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

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

Abstract

This paper discusses opportunities for developments in spatial clustering methods to help leverage broad scale community science data for building species distribution models (SDMs). SDMs are tools that inform the science and policy needed to mitigate the impacts of climate change on biodiversity. Community science data span spatial and temporal scales unachievable by expert surveys alone, but they lack the structure imposed in smaller scale studies to allow adjustments for observational biases. Spatial clustering approaches can construct the necessary structure after surveys have occurred, but more work is needed to ensure that they are effective for this purpose. In this proposal, we describe the role of spatial clustering for realizing the potential of large biodiversity datasets, how existing methods approach this problem, and ideas for future work.

1. Introduction

Species distribution models (SDMs) are critical tools for mitigating the impact of climate change on biodiversity. SDMs link environmental variables to species observations to infer key characteristics of habitat requirements and to predict where species can persist. As ecosystems change, scientists and natural resource managers rely on SDMs to inform conservation policy decisions like designing reserves for threatened species and to predict how species will react to global change. Our ability to effectively protect species against extinction relies on high-quality SDMs. This tool can be built from a variety of biodiversity datasets, but community science programs like eBird and iNaturalist are growing in size, quality, and importance. A strength of these programs is that they welcome contributions from a diverse population of wildlife observers with varying experience. These crowdsourced datasets present exciting opportunities to understand the Earth's changing ecosystems at scales previously unattainable, but further methodological research is needed to fully leverage their potential. In this proposal, we outline the role that spatial clustering algorithms can play in improving SDMs built with community science data.

Community science data are impacted by *imperfect detection*: the common phenomenon in which observers do not detect all individuals and/or species present during a survey. Even expert observers encounter this issue—sometimes, birds are silent and hidden, for example—but community scientists with less training are even more impacted. To account for imperfect detection, studies conducted by experts use a careful sampling design that allows them to simultaneously estimate the probability that a species occupies a location and the probability that the observer detects the species given that it is present. This methodology is referred to as *occupancy modeling*, and it has become the dominant approach in statistical ecology for correcting the effects of imperfect detection (MacKenzie et al., 2002; Bailey et al., 2014). Ignoring imperfect detection can lead to biased estimates of species distributions, so occupancy models are a necessary tool for drawing accurate conclusions from biodiversity data (Guillera-Arroita et al., 2014; Lahoz-Monfort et al., 2014). Our focal problem in this paper is the fact that community science data are not collected according to the sampling design prescribed by occupancy models. Instead, contributors report observations at the times and places of their choosing.

We present the first attempt to frame the challenge of creating sites for occupancy models from unstructured community science data as a spatial clustering problem, which we introduce as the **Site Clustering Problem**. Below, we examine existing methods that at least partially address the problem and explore some spatial clustering approaches to solve it in a case study. We illustrate where this research lies on the critical path that begins at data collection and ends with climate change mitigation (Figure 1). Improved solutions to this problem will produce more accurate models and have direct impacts to biodiversity conservation and

[1]Department of Electrical Engineering Computer Science, Oregon State University, USA [2]Department of Fisheries, Wildlife, Conservation Sciences, Oregon State University, USA [3]Swiss Ornithological Institute, Sempach, Switzerland. Correspondence to: Mark Roth <rothmark@oregonstate.edu>.

Proceedings of the 38^{th} International Conference on Machine Learning, PMLR 139, 2021.

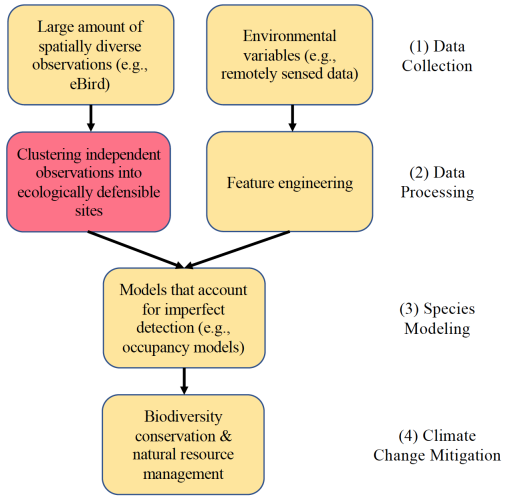

Figure 1. The proposed **Site Clustering Problem** (in red) lies on the critical path to informing action on biodiversity conservation.

natural resource planning.

