Clustering Methods and Occupancy Models For Better Detection Estimates and Occupancy Probabilities

AnaPatricia Olvera Medina
Lehman College
anapatricia.medina@lc.cuny.edu

Demetrius Hernandez
The University of Texas at El Paso
dhernandez79@miners.utep.edu

Dr. Rebecca A. Hutchinson
Oregon State University
Rebecca.Hutchinson@oregonstate.edu

Dr. Eugene Seo
Oregon State University
seoe@oregonstate.edu

Abstract

Community science projects are a good resourceful way to access huge amounts of data. Since it has low barriers to contribute, it is voluntary, and crowdsource data from different locations. However, one of the disadvantages that are encountered with community science data is *imperfect detection*; the phenomenon in which observers do not detect all individuals and/or species present during a survey [1]. Occupancy models are used to account for imperfect detection of organisms in surveys and to determine the probability of the true presence or absence of a species at a site [2]. In this paper, different clustering methods are explored to find a method that helps with obtaining better detection and occupancy estimates.

Introduction

Species Distribution Models (SDM) are tools that allow scientists and natural resource managers to make informed decisions regarding real-world policies to mitigate the impact of climate change and aid in wildlife conservation efforts. Occupancy models help to correct for imperfect detection, so they are crucial to creating SDM. Without occupancy models the data being fed into SDMs would be faulty, thus the information given to policymakers will also be flawed, leading to important decisions being made using incorrect data. The success of occupancy models is paramount to produce correct and accurate predictions because we want to avoid making real-world decisions that can negatively impact our environment.

Occupancy models were developed for expert surveys that had pre-defined sites, where experts can define the sites to ensure closure. Closure can be explained as “if a site is occupied during at least one survey, it is assumed to have been occupied during all surveys [3], i.e., no changes in occupancy between surveys. So the new problem is how do we adapt this pre-existing framework to the new community science datasets. In this document, work from this past summer is outlined, specifically regarding the impact of clustering methods on occupancy models’ ability to produce better detection estimates and occupancy probabilities.

Experiment 1

The main question attempted to be answered with experiment 1 is assuming that closure holds, what trade-offs should be considered about the number of sites and number of visits per site. For example, are the parameter estimates better with 100 sites and 3 visits, or 150 sites 2 visits.

Approach

The first step in designing this experiment was data collection. It was decided to use fully simulated data when running this experiment because a simulation enables one to compare actual target processes and not have concerns with the "messiness" of real-world data. With simulated data there is control of the probability coefficients being used, so the conditions of the experiment can be varied and the outcomes can be investigated accordingly. Creating simulated data was a very important step because if the data was not reliable then the results would not be useful.

All code was written in R for this experiment. First, a randomly ground-truth dataset was generated. This ground-truth dataset was used to repeatedly change the number of visits and number of sites by partitioning the data. This experiment starts with a dataset of size M, where M is some constant, and 16 visits per site. Then the number of sites is halved(M/2) and the number of visits doubled. Below we see an example with a ground truth dataset of size M = 200 with 16 visits per site:

200 sites / 2 visits
100 sites / 4 visits
50 sites / 8 visits
25 sites / 16 visits

At each partition, an analysis was run to determine the root mean square error (RMSE) [4] value for both occupancy and detection predictions. RMSE value indicates the error between our predicted model and the values observed. The experiment was conducted with 4 different sizes of M and 10 replications per size.

Results

It was found that to have really good parameter estimates one must ensure a good balance of the number of sites and the number of visits per site. Below is a graph of one of the first test trials (Fig 1).

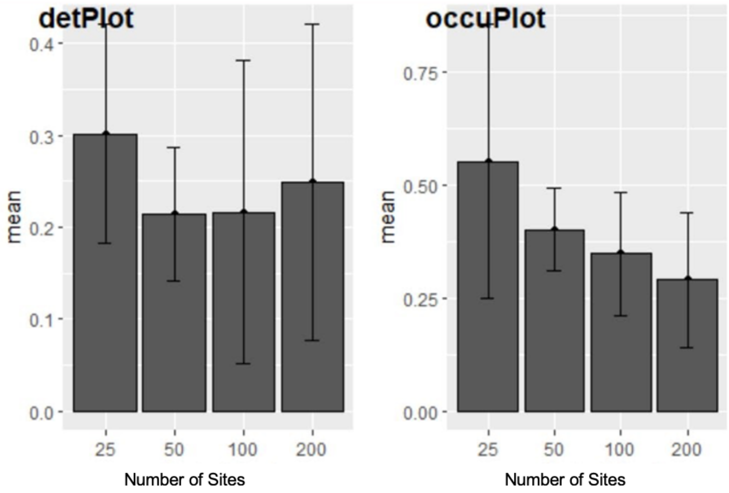


Figure 1. 200 sites and 2 visits showed a lower RMSE value

It was decided to run again with larger data points to find more prominent patterns in the results (Fig 2).

