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

Brett Keeler

Undergraduate Biology, Stony Brook University, Stony Brook, NY 11790

Tim Green

Environmental Protection Division, Brookhaven National Laboratory, Upton, NY 11793

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. This model can have significant management benefits for the lab, as it can be used by anyone to determine long-term trends in the population in the future when more data is collected. Future refinements can also make it site-specific- allowing BNL to know what future construction or disturbance can do to the environment. Through this work, I've learned the value of scientific collaboration through working with graduate students at Stony Brook University, as my work would not have been possible without them. I've also learned great skills that will apply to my future in biology regardless of what path I take in the future, and I now know what it will takes to succeed at the next level of my career.

Introduction

Ecologists have two main questions when studying a species. First, they want to know how many of a species there are, and second, they want to know where those species are located. These two

terms can be defined as abundance and distribution, and they are essential in understanding life on earth.

Census studies, though, are difficult for humans, with one only taking place in the United States every ten years. While plants are sessile, animals are always on the move, and neither group is particularly responsive to mailed letters about themselves anyway. This leads to census studies of animals to be very expensive and very time consuming – two things that ecologists want to avoid.

However, a relatively new technique has been developed that can take advantage of smaller data sets to produce reliable results: occupancy modelling. Occupancy modelling uses Bayesian statistics in order to turn presence-absence data, which can be as small as just a few visits to a site per year, into reliable, accurate estimates of the abundance of a population over the sites wherein the population was measured. My internship focused on developing an abundance model (which uses counts of data, while occupancy models use strictly "presences" and "absences") to assess the avian populations at BNL, of which we have 19 years of high quality data for. This paper will describe the models in general, of which learning about was a significant amount of time in my internship, how the model fits with our data, and what steps we have undertaken to build our own. It will then talk about our future steps we need to take in order to finish the model.

Overview of Bayesian Modeling

Bayesian modeling became prominent in ecology around the turn of the century, allowing ecologists to use strictly detection/non-detection data to successfully model the occupancy of a site[2]. Over time, this model would be developed for even greater results- species-species interactions, abundance, and metapopulational ideas would all come in future years[1]. Ecologists rely on the programming language BUGS, Bayesian Inference Using Gibbs Sampling, and one of two compilers, either WinBUGS or JAGS, typically running those compilers through the programming language R. BUGS

is amazing in that it is simple to write and understand, allowing ecologists with little experience in programming to learn it and use it for their studies. BUGS is used in other fields besides ecology, as well, though we will not discuss its impact there. For this model, we used JAGS, though it is relatively simple to switch between JAGS and WinBUGS.

The model uses Monte Carlo Chains (MCMC sampling) to function. The model takes inputs through the forms of parameters – either an exact value or a range of values, which can come in the form of a statistical distribution. There are two main parameters that most models try to observe- the abundance of the species in question, and the detection probability of the species in question. The abundance model is typically modeled through a discrete random distribution – like the Poisson – as there will always be a whole number of organisms present. The detection can be modeled in a few different ways, but it must be constrained through 0 and 1 – it is a probability – so the binomial or Bernoulli distributions typically work the best for it. While the mathematical workings of the model are important, they aren't vital for understanding the model as an outside viewer, and arguably even as someone building and using the model, too.

Though this kind of modelling is becoming more popular, it is still not quite in the mainstream of biology yet. I hadn't heard of it in my undergraduate curriculum, and thus spent much of my time reading papers, one very good textbook[2], and working with graduate students at Stony Brook University in order to learn and develop my skills.

Our Built Model

The data for the model must be structured in a way that makes sense for what you want to know. We were interested in learning about the population changes in the avian populations at BNL over time – after all, we had 19 years of data to look at. This data was collected by the same few trained surveyors, meaning that the quality of measurements was high. Each survey was done on one of seven

different transects – each with a varying number of sites – and birds were counted about 150 meters away from the central point of the transect. For the model, we created a matrix with the rows consisting of years and the columns consisting of the species – over 120 of them – and either had the count of the birds seen or a zero filled in. This lost some resolution in the model, as the sites were lumped together, but it allows for an easy way to see long term trends. We also constricted our data to the months between May and August to assume the population is closed.

While we have over 120 species to model, we don't have enough data for many of them. Some species have simply been recorded once and then never seen again, while some are common enough to appear at every transect each yeah. About half of the species have enough data to model, and probably about half of those are interesting. Since our model is species-specific, we had to start with just one, so we chose the American Crow. It has a few advantages – crows are common, easy to identify, and already have a known history of population decline with the West Nile disease[3]. This known decline helps us to have a check on our model – once we get the results, we know that this population should decline some time in the mid 2000's, allowing for us to have a reliable check with something other than simulated data to ensure that our model is working properly. It is important to note, though, that changing the species in the model should be straightforward- the actual model won't change, just the data that is being inputted.

We are modeling three covariates in the model right now. These covariates can affect either the detection parameter or the occupancy parameter. The day of year we have affecting the detection parameter, as we think that it is more likely for the birds to sing and be heard in the spring than the late summer. The wind speed also impacts the detection parameter, as birds are less likely to fly in higher winds. We have the habitat sorted into categories, either forested or wetlands, and that affects our

abundance parameter. Birds will only be present in habitats they prefer, so having it affect that instead of detection makes sense.

Additionally, our model is zero-inflated. This means that we have a higher number of non-detections than what we would typically expect from the environment. A few straightforward additions to the model allow an additional parameter to let JAGS have an easier time giving us an accurate model.

There hasn't been much discussion about the mathematics of the model. That is because our model is incomplete, and the structure of the model itself is not set in stone yet. Currently we are having trouble initializing our model: though we would expect the model to be able to start at any number in some random range, we can only get it to run at a very limited range right now. Even then, the model does not converge and yield accurate results. We think that either the initial values or the prior values are the problem. I have investigated the covariates that we are modelling – the wind and day seem to have a low correlation with the abundance of the crow, though the habitat seems to have a significant one. A good next step for the model can be to run it without those covariates and see where we go from there.

Future Implications & Conclusion

While the model isn't done, we are planning to have it completed by June. The model will be able to tell a lot about the history of the avian populations of BNL, leading to valuable insights about how climate change, population fluctuations, and major disturbances (such as the building of the Solar Farm) affected some of our wildlife on the lab. The model is valuable enough that it should be fit for publication as well.

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 Education, Brookhaven National Laboratory, Office of Science, Office of Workforce Development for Teachers and Scientists (WDTS) under the Science Undergraduate Laboratory Internships Program (SULI).lllf

References

Kery, Marc and Royle, Andy. 2017. Applied hierarchical modeling in Ecology. Volume 1. Associated Press, 978-0128013786.

MacKenzie, D.I., Nichols, J.D., Hines, J.E., Knutson, M.G., Franklin, A.B., 2003. Estimating site occupancy, colonization and local extinction when a species is detected imperfectly. Ecology 84, 2200 2207.

Yaremych, S.A., et al. West Nile Virus and High Death Rate in American Crows. 2004. Emerg. Infect. Dis., Apr; 10(4), 709-711.