Discovering Hotspots and Coldspots of Species Richness in eBird Data

Travis Moore and **Weng-Keen Wong**
School of EECS
Oregon State University
{moortrav,wong}@eecs.oregonstate.edu

Abstract

Quantifying biodiversity is an important task related to ecological research. One way to measure biodiversity is through species richness, which measures the number of unique species found in an area. Recently, citizen science biodiversity datasets such as eBird allow the calculation of species richness over an unprecedented spatial and temporal extent. However, several confounding factors associated with the unstructured observation process, such as observer effort, affect the number of species reported by citizen scientists. In this work, we develop an algorithm for discovering hotspots and coldspots of species richness using eBird data while accounting for these confounding factors.

Introduction

Quantifying the biodiversity of a region is a critical aspect of many important ecological research problems, including designing reserves for conservation, building models of species extinction and measuring the effects of climate change. One way to measure biodiversity is to use species richness, which is the count of unique species found in a region. Since it is difficult to detect all species within a given area, ecologists have traditionally estimated species richness from species accumulation curves (Gotelli and Colwell 2001). These curves model the rate of change in the accumulation of new species and use the asymptote of this curve to estimate the total number of species. Collecting data to produce species accumulation curves is labor intensive and is typically performed by trained scientists. Due to the manual labor involved, the data are usually collected over small regions, thereby limiting the conclusions that can be drawn from these studies.

Recently, the citizen science paradigm has opened new doors for ecological research. Citizen science projects for biodiversity, such as eBird (Sullivan et al. 2014) and eButterfly (Larrivee et al. 2014), encourage participants to submit their species observations and thus create a global human sensor network. The eBird project, which is the context for our work, is currently one of the largest active citizen science projects. Participants in eBird submit checklists of bird species observed during searches, along with information describing the amount of effort they expended (eg. distance traveled and time spent birding). These checklists can be used to estimate species richness over broad spatial and temporal extents that were not previously possible. However, this data is collected by citizen scientists instead of trained scientists. Although critics bring up concerns over data quality (eg. due to observer variability in identifying birds by species), recent studies have shown citizen science data to be informative (Munson et al. 2010) and machine learning approaches can help separate the signal from the noise (Yu, Wong, and Hutchinson 2010).

In our work, we present an algorithm for discovering hotspots or coldspots for species richness based on data reported by eBird participants. Our work is related to approaches in spatial statistics for discovering spatial hotspots such as the GI* statistic (Getis and Ord 1992) and the Spatial Scan Statistic (Kulldorff 1997). However, because the number of unique species reported on a checklist is influenced by numerous factors (eg. the duration of time spent birding), a novel aspect of our approach is spatial hotspot/coldspot discovery based on citizen science observations taking into account the confounding factors associated with the unstructured observation process. We evaluate the algorithm using simulated data designed to resemble eBird data.

Methodology

Negative Binomial Regression

Our detection algorithm represents the ith eBird checklist as a tuple $(\boldsymbol{x_i}, y_i)$, where y_i is the count of unique species observed and $\boldsymbol{x_i}$ are the covariates affecting y_i. First, we model the distribution of y_i as a negative binomial distribution. We use the negative binomial because it acts as an overdispersed Poisson distribution, modeling count data without having the variance tied directly to the mean (Hilbe 2007). Capturing this overdispersion is important for eBird data, which is very noisy.

The PDF of the negative binomial distribution can be parameterized as

$$P(Y=k) = \frac{\Gamma(k+1/\alpha)}{\Gamma(k+1)\Gamma(1/\alpha)} \left(\frac{\alpha\mu}{\alpha\mu+1}\right)^k \left(\frac{1}{\alpha\mu+1}\right)^{\frac{1}{\alpha}} \quad (1)$$

This gives the log likelihood of the data as

$$l(y, \mu, \alpha) = \sum_{i=1}^{n} \ln(\Gamma(y_i + 1/\alpha)) - \ln(\Gamma(y_i + 1))$$

$$- \ln(\Gamma(1/\alpha)) + y_i \ln\left(\frac{\alpha\mu_i}{\alpha\mu_i + 1}\right) - \frac{1}{\alpha}\ln(\alpha\mu_i + 1) \quad (2)$$

The parameter μ_i is subscripted here because it will be replaced by a regression on the covariates $\boldsymbol{x_i}$. We use the canonical log link for negative binomial regression, allowing us to rewrite $\mu_i = \frac{1}{\alpha(\exp(-\boldsymbol{x_i}\cdot\boldsymbol{\beta})-1)}$. As (Hilbe 2007) shows, making this substitution gives us the likelihood

$$l(y, \boldsymbol{\beta}, \alpha) = \sum_{i=1}^{n} y_i(\boldsymbol{x_i}\cdot\boldsymbol{\beta}) + \frac{1}{\alpha}\ln(1 - \exp(\boldsymbol{x_i}\cdot\boldsymbol{\beta})) +$$

$$\ln(\Gamma(y_i + 1/\alpha)) - \ln(\Gamma(y_i + 1)) - \ln(\Gamma(1/\alpha)) \quad (3)$$

In this work, $\boldsymbol{x_i}$ consists of a single covariate, namely the time spent observing, and an intercept term. As such, the regression on $\boldsymbol{x_i}$ accounts for the trend that observers report a greater number of species when observing for a longer period of time. To simplify notation, we will refer to the time spent observing on the ith checklist as the single covariate x_i.

