupancy models centers instead on estimates of the latent variable, which are challenging to evaluate. We can at least visualize differences in the estimates provided by occupancy models when supplied with data shaped by the different site clustering approaches. Fig. 5 provides an example of the variation across methods for Northern Flicker (*Colaptes auratus*). While further expert analysis is required to gauge reliability of these maps, it is worth noting the variability in overall magnitude and spatial distribution of occupancy probability across clustering approaches. For most study species, the clustering approach had visually apparent effects on the occupancy estimates that inform science and policy in ecology and conservation.

Conclusion

This study explored the role of clustering of opportunistic biodiversity observations as a precursor to species distribution modeling. We evaluated ten approaches to this task and provided insight for future directions. Both the predictive and qualitative results show that models are sensitive to the design choices made at the clustering stage of the analytic workflow. Corroborating other work in the ecology literature, we find that clustering approaches which exclude some data points are outperformed by those that do not. Spatial clustering algorithms from the machine learning literature can incorporate environmental feature space as well as geographic space, and they show promising results in our comparative evaluation. Future work on this topic should focus on species-specific selection of clustering parameters while minimizing additional burden to modeling practitioners.

Acknowledgments

This research was supported by the National Science Foundation (NSF) under Grant No. III-2046678 (NA, MR, RAH), and the Bob and Phyllis Mace professorship (WDR).

References

Baig, M. H. A.; Zhang, L.; Shuai, T.; and Tong, Q. 2014. Derivation of a tasselled cap transformation based on Landsat 8 at-satellite reflectance. *Remote Sensing Letters*, 5(5): 423–431.

Bailey, L. L.; MacKenzie, D. I.; and Nichols, J. D. 2014. Advances and applications of occupancy models. *Methods in Ecology and Evolution*, 5(12): 1269–1279.

Barlow, M. M.; Johnson, C. N.; McDowell, M. C.; Fielding, M. W.; Amin, R. J.; and Brewster, R. 2021. Species distribution models for conservation: identifying translocation sites for eastern quolls under climate change. *Global Ecology and Conservation*, 29: e01735.

Barnes, R.; and Sahr, K. 2017. dggridR: Discrete Global Grids for R. *R package version 2.0.4*.

Beery, S.; Cole, E.; Parker, J.; Perona, P.; and Winner, K. 2021. Species distribution modeling for machine learning practitioners: A review. In *Proceedings of the 4th ACM SIGCAS Conference on Computing and Sustainable Societies*, 329–348.

Betts, M. G.; Yang, Z.; Hadley, A. S.; Smith, A. C.; Rousseau, J. S.; Northrup, J. M.; Nocera, J. J.; Gorelick, N.; and Gerber, B. D. 2022. Forest degradation drives widespread avian habitat and population declines. *Nature Ecology & Evolution*, 6(6): 709–719.

Billerman, S.; Keeney, B.; Rodewald, P.; Schulenberg, T.; et al. 2020. Birds of the World. *Cornell Laboratory of Ornithology, Ithaca, NY, USA*.

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

Cole, E.; Van Horn, G.; Lange, C.; Shepard, A.; Leary, P.; Perona, P.; Loarie, S.; and Mac Aodha, O. 2023. Spatial implicit neural representations for global-scale species mapping. In *International Conference on Machine Learning*, 6320–6342. PMLR.

Elith, J.; and Leathwick, J. R. 2009. Species distribution models: ecological explanation and prediction across space and time. *Annual review of ecology, evolution, and systematics*, 40(1): 677–697.

Fiske, I.; Chandler, R.; Miller, D.; Royle, A.; Kery, M.; Hostetler, J.; Hutchinson, R.; and Royle, M. A. 2015. Package 'unmarked'. *R Project for Statistical Computing*.

Garnett, R. 2023. *Bayesian optimization*. Cambridge University Press.

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

Hochachka, W. M.; Ruiz-Gutierrez, V.; and Johnston, A. 2023. Considerations for fitting occupancy models to data from eBird and similar volunteer-collected data. *Ornithology*, 140(4): ukad035.

Hutchinson, R.; He, L.; and Emerson, S. 2017. Species distribution modeling of citizen science data as a classification problem with class-conditional noise. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31.

iNaturalist. n.d. Available from https://www.inaturalist.org. Accessed: 14 August 2024.

Johnston, A.; Hochachka, W. M.; Strimas-Mackey, M. E.; Ruiz Gutierrez, V.; Robinson, O. J.; Miller, E. T.; Auer, T.; Kelling, S. T.; and Fink, D. 2021. Analytical guidelines to increase the value of community science data: An example using eBird data to estimate species distributions. *Diversity and Distributions*, 27(7): 1265–1277.