2. Background

Occupancy models rely on a few key assumptions to correct for imperfect detection. They expect that data consist of multiple observations collected at each of a set of sites. The *closure assumption* states that the occupancy status of the species remains constant across the observations at a site. For example, a site could be a section of a state park surveyed along three different transects, and the occupancy model would assume that the entire section (site) was either occupied or unoccupied, even if the observations from the transects vary. Classic occupancy models also assume *no false positives*; that is, observers may miss the species, but positive reports are reliable. These assumptions in tandem allow the probability of detection to be estimated. In our example, if the three observations were $[0, 1, 0]$, it is implied that this site is occupied, and the detection probability is $\frac{1}{3}$. This is just the intuition—in the statistical model, many sites and repeated observations are aggregated and the probabilities are linked to covariates (habitat features for occupancy probabilities and survey features for detection probabilities). Scientists planning to use occupancy models design surveys with repeated observations over which closure can reasonably be assumed, but community scientists do not have this structure imposed upon their contributions in advance.

As a case study, we focus on the eBird community science program. eBird, which has over one billion birding observations across the globe, provides researchers with the data necessary to construct avian species models at vast spatial and temporal scales (Sullivan et al., 2014). eBirders report *checklists* of how many of each species they observed, and they indicate whether they are reporting everything they saw, in which case absences can be inferred for all other species.

3. Problem Statement and Related Work

We introduce the **Site Clustering Problem** for grouping a set of geospatial wildlife surveys into sites for occupancy modeling. Our objective is to create sites that satisfy the closure assumption for occupancy modeling. This is challenging because closure does not have a mathematical definition amenable to direct optimization, and because it will vary across species and regions. We expect that successful solutions to this problem will: 1) discover the optimal number of clusters automatically, 2) respect geospatial constraints imposed by species behavior 3) consider similarity in environmental feature space, and 4) run efficiently on large datasets.

The initial state of our case study begins with each eBird checklist as an independent birding observation that has associated spatial coordinates and environmental features. In existing literature, there are two main approaches to preparing community science data for occupancy modeling. Johnston et al. (2019) suggested best practices for analyzing eBird data, including defining a site as a set of two to ten checklists submitted by the same observer at the same exact latitude-longitude coordinates. While this definition is conducive to assuming closure, it reduces the number of visits per site, which occupancy models need for identifying occupancy and detection parameters. Consider the scenario in which an observer conducts consecutive birding surveys at two adjacent, nearby locations in the same habitat. Since the coordinates are not identical, these two checklists would be considered independent observations of two distinct sites instead of leveraging them as replicate observations. Furthermore, this method discards all sites with only a single visit, which can reliably estimate site occupancy (Lele et al., 2012). In our preliminary study, this definition of a site retains **less than 25% of available checklists**. A second approach places a grid over the study region. All surveys within the same grid are considered repeated observations of the same site. A common choice of resolution is 1 km^2 (Dennis et al., 2017; vanStrien et al., 2013). Because this approach ignores the spatial and environmental information, checklists in two different habitats can be grouped into the same site, which is likely to violate the closure assumption.

A variety of existing spatial clustering algorithms (reviewed further by Liu et al. (2012)) might be applied to the Site Clustering Problem, but none clearly meet all the success criteria laid out above. Partitioning methods, such as spatial k-means, and CLARANS (Ng & Han, 2002), require a user-defined number of clusters. Density-based approaches, such as DBSCAN (Ester et al., 1996), and regionalization algorithms, such as SKATER (Assunção et al., 2006) and

	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

Table 1. Several off-the-shelf algorithms and simple heuristics outperform existing practices (red) in terms of similarity (ARI, AMI, NID) and parameter estimation (occ MSE). Higher values of ARI and AMI and lower values of NID and occ MSE indicate better performance.

REDCAP (Guo, 2008), are guided by the density of points, but closure is independent of density (i.e., clusters may be very close in space).

4. Preliminary Experiments

We simulated a case study inspired by 2,146 eBird checklists from the southwestern quadrant of Oregon, USA collected between May and July 2017. We used simulated data in order to examine the impact of different site clusterings with access to ground truth. To create it, we constructed ecologically defensible sites by manually examining the environmental and spatial proximity of eBird checklists. Then we simulated species data at these sites from an occupancy model, which produced present/absent values for each site and detection/non-detection values for every checklist.