Grid Search

In order to deal with the spatial aspects of this problem, our algorithm partitions the geographic area into a set of fixed-size grid cells (see Figure 1). For each grid cell, we consider two sets of checklists: 1) the checklists within a grid cell and 2) the checklists from the grid cell's eight immediate neighbors. We fit a negative binomial regression to each set of checklists. This regression has the number of species as the response variable and the observation time as the covariate. We end up with two models, one for the number of species within the grid cell and the other for the neighbors of the grid cell. In order to compare the two models, we perform a hypothesis test on the two regression models:

$$H_0 : \mu_i = \beta_0 + \beta_1 x_i$$
$$H_1 : \mu_i = \beta_0 + \beta_1 x_i + \beta_2[I(i)] + \beta_3[I(i)x_i]$$
$$s.t.\ \beta_2 \neq 0 \text{ or } \beta_3 \neq 0 \quad (4)$$

In the equations above, $I(i)$ is an indicator function that is 1 if the ith checklist comes from within the grid cell, and 0 if it comes from one of the cell's neighbors. Under the null hypothesis, the regression model for the grid cell and its neighbors are identical. Under the alternative hypothesis, the checklists inside the grid cell may have an different conditional mean relative to its neighbors.

We use the likelihood ratio test to determine if the null hypothesis is to be rejected. Using Wilks theorem (Wilks 1938) the test statistic $D = 2l(H_a) - 2l(H_0)$ follows a chi-squared distribution with two degrees of freedom, since there are two extra parameters to fit in the alternative model.

Since this test is performed on each grid cell, we must account for the multiple hypothesis test problem. Even with a set error level of $\alpha = 0.05$ for each test, the probability of a false positive increases as the number of tests increases. We correct for this by using the False Discovery Rate (FDR) (Benjamini and Hochberg 1995). To explain the FDR correction, we define the number of *discoveries* to be the number of times we reject the null hypothesis. Furthermore, we define a *false discovery* to occur when we reject the null hypothesis when it is in fact true. If we use $\alpha = 0.05$, then the FDR ensures that the expected number of false discoveries divided by the number of discoveries will be 0.05. Using FDR in conjunction with our algorithm allows us to reduce the probability of falsely reporting a grid cell to be different from its neighbors.

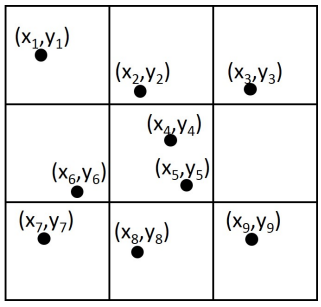

Figure 1: An example of a grid used in our algorithm. Each checklist (represented as a dot) falls into a single grid cell, and consists of the observation duration (x_i) and the number of species seen (y_i). Under the alternative hypothesis, checklists inside the center grid cell $\{(x_4, y_4), (x_5, y_5)\}$ have a different conditional mean from the negative binomial regression than those in the outer 8 grid cells.

Synthetic Data Generation

For the purpose of establishing ground truth, we generate simulated data to evaluate our approach. The simulated data models the number of species that occupy the area of observation as a Gaussian process (Rasmussen 2006). We used a Gaussian process with a constant mean and periodic kernel to create a smooth distribution space with many local "bumps" (see Figure 2(a)). Simulated hotspots are then injected by selecting a rectangular area and uniformly boosting the counts in this area by a constant (see Figure 2(b)). We chose rectangular hotspots to better match the assumptions of the grid search, though in the future we hope to move away from any assumptions on the shape of hotspots. Although our algorithm is equally capable of detecting hotspots and coldspots, in this paper we only evaluate its ability to detect injected hotspots.