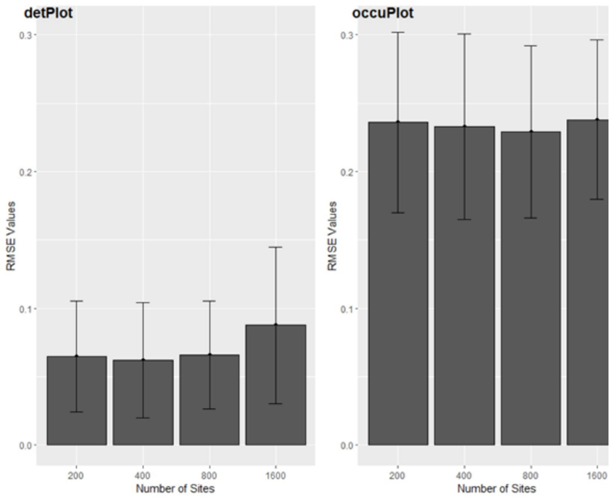


Figure 2. RMSE values for detection lowered ad the number of sites increas

In this experiment, 200 sites was an effective cutoff point, as the results are not significantly better when the number of sites is increased. 200 sites may not always be an effective cutoff, depending on many different variables such as how you slice the data, what is the criteria for success of the experiment, etc. Limitations of this study can be how much data is available for analysis. In this experiment, there are infinite data points because of the use of simulated data and not real-world data. The result of this experiment may not hold true for all similar trials, but the results are important to emphasize the significance of finding an effective cutoff point between the number of sites and the number of visits.

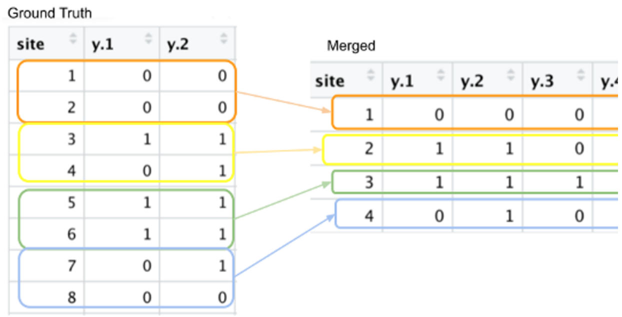
Experiment 2

For this second experiment, the assumption of closure would be broken on purpose by merging presence and absence sites into one. The purpose of this was to explore the question, how bad estimates could be when merging sites? Is there an instance in which it is possible to get a lower RMSE?

Approach

The procedure followed to merge sites was the following: first, a ground-truth dataset was created with two covariates, elevation and temperature. In the first scenario, the ground-truth dataset was 200 sites and 2 visits sized. The merged dataset had to have half the number of sites and double the number of visits (100 sites and 4 visits). To obtain this outcome sites were merged in the following way; in Figure 3, the left side is a snapshot of the ground-truth dataset, sites 1 and 2 merge into site 1 on the right side, sites 3 and 4 merge into site 2, and so on. By looking at sites 1 and 2 of the ground-truth data, it is noticeable that the detection for all visits is zero, however, the occupancy value for each site might be different. Therefore merging both sites into one violates the closure assumption. The date variables are also merged in the same way observations are.

To calculate the values of the covariates for the merged data the average of the sites on the ground-truth data is taken. For example, in Figure 4, the average of sites 1 and 2 for elevation in the ground-truth dataset becomes the elevation value of the merged data, the same happens with temperature. For the second scenario, the ground-truth dataset is 100 sites and 4 visits sized.