We examined a range of spatial clustering algorithms and baselines and compared the following approaches:

- **eBird-BP**: groups checklists with identical coordinates and observers; 2-10 visits per site (Johnston et al., 2019)
- **1-kmSq**: groups checklists falling within the same cell of a grid of 1 km^2 squares
- **lat-long**: groups checklists with identical coordinates, any number of visits per site are allowed
- **rounded-4**: groups checklists with the same coordinates, rounded to the 4^{th} decimal place
- **density-based**: groups checklists by considering each checklist's spatial neighbors and the similarity in environmental feature space (Liu et al., 2012)
- **clustGeo**: a hierarchical agglomerative clustering method (Chavent et al., 2018)

In addition, we implemented two consensus clustering solutions, **CC-balls** and **CC-agglom** (Gionis et al., 2007). Consensus clustering (or ensemble clustering) combines the clusters of multiple clustering algorithms into a single result (Vega-Pons & Ruiz-Shulcloper, 2011). This is a promising method for our domain because a combination of algorithms with different spatial constraints may better approximate a spatial requirement that is not well-defined (closure). In both **CC-balls** and **CC-agglom**, the inputs were **lat-long**, **rounded-4**, and **density-based**.

We considered two strategies for evaluating these algorithms. First, we asked whether the sites returned by the clustering algorithms accurately reflected the ground truth assignments. We measured Adjusted Rand Index (ARI), Adjusted Mutual Information (AMI), and Normalized Information Distance (NID) to assess this (Vinh et al., 2010). These metrics serve as a proxy for a measurement of closure because we can assume closure holds for our ground-truth clustering. Second, we asked whether the site assignments produced differences in the ability of the occupancy model to estimate the ground truth parameters. To this end, we measured the average mean squared error of the predicted and true occupancy probabilities across all checklists.

5. Results and Discussion

For all evaluation metrics, the new methods we applied outperformed the eBird best practice and the common practice of imposing a grid (Table 1). The improvements exceed two standard deviations, with the exception of **density-based**. Even very simple solutions (i.e., **rounded-4** and **lat-long**) substantially improved on existing practices, and the off-the-shelf spatial clustering algorithms showed similar improvements. The similar performance among these methods suggests room for improvement.

Future work on this project will focus on designing a novel spatial clustering algorithm to improve on the methods presented here. A successful algorithm will integrate domain knowledge and we have begun to develop an ecology-informed distance metric that can be inserted into existing clustering algorithms. Eventually, we plan to expand this work to spatial-temporal considerations rather than fixing the temporal period of closure in advance. Our preliminary work suggests that better solutions to the Site Clustering Problem for community science datasets may produce sig-

nificant improvements in SDMs. With more accurate SDMs, conservationists and land managers can identify regions of concern with greater precision and dedicate limited resources to remedy population declines more effectively.

References

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

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

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.

Dennis, E., Morgan, B., Freeman, S., Ridout, M., Brereton, T., Fox, R., Powney, G., and Roy, D. Efficient occupancy model-fitting for extensive citizen-science data. 2017.

Ester, M., Kriegel, H.-P., Sander, J., and Xu, X. A density-based 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.

Johnston, A., Hochachka, W., Strimas-Mackey, M., Gutierrez, V. R., Robinson, O., Miller, E., Auer, T., Kelling, S., and Fink, D. Best practices for making reliable inferences from citizen science data: case study using ebird to estimate species distributions. *bioRxiv*, 2019. doi: 10.1101/574392.

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

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

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.

MacKenzie, D. I., Nichols, J. D., Lachman, G. B., Droege, S., Royle, J. A., and Langtimm, C. A. Estimating Site Occupancy Rates When Detection Probabilities Are Less Than One. *Ecology*, 83(8):2248–2255, aug 2002.

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.

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., Fink, D., Fitzpatrick, J. W., Fredericks, T., Gerbracht, J., Gomes, C., Hochachka, W. M., Iliff, M. J., Lagoze, C., La Sorte, F. A., Merrifield, M., Morris, W., Phillips, T. B., Reynolds, M., Rodewald, A. D., Rosenberg, K. V., Trautmann, N. M., Wiggins, A., Winkler, D. W., Wong, W. K., Wood, C. L., Yu, J., and Kelling, S. The eBird enterprise: An integrated approach to development and application of citizen science. *Biological Conservation*, 169:31–40, 2014.

vanStrien, A., van Swaay, C., and Termaat, T. Opportunistic citizen science data of animal species produce reliable estimates of distribution trends if analysed with occupancy models. *Journal of Applied Ecology*, 50:1450–1458, 2013.

Vega-Pons, S. and Ruiz-Shulcloper, J. A survey of clustering ensemble algorithms. *International Journal of Pattern Recognition and Artificial Intelligence*, 25(03):337–372, 2011.

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.