To generate the checklists, we randomly sample locations and observation durations. The durations are modeled as a beta distribution with a mean at 60 minutes, as this was the empirical distribution we found on actual eBird data. Each duration x_i is transformed into a proportion p_i using the equation $p_i = Pr(Z \leq x_i)$, where Z follows an exponential distribution. This maps the duration values to a value between 0 and 1 along the exponential CDF curve. These proportions are then multiplied by the true occupancy at the

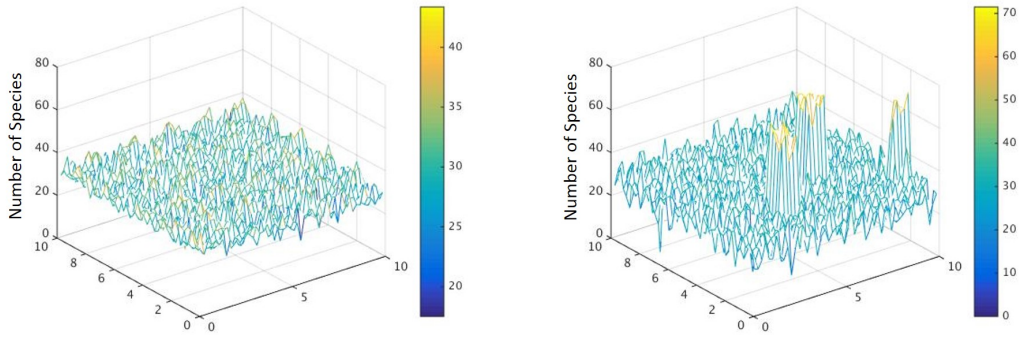

Figure 2: An example of synthetic data without injected hotspots (left) and with injected hotspots (right).

location to get the observed number of species for the checklist. Under this procedure, checklists with longer durations capture a greater percentage of the true number of species. This downsampling makes the injected peaks harder to find, since the algorithm must account for the *detected* number of species instead of the actual number.

We tested our algorithm on 30 different randomly generated synthetic datasets of 20,000 checklists each. Each dataset had 1 to 5 injected hotspots of varying height and width. The data was partitioned into square grids, with either 10, 20, 25, or 40 grids to a side. If a grid cell was found to be a hotspot, the algorithm labels all the checklists within the grid cell as coming from a hotspot. We compared this labeling against the ground truth from the simulation, and reported the following metrics: *true positive rate* (the percent of checklists correctly identified to be coming from hotspots), *false positive rate* (the percent of non-hotspot checklists incorrectly identified as coming from hotspots) and *false discovery rate* (the percent of checklists identified to be from hotspots that are *not* actually from hotspots).

Results and Discussion

As a baseline, we tested our algorithm with negative binomial regression, which accounted for the observation duration, against the algorithm with just a negative binomial. We abbreviate the detector with regression as NBR, and the detector without regression as NB. In the non-regression version, the model treats the detected number of species as the actual number of species and we still use likelihood ratio to test if each grid has an elevated mean with respect to its neighbors. For the baseline, this test statistic is a chi-squared distribution with one degree of freedom. We also report the algorithm results with the FDR correction, with $\alpha = 0.05$. These variants are abbreviated as NBR-FDR and NB-FDR.

Tables 1 and 2 show the true and false positive rates for the algorithms, averaged over the 30 datasets. Bolded values are significantly better then their counterparts (NBR vs NB, and NBR-FDR vs NB-FDR) based on a paired t-test with $\alpha = 0.05$. Table 3 shows the false discovery rate for the algorithms, and how it changes when using the FDR correction.

True Positive Rate

Grid Size	NBR	NB	NBR-FDR	NB-FDR
10x10	**0.9997**	0.9958	0.4733	0.5038
20x20	**0.9912***	0.9781	0.7382*	0.609
25x25	**0.6277***	0.4918	0.1262*	0.0497
40x40	**0.3271***	0.1796	0	0

Table 1: True positive rates of per-checklist hotspot detection for each algorithm. Results averaged over 30 randomized experiments. Starred values are significantly better from their non-regression counterparts (paired t-test, α=0.05)

False Positive Rate

Grid Size	NBR	NB	NBR-FDR	NB-FDR
10x10	0.0562	0.0403*	**0.006**	0.0062
20x20	0.0413	0.0184*	0.0007	**0.0003**
25x25	0.0464	0.0177*	0.0002	**0**
40x40	0.0501	0.0087*	**0**	**0**

Table 2: False positive rates of per-checklist hotspot detection for each algorithm. Results averaged over 30 randomized experiments. Starred values are significantly better from their regression counterparts (paired t-test, α=0.05)

In 3 of the cases NBR has a significantly higher true positive rate than NB, while in all cases NB has a significantly lower false positive rate. These results indicate that NB is much more conservative in finding hotspots, which is to be expected, since without the additional input of observation time and the majority of observations being at relatively low durations, NB will be inclined to underestimate the population mean. The addition of the FDR step makes both algorithms more conservative, lowering the false positive rate but also substantially reducing the true positive rate relative to its non-FDR counterparts. However it does make the reported hotspots more reliable, as a smaller percentage of them are false after the correction. In future work, we will look deeper into finding the right tradeoff between the true and false positive rates.