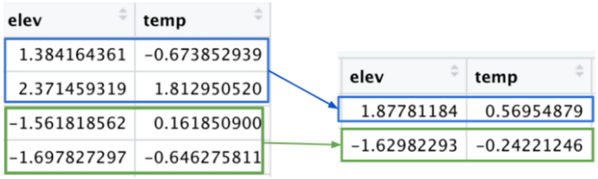
Ground Truth

site	y.1	y.2
1	0	0
2	0	0
3	1	1
4	0	1
5	1	1
6	1	1
7	0	1
8	0	0

Merged

site	y.1	y.2	y.3	y.
1	0	0	0	
2	1	1	0	
3	1	1	1	
4	0	1	0	

Figure 3. The number of sites was half by merging sites together



elev	temp
1.384164361	-0.673852939
2.371459319	1.812950520
-1.561818562	0.161850900
-1.697827297	-0.646275811

elev	temp
1.87781184	0.56954879
-1.62982293	-0.24221246

Figure 4. Covariates were average to obtain a new covariate value for the merged

After having two datasets, the ground-truth and merged, they were converted into CSV files, occupancy models were fitted into both datasets, and the root square means error of occupancy and detection were calculated. This process was repeated 10 times and all RMSE values were store into matrices for visualizations.

Results

In both scenarios, 200 sites and 2 visits (Fig 5) and 100 sites and 4 visits (Fig 6), the detection parameter estimates appeared to be more affected with the merged data compared with the occupancy parameter estimates. This is due to the closure assumption being broken when merging sites together. As previously said, two sites might have the same detection values, but different occupancy statuses (i.e., one is occupied and the other is not occupied by a species). When merging the sites the new detection value leads to an incorrect estimation, as well as occupancy.

The 200 sites and 2 visits model showed to have lower RMSE than the 100 sites and 4 visits multiple instances with the ground-truth dataset. In previous research by Lele et al. *(*2012*)*, repeated visiting to estimate detection errors may not be productive all the time. Repeated visiting can be expensive since it reduces the number of sites that can be visited within "the same sampling season for the same cost," and can affect the environment in which the species is located [5]. This may justify why the 200 sites and 2 visits model performed much better than 100 visits and 4 visits. However, this statement may not be true since the experiment used generated data.

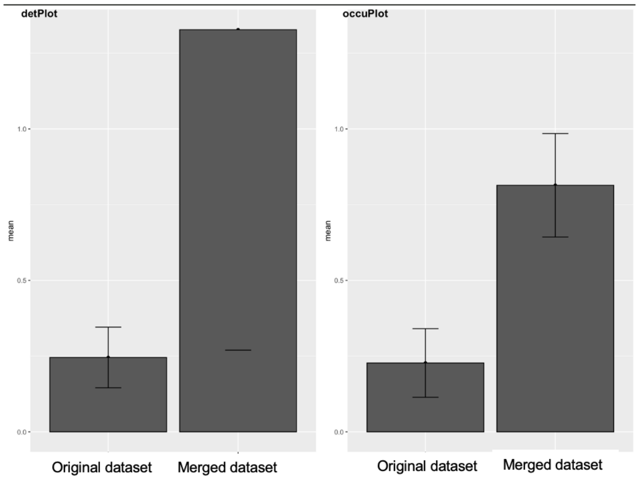


Figure 5. 200 sites and 2 visits results

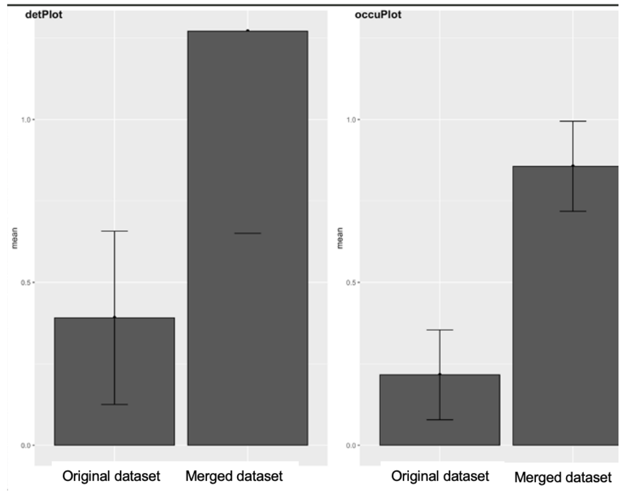


Figure 6. 100 sites and 4 visits results

eBird Data Predictions

After both experiments were concluded, eBird data was used to make predictions and visualize them. It is important to notice that with the previous experiments, the simulated data was created based on set beta-coefficients which represented the correct estimated values. These beta-coefficients were then later compared to the results after fitting the occupancy model to see which trade-offs between the number of sites and the number of visits were better. Trying to repeat this process with eBird data would not be possible since the data does not have any beta coefficients.

