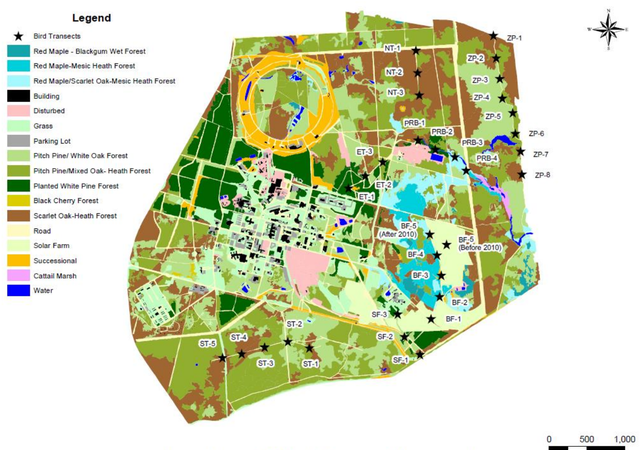
Building a zero-inflated abundance model for avian populations at BNL

Brett Keeler, Stony Brook University, Stony Brook, NY 11790

Mentor, Tim Green, Environmental Protection Division, Brookhaven National Laboratory, Upton, NY 11973



Location of Bird Survey Transects

Figure 1. Locations of the various transects and sites at BNL.

Abstract

As an approach, Bayesian modeling has provided many advantages for ecologists when compared to classical statistics: it is easier to fit smaller data sets, makes it simpler to propagate error, and allows for the use of prior information in order to make better informed determinations. Though the approach is conceptually and computationally difficult, relatively new programs such as WinBUGS and JAGS have allowed Bayesian modeling to become much more widespread in use, especially with packages that combine them with R. The Environmental Protection Division has maintained an observational data set for 19 years for all avian species at marked spots in BNL. We set out to create an abundance model to determine the total population of birds at BNL, starting with the American Crow, *Corvus brachyhrynchos*, while accounting for their detection probability and other confounding variables. Though the model has yet to be completed, model building is an iterative process and progress is being made. Instead, we present information about the process of building the model as well as some preliminary thoughts about relationships that we are investigating.

Figure 2. An American Crow.
https:/ / www.thoughtco.com/ thmb/ 2D5LXr6vKv9Axc UKpcBv4jpFv8g=/ 768x0/ filters:no_upscale():max_bytes (150000):strip_icc()/ GettyImages-571342903-5a495f2d482c52003608a6d3.jpg

Introduction

Bayesian Modeling

BUGS, Bayesian Inference Using Gibbs Sampling, is often the language of choice for ecologists who use Bayesian modeling. JAGS, a companion program, is used to compile BUGS and must be used alongside R. This kind of modeling comes in two main flavors- occupancy modelling, which focuses on whether or not the species was present at all, and abundance modelling, which accounts for the number of each species present. Since we have the proper data for it, we are building an abundance model.

Our current model accounts for two states, the first being the observation state, which is what we can actually see. This is just a reflection of the latent state, which is the true state of the system. While our data set perfectly covers the observation state, determining the latent state is the true goal of the model. These states are modeled via distributions, such as the Poisson and Bernoulli, and simulated using Monte Carlo (MCMC) chains with thousands of iterations in order to find convergence. However, because of the complexities of the model, including the choice of distributions, initial and prior values, and structure, fitting them is often difficult and time consuming (2). Finally, repeated visits to the same site are required for valid results.

Why create the model?

While abundance is one of the more interesting measurements in biology, full census surveys are difficult and expensive to run.

Occupancy modeling relies on simple measurements, either just presence/ absence or count-based data, that are easy to collect and can even be gleaned from historical records. Many projects use data that is volunteer-sourced as well. This data is much cheaper, less researcher intensive, and still viable to work with.

Outcomes of occupancy models include, but are not quite limited to (3):

Informing species monitoring programs, including how species are distributed and what habitat they prefer.

Determining meta-population dynamics.

Determining species-species interactions.

Determining the species richness of a community.

The strength of the technique has caused the field to become very popular in recent years!