Johnston, A.; Matechou, E.; and Dennis, E. B. 2023. Outstanding challenges and future directions for biodiversity monitoring using citizen science data. *Methods in Ecology and Evolution*, 14(1): 103–116.

Knape, J.; and Korner-Nievergelt, F. 2015. Estimates from non-replicated population surveys rely on critical assumptions. *Methods in Ecology and Evolution*, 6(3): 298–306.

Kuznetsova, A.; Brockhoff, P. B.; and Christensen, R. H. B. 2017. lmerTest package: tests in linear mixed effects models. *Journal of statistical software*, 82(13).

Lahoz-Monfort, J. J.; Guillera-Arroita, G.; and Wintle, B. A. 2014. Imperfect detection impacts the performance of species distribution models. *Global ecology and biogeography*, 23(4): 504–515.

Lele, S. R.; Moreno, M.; and Bayne, E. 2012. Dealing with detection error in site occupancy surveys: what can we do with a single survey? *Journal of Plant Ecology*, 5(1): 22–31.

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

MacKenzie, D. I.; Nichols, J. D.; Lachman, G. B.; Droege, S.; Andrew Royle, J.; and Langtimm, C. A. 2002. Estimating site occupancy rates when detection probabilities are less than one. *Ecology*, 83(8): 2248–2255.

Miller, D. A.; Nichols, J. D.; McClintock, B. T.; Grant, E. H. C.; Bailey, L. L.; and Weir, L. A. 2011. Improving occupancy estimation when two types of observational error occur: Non-detection and species misidentification. *Ecology*, 92(7): 1422–1428.

Ng, R. T.; and Han, J. 1994. Efficient and effective clustering methods for spatial data mining. In *Proceedings of VLDB*, 144–155. Citeseer.

Prudic, K. L.; McFarland, K. P.; Oliver, J. C.; Hutchinson, R. A.; Long, E. C.; Kerr, J. T.; and Larrivée, M. 2017. eButterfly: leveraging massive online citizen science for butterfly conservation. *Insects*, 8(2): 53.

Rosenberg, K. V.; Dokter, A. M.; Blancher, P. J.; Sauer, J. R.; Smith, A. C.; Smith, P. A.; Stanton, J. C.; Panjabi, A.; Helft,

L.; Parr, M.; et al. 2019. Decline of the North American avifauna. *Science*, 366(6461): 120–124.

Rota, C. T.; Fletcher Jr, R. J.; Dorazio, R. M.; and Betts, M. G. 2009. Occupancy estimation and the closure assumption. *Journal of Applied Ecology*, 46(6): 1173–1181.

Roth, M.; Hallman, T.; Robinson, W. D.; and Hutchinson, R. 2021. On the Role of Spatial Clustering Algorithms in Building Species Distribution Models from Community Science Data. In *ICML 2021 Workshop on Tackling Climate Change with Machine Learning*.

Rousseeuw, P. J. 1987. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20: 53–65.

Royle, J. A.; and Link, W. A. 2006. Generalized site occupancy models allowing for false positive and false negative errors. *Ecology*, 87(4): 835–841.

Rugg, N. M.; Jenkins, J. M.; and Lesmeister, D. B. 2023. Western screech-owl occupancy in the face of an invasive predator. *Global Ecology and Conservation*, 48.

Snoek, J.; Larochelle, H.; and Adams, R. P. 2012. Practical bayesian optimization of machine learning algorithms. *Advances in neural information processing systems*, 25.

Sólymos, P.; and Lele, S. R. 2016. Revisiting resource selection probability functions and single-visit methods: Clarification and extensions. *Methods in Ecology and Evolution*, 7(2): 196–205.

Stoudt, S.; de Valpine, P.; and Fithian, W. 2023. Nonparametric Identifiability in Species Distribution and Abundance Models: Why it Matters and how to Diagnose a Lack of it Using Simulation. *Journal of Statistical Theory and Practice*, 17(3): 39.

Sullivan, B. L.; Aycrigg, J. L.; Barry, J. H.; Bonney, R. E.; Bruns, N.; Cooper, C. B.; Damoulas, T.; Dhondt, A. A.; Dietterich, T.; Farnsworth, A.; et al. 2014. The eBird enterprise: An integrated approach to development and application of citizen science. *Biological conservation*, 169: 31–40.

Syfert, M. M.; Joppa, L.; Smith, M. J.; Coomes, D. A.; Bachman, S. P.; and Brummitt, N. A. 2014. Using species distribution models to inform IUCN Red List assessments. *Biological Conservation*, 177: 174–184.

von Hirschheydt, G.; Stofer, S.; and Kéry, M. 2023. "Mixed" occupancy designs: When do additional single-visit data improve the inferences from standard multi-visit models? *Basic and Applied Ecology*, 67: 61–69.