False Discovery Rate

Grid Size	NBR	NB	NBR-FDR	NB-FDR
10x10	0.7811	0.7198	0.4459	**0.4386**
20x20	0.7257	0.5443	0.0568	**0.0303**
25x25	0.8243	0.6955	0.0914	**0**
40x40	0.9067	0.7546	**0**	**0**

Table 3: False discovery rates of per-checklist hotspot detection for each algorithm. Results averaged over 30 randomized experiments. False discovery rate is the proportion of reported hotspots that are false.

Algorithm performance peaks at the 20 by 20 grid split, and then degrades as the grid cells get finer and finer. When the grid cells are much smaller than the injected peaks, cells deep within the injected peak appear to be similar to their neighbors, and hence not reported as a local hotspot. Data sparsity may also be a contributing factor, as the smaller grid cells will have less data points to fit the regression model. The algorithm is clearly sensitive to the choice of grid size, and it would be beneficial in future versions to automate this choice in some intelligent way.

Conclusion and Future Work

We investigated an algorithm to detect biodiversity hotspots from citizen science data. Our results on simulated data indicate that the negative binomial regression model, which models how the number of species detected changes with time spent observing, helps improve the true positive rate. The addition of the False Discovery Rate to correct for multiple hypothesis testing helps reduce false positives, but does so at the expense of a substantially reduced true positive rate. In addition, the size of the grid cells clearly has an effect on the performance of the model.

In future work, we first plan to discover spatial regions rather than being restricted to considering a set of predefined grid cells. We will employ a similar strategy to the spatial scan statistic, which searches over spatial regions such as circles (Kulldorff 1997). This search will increase the computational complexity but will improve the flexibility of the approach. Second, we would like to explore a one-sided likelihood ratio test to specifically detect species-rich hotspots, but these one-sided tests are more complex and computationally expensive. Third, another issue with our approach is that it uses a regression estimate of the conditional mean of the distribution as a conservative estimate of species richness. We are actually interested in the maximum number of species in an area (for hotspots) instead of the mean. As an alternative, we will explore estimating the maximum number of species in an area, but this estimate is well-known to lack robustness as it is very sensitive to noise.

Finally, in this work, we are only using the total count of species on each checklist. In reality, checklists from eBird are more detailed as they list the actual species observed. We can leverage these details to develop new hotspot detection algorithms based on biodiversity metrics other than species richness, such as entropy (Shannon 1948) and β diversity (Whittaker 1972), which measures the change in species composition.

Acknowledgements

This work was supported by the National Science Foundation under Grant No. 1209714.

References

Benjamini, Y., and Hochberg, Y. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B* 57:289300.

Getis, A., and Ord, J. K. 1992. The analysis of spatial association by use of distance statistics. *Geographical Analysis* 24:189–206.

Gotelli, N., and Colwell, R. 2001. Quantifying biodiversity: procedures and pitfalls in the measurement and comparison of species richness. *Ecology Letters* 4:379–391.

Hilbe, J. M. 2007. Negative binomial regression. In *Negative Binomial Regression*. Cambridge University Press. 77–98. Cambridge Books Online.

Kulldorff, M. 1997. A spatial scan statistic. *Communications in Statistics: Theory and Methods* 26(6):1481–1496.

Larrivee, M.; Prudic, K. L.; McFarland, K.; and Kerr, J. 2014. ebutterfly: a citizen-based butterfly database in the biological science s. http://www.e-butterfly.org.

Munson, M. A.; Caruana, R.; Fink, D.; Hochachka, W. M.; Iliff, M.; Rosenberg, K. V.; Sheldon, D.; Sullivan, B. L.; Wood, C.; and Kelling, S. 2010. A method for measuring the relative information content of data from different monitoring protocols. *Methods in Ecology and Evolution* 1(3):263–273.

Rasmussen, C. E. 2006. *Gaussian processes for machine learning*. MIT Press.

Shannon, C. E. 1948. A mathematical theory of communication. *The Bell System Technical Journal* 27:379423, 623656.

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. 2014. The ebird enterprise: An integrated approach to development and application of citizen science. *Biological Conservation* 169:31–40.

Whittaker, R. H. 1972. Evolution and measurement of species diversity. *Taxon* 21:213–251.

Wilks, S. S. 1938. The large-sample distribution of the likelihood ratio for testing composite hypotheses. *The Annals of Mathematical Statistics* 9:60–62.

Yu, J.; Wong, W.-K.; and Hutchinson, R. 2010. Modeling experts and novices in citizen science data for species distr ibution modeling. In *Proceedings of the 2010 IEEE International Conference on Data Mini ng*, 1157–1162.