Detection Probability & Closure

One particularly important factor to account for is how likely it is to see a species provided that it is in the environment. We can model this through the use of covariates – factors that change along with the observations – that we think effect how likely it is to see or hear a species. Since our model is single-species, they can be species-specific. Our factors currently include:

Habitat Preference
Crows tend to prefer forests as opposed to wetlands.

Julian Day
Many species are louder in the spring than the early fall.

Wind Speed
Many species fly less often in windy conditions.

Other potential factors could include the observer, species interactions, or disturbance.

Additionally, we must assume the population is closed. We do this by restricting our data to the non-migratory months (May through August). We also use the entirety of BNL as our sample size, though specific methods are valid as well.

Why study the crow?

The American Crow (four letter code AMCR, as it is referred to as throughout the poster) is a good starting point because:
- It is fairly common, allowing for many observations.
- It is well known, allowing for easier interpretation.
- It has been documented that many crows were lost to the West Nile Virus; allowing for historical reference points in our observations (1). This provides a built in "check" to the model.

Priors & Initial Values

Bayesian models need to be told where to start, which can be a powerful tool when used properly. Priors, determined *a priori*, can inform the model of a range that you think the final value is likely to fall. For example, we know there can't be a negative number of crows at BNL, so we can restrict it to positive values to yield the broadest yet informative prior.

Initial values are where the model starts its MCMC simulation - restricting them to values that are both near the final value and within the realm of possibility are vital to the success of the model.

Methods & Materials

Data was collected using trained observers to reduce misidentification. Each site was visited multiple times per year, in a 150 meter radius around a site, shown in Figure 1. Data was recorded using Microsoft Excel ®.

The model was built using R 3.5.2 through R Studio and JAGS version 4.3.0. with the package R2jags. Plots shown were made using R.

Extensive data manipulation was undertaken in order for the data to be usable. Additionally, the model was also built to account for the high number of null values in our data set. Distributions and other facets of the model are still being determined.

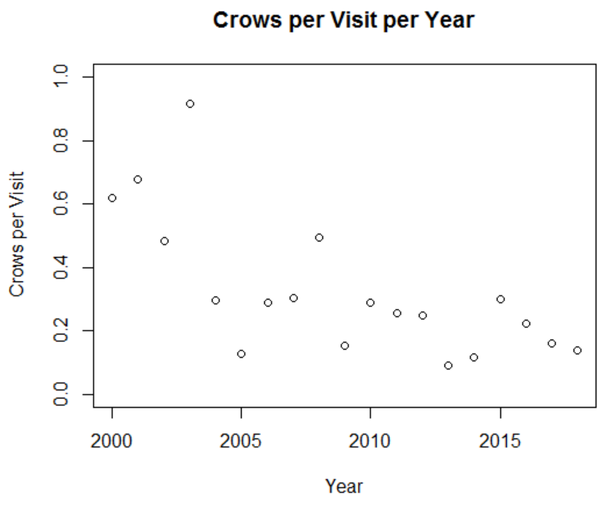


Figure 3. Since each year had a variable number of visits, a rate based metric was used to determine how many crows were seen on average per visit. This should correlate with the estimated population of the model in the end. It is noteworthy that the population seems to be declining, and that 2003 was likely a higher year due to one "40" data point, likely due to a flyover.

Initial Results and Discussion

Even though our model has been under progress for many weeks, it is still a few iterations away from completion. Unfortunately, those iterations aren't interesting enough to report. Instead, we have some preliminary work that can inform two places our model failing: the priors and the initial values.

