A data-intensive assessment of the species-abundance distribution.

Presented by Elita Baldridge in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Ecology at Utah State University, 21 April 2015.

Begin transcript:

Slide 1: *Usage notes: Feel free to redistribute this talk, etc. with proper attribution.*

Slide 2: *Title slide*

Slide 3: Let's begin with an example. Here we have a simulated beetle community, in which different colors represent different species of beetle. As you can see, in this community, a majority of species are rare, and then we have a few species that are very abundant. We call this general pattern...

Slide 4: ... the species abundance distribution. The species abundance distribution characterizes the distribution of commonness and rarity in ecological communities, and this pattern of a few species being abundant and the majority being rare is observed not only in this simulated community, but for many different taxonomic groups in many different ecosystems.

Slide 5: Because this pattern is so general, and we do not typically observe other possible forms of the distribution (for example, all species being equally abundant), but instead observe this characteristic hollow curve distribution, it is thought that the shape of the species abundance distribution can reveal mechanisms that generate the observed pattern of commonness and rarity in ecological communities.

Slide 6: Because this pattern is so general, and is thought to be able to potentially provide insight into pattern-generating mechanisms, many different processes have been proposed to generate the species abundance distribution. An example of one process that has recieved a great deal of attention in recent years is neutral theory. Neutral theory is based on the assumption that all species are ecologically and demographically equivalent, and that stochastic variation in birth, death, and immigration processes results in the observed species abundance distribution.

Slide 7: However, there are many other mechanisms that have been proposed to generate the species abundance distribution, and, as yet, no clear concensus as to which process best describes species abundance.

Slide 8: Macroecology is one approach for identifying general ecological patterns and processes like the species distribution. It is characterized by being extremely data-intensive, and macroecologists tend to use large ecological datasets for many different ecosystems and many different taxonomic groups.

Slide 9: The traditional approach to ecology has been to focus on a single site, or a few sites, for a single taxonomic group. This has allowed to learn a great deal about the pattern generating mechanisms at that site; however, it is difficult to generalize these results to other systems and other taxa.

Slide 10: The macroecological approach has tended to use large ecological datasets that cover a broad geographic area, but still have tended to focus on a single taxonomic group.

Slide 11: This brings us to the idea of signal and noise. It's thought that applying more data to a problem can provide a clearer picture (or signal) of what's going on in that particular system.

Slide 12: I'm using a very data-intensive approach to understand the species abundance distribution. First, I'll talk a little bit about the currently available large macroecological datasets suitable for testing species abundance, and then I'll talk a bit about how I leveraged existing ecological data in the literature to fill in some of the gaps in the macroecological datasets and compile the largest dataset for testing species abundance that has been currently assembled. Next, I'll talk about the species abundance distribution, and two different approaches for exploring questions about species abundance. The first of

these approaches is a statistical approach, where I seek to identify the model of species abundance that provides the best characterization of the species abundance distribution for my empirical data. Next, I explore mechanisms that could potentially generate the species abundance distribution by determining whether a neutral or non-neutral model of species abundance best describes the empirical species abundance distribution for the dataset I compiled.

Slide 13: Here is a map of some of the currently available large macroecological datasets suitable for testing species abundance. Out of six datasets, there are four taxonomic groups represented. For birds, we have the North American Breeding Bird Survey, a publicly available dataset collected by citizen scientists along transects over North America in summer. Next, we have the Christmas Bird Count, which is unfortunately not publicly available, but can be obtained through a memorandum of understanding with the Audubon Society and is collected by citizen scientists as a winter bird census. For trees, we have the Forest Inventory Analysis, which is a governmental effort to census trees in the United States, which is publicly available, and then also publicly available, the Alwyn Gentry Forest Tree Transects, which are global in scope. We have one invertebrate taxon, the butterflies, which is collected by citizen scientists for the North American Butterfly Association, and can be obtained through a memorandum of understanding with that organization. The last dataset is the Mammal Community Database, and this is a compilation from the literature that was put together by current and former members of the Weecology lab. As you can see, there is a strong North American bias, and out of six datasets, there are only four taxonomic groups, all terrestrial.

Slide 14: One challenge that macroecology has faced has been a lack of identification of pattern-generating mechanisms. One best practice recommendation to address this challenge is to test with many different taxonomic groups across many different ecosystems. I decided to address some of the gaps in the available macroecological data by compiling a database for underrepresented taxonomic groups from the literature.

Slide 15: This brings us to the first rule of ecoinformatics, which is the same as in programming: garbage in, garbage out. All data are good; they were undoubtedly useful for what they were originally intended for, but not all data are appropriate to all questions. It's extremely important to fit the data to the question and not the question to the data. Using data that are not appropriate for the question can generate a false signal, or prevent a researcher from being able to detect any signal from pattern or process.