Approach

First, the Western Tanager [6] data was merged with 2017_UPDATED_COVS_df [7] data into a single data frame utilizing checklist_id as a common factor. Also, the data frame was filtered so only 2017 checklists could be used. With this, some visualization was done to understand the data (Fig.7, 8, 9). Sites were created based on latitude and longitude, so if two checklists had the same latitude and longitude then their site ID would be the same. To make the predictions, the data was split 20:80 to be trained. The data frame had 41 variables, which caused some issues when trying to use the *cvsToUMF()* function [8]. It was decided to only keep the following variables:

- Site ID
- Detection values(y.1, y.2)
- Simulated covariates(TCA_mean_75, TCB_mean_75, TCW_std_150, TCW_std_1200, aspect_mean_1200)
- Date values (date.1, date.2)

After dropping all the variables that were not needed, and being able to fit the occupancy model into the data, the *predict()* function [9] was used to make predictions shown in Fig 10.

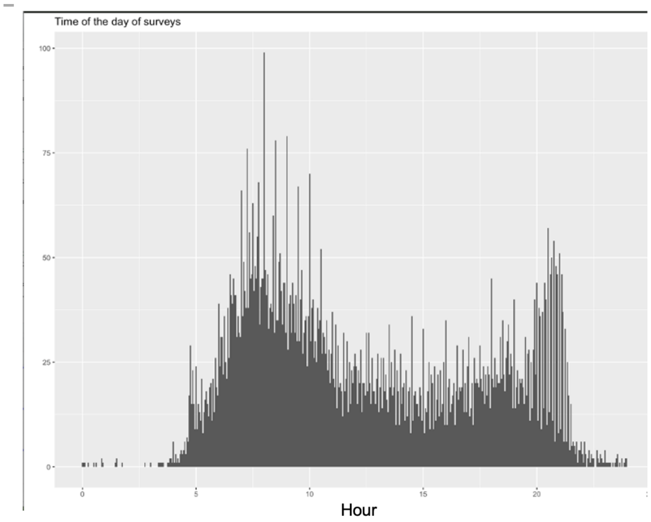


Figure 7. Most surveys took place between 5AM-10AM

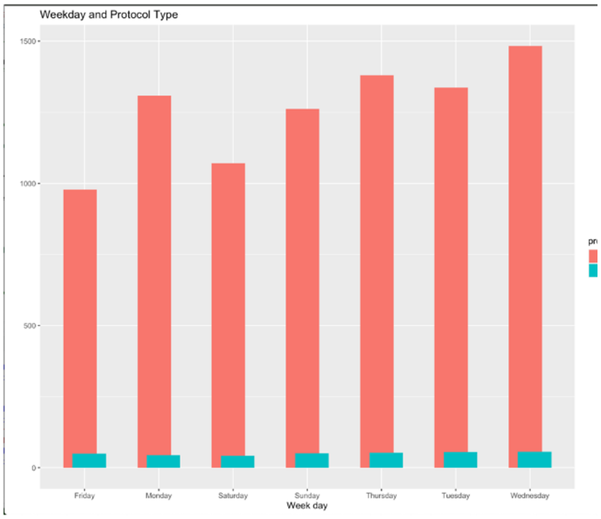


Figure 8. Most surveys were taken on Wednesday and were stationary

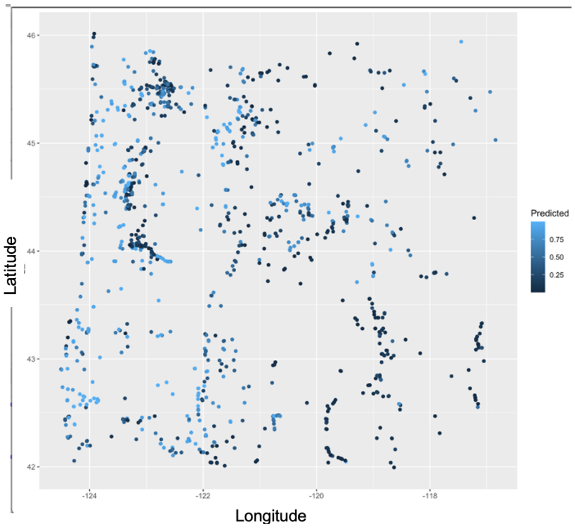


Figure 10. Predicted values showed low occupancy in cities like Portland

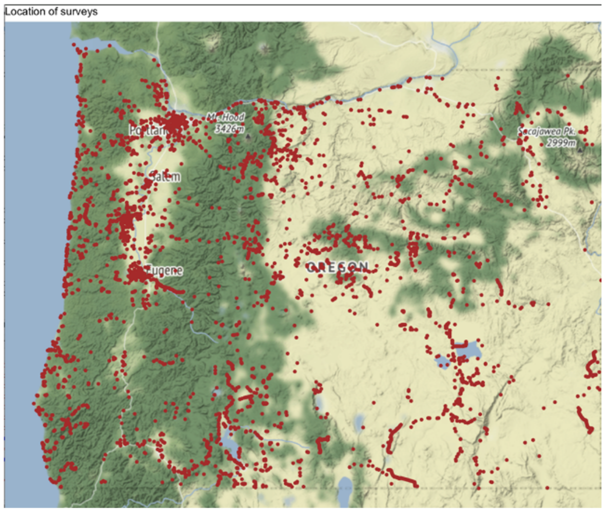


Figure 9. Location of surveys in the State of Oregon

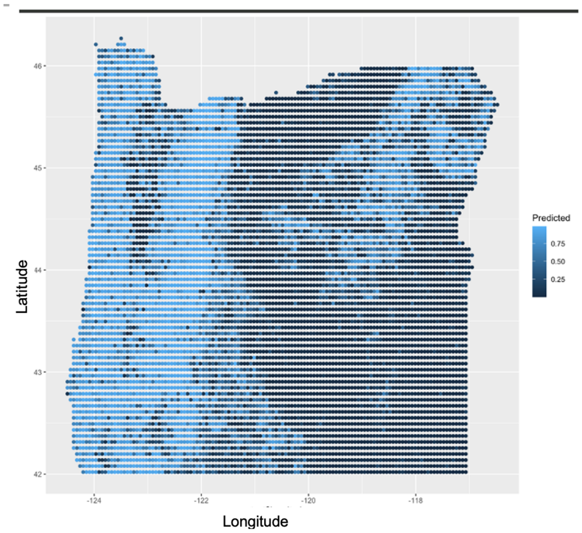


Figure 11. Low occupancy in the South-East side of Oregon

Results

The predictions show that it is more likely to see species present in the Southwest part of Oregon. The size of the data used to make this prediction had only 9,170 checklists. Going through the same process, the data for all Oregon areas was used to make a second prediction map and see differences if any. As shown in Fig, 11, the Southwest remains the best place for occupancy.

Future work

It would be interesting to explore different questions such as: assuming closure holds and the number of sites and visits stays the same, what trade-offs should be considered between the number of covariates? Is it more beneficial to have more or fewer covariates? While closure holds, is there an instance in which the number of sites, visits, and covariates approximate lower occupancy and detection values?