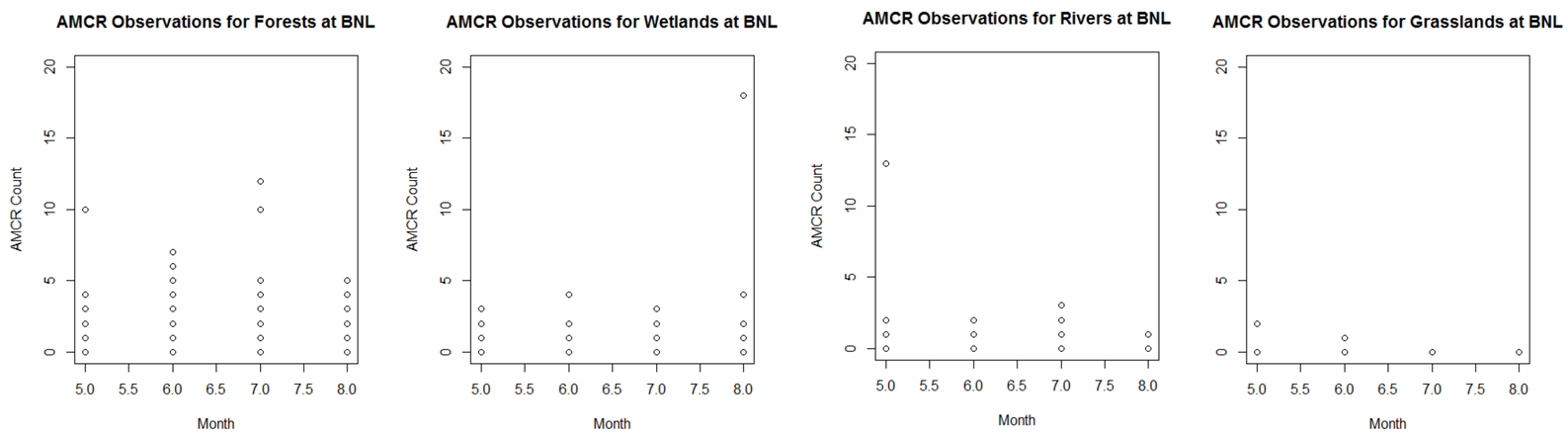


Figure 4. Habitat is likely the most important covariate for occupancy for any species of bird. Forest dwelling birds like the American Crow should be more likely to be found in the forest, and less likely in other areas. These scatterplots show the high counts for each month over the course of the entire data set, leaving out many null data points, but giving a good idea of the preference of the species overall. While one should expect there to be more crows in the grasslands, there are very few sites that can be classified as that – just the solar farm – so that part should be taken with caution.

Table 1. These correlation values for our two measured covariates. These values are likely so low due to the large proportion of null values in the data set, though it does suggest that there is not a particularly impact and should be modeled as such.

Correlation Values for Detection Covariates		
Wind	*r = -.0151*	R^2 = .0004
Julian Date	*r= -.0314*	R^2 = .0103

Figure 5. https://i.pinimg.com/originals/8e/1e/66/8e1e66b0713bc96b9dddb1336f397460.jpg

Future Implications

The values that we are suggesting impact detection are wind speed and the date of the year. Both shouldn't effect whether a bird actually nests in an area, just whether we observe them. Both of their respective impacts seem to lower the detection rate as they get larger, but the results suggest that habitat has a more profound impact overall.. We are assuming that each year has the same detection probability, and that each covariate has the same impact at each habitat.

The American Crow isn't the only species that we are interested in. There are over 119 species in our data set, and around half of them likely have enough data to be modeled, including waterfowl, warblers, and woodpeckers. Birds can be a useful indicator of habitat quality, pollution levels, and other disturbances. The ability to model each individual species at the lab will provide great insight into these factors.

Further developments in the model could focus on being able to distinguish specific sites in the lab, trading a larger sample size for a smaller snapshot of a specific site. If this site is near a disturbance zone in the lab, such as the Solar Farm that was built in the middle of a transect, it could yield meaningful management results as well.

Citations

1. Yaremych, S.A., et al. West Nile Virus and High Death Rate in American Crows. 2004. Emerg. Infect. Dis., Apr; 10(4), 709-711.
2. Welsh, A.H. Fitting and Interpreting Occupancy Models. 2013. PLOS ONE 8(4).
3. Kery, M and Royle, J.A.. Applied hierarchical modeling in Ecology. Volume 1. 2017. Assosciated Press.

Acknowledgements

I would like to thank Tim Green for taking me onboard for the internship and letting me run free on one of the most difficult yet exciting tasks I've attempted.
I would also like to thank Michael Schrimpf, whose assistance and insight was vital to the project and will still be as we finish the model.

This project was supported in part by the U.S. Department of Energy, Office of Science, Office of Workforce Development for Teachers and Scientists (WDTS) under the Science Undergraduate Laboratory Internships Program (SULI).