Conclusion

Experiment 1 showed the importance of having a good balance between the number of sites and visits because too many of either will be problematic or too costly. Also having too many repeated visits to the environment may scare certain species away, so fewer visits are preferred so that these species are more likely to be detected.

Experiment 2 taught the importance of being cautious while merging sites. Merging sites can cause huge issues because, for example, one site can have a different occupancy probability than another, merging the two sites creates an incorrect occupancy estimate for the newly merged site.

Overall, with community science growing in size, quality, and the importance it must ensure that researchers continue to improve. As the data changes, it is important to continue to adapt existing frameworks to the new community science datasets to provide conservationists and land/resource managers with accurate information that will help them make informed decisions.

References

[1] Roth M., Hallman T., Robinson W.D., Hutchinson R.A. (2021) On the Role of Spatial Clustering Algorithms in Building Species Distribution Models from Community Science Data.

[2] Nick. (n.d.). Occupancy models. Retrieved September 20, 2021, from https://fukamilab.github.io/BIO202/09-C-occupancy-models.html.

[3] Rota, C. T., Fletcher Jr, R. J., Dorazio, R. M., & Betts, M. G. (2009). Occupancy estimation and the closure assumption. *Journal of Applied Ecology*. https://doi.org/10.1111/j.1365-2664.2009.01734.x

[4]*Rmse: Root mean square error*. Statistics How To. (2021, May 31). Retrieved September 14, 2021, from https://www.statisticshowto.com/probability-and-statistics/regression-analysis/rmse-root-mean-square-error/.

[5] Lele, S. R., Moreno, M., & 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. https://doi.org/10.1093/jpe/rtr042

[6]Western Tanager. (n.d.). https://ebird.org/news/western-tanager.

[7] Mark Roth, pers. comm.

[8] *Csvtoumf: Convert .csv file to an unmarkedframe*. RDocumentation. (n.d.). https://www.rdocumentation.org/packages/unmarked/versions/1.1.1/topics/csvToUMF.

[9] *Predict: Model predictions*. RDocumentation. (n.d.). https://www.rdocumentation.org/packages/car/versions/3.0-11/topics/Predict.