Slide 16: It's extremely important, when building a database, to record all decisions that are made at all steps. No database or dataset is complete without metadata. Metadata are not only important if you want to share your data with other researchers; they are also important for allowing you to reuse your database six months down the road, two years down the road; twenty years down the road, when you want to revisit that dataset that you collected when you were a PhD student. Without good metadata, it will be extremely difficult, if not impossible to remember everything you need to know about the dataset to use it appropriately.

Slide 17: I used the following criteria to determine which paper to include in my database. While abundance can be measured by percent cover, or biomass, the currently available large macroecological datasets use count data, and I wanted my database to be able to be combined with the currently existing data. Because I'm interested in species abundance distributions, I needed data for complete ecological communities, not just for the most common or the most rare species. I wanted the raw data, not a summarized form of the data, and because I'm interested in pattern-generating mechanisms of species abundance, I wanted observational data. Using experimental data could create a signal generated by the experimental treatment that would not be appropriate for this study.

Slide 18: I collected the following variables: taxonomic information, abundance data, of course, information about the starting and ending year of collection, site information, including any notes that were important to understand the data for a particular site and the biogeographic region the site occurred in, and, of course, a citation.

Slide 19: So, here we have a table from one of the papers I collected for my database; actually the first

paper I entered. This table is good for publication, but the data can't be used directly from the paper.

Slide 20: Instead, the data from the table have to be hand-entered from the paper into each of three separate tables: a citation table, a sites table, and the main abundance table.

Slide 21: After all the data entry was complete, I saved the tables as plain text; I used a comma delimited format. Saving the data in a non-proprietary file format allows for maximum reusability and direct import into a variety of software and prevents the data from being locked in a format that may become obsolete. At this point, I also wrote up the metadata, because no database, or dataset, is complete without metadata.

Slide 22: Once I was finished building the database, I had data that focused on five taxonomic groups, fish, spiders, beetles, reptiles, and amphibians.

Slide 23: Unfortunately, there is still a strong North American bias, but I have data for all major biogeographic regions except Antarctica.

Slide 24: In addition, I was able to address some of the terrestrial bias with data for fish. I also had a good amount of data for reptiles and some data for amphibians, spiders and beetles.

Slide 25: I've made the data publicly available and open access through figshare, and it's also directly importable through the EcoData Retriever. For the R users out there, I've also provided a snippet of code to import my database directly into R with the EcoData Retriever.

Slide 26: So, I collected data for five taxonomic groups, fish, spiders, amphibians, beetles, and reptiles...

Slide 27: ...and to that, I added data for birds, trees, butterflies, and mammals.

Slide 28: All together, I have data for over 16,000 different communities, for nine different taxonomic groups, over all biogeographic regions except Antarctica. The majority of this data is publicly available, except for the North American Butterfly Association and Christmas Bird Count data, which were obtained through memorandums of understanding with their organizations.

Slide 29: The species abundance distribution characterizes the distribution of commonness and rarity in ecological communities, with the majority of species being rare, and a few species being very abundant. Because this pattern is so general, and it's thought that the shape of the distribution can reveal insight into pattern-generating mechanisms, there have been many different models proposed to explain the species abundance distribution.

Slide 30: These models fall into two major categories, statistical descriptions, which only seek to describe the shape of the species abundance distribution, without inferring mechanism, and process-based models, which seek to identify the major pattern-generating mechanism for species abundance distributions.

Slide 31: Most comparisons of species abundance distribution models to date have used a small subset of the available models, have tended to focus on a single taxonomic group or ecosystem, and have failed to use the most appropriate or current statistical approaches for species abundance distribution analysis.

Slide 32: So, using data for over 16,000 communities for nine taxonomic groups...

Slide 33: ... I tested five species abundance distribution models from four classes. I tested two purely statistical models, the logseries and the Poisson lognormal. For process-based models, I tested the Zipf distribution, a branching process model, the negative binomial distribution, a population dynamics model, and a niche partitioning model, the geometric series.

Slide 34: I did all my model fitting with maximum likelihood estimation, considered to be the current best practice for fitting species abundance distributions. Maximum likelihood estimation provides an estimate for parameters that provide the most likely characterization of the data for a given model.

Slide 35: I then used maximum likelihood based model selection to provide an measure of absolute fit of

the model to the data. This does not take into account the number of parameters, and just describes how well that model fits the data.

Slide 36: To compare how the models fit relative to the other candidate models, I used AICc weights, which describe how well the model fits the data relative to the number of parameters and the other models. I used AICc, instead of AIC, because AICc is more robust to small sample sizes, which was a factor for some of our communities. AICc weights range from zero to one, and the best fitting model from the set of candidate models had the AICc weight closest to one.

Slide 37: All maximum likelihood and AICc calculations were performed with the macroecotools Python package. Xiao Xiao was invaluable in making sure that everything was implemented correctly in macroecotools. All of the code to fully reproduce the analyses, plus a complete description of workflow is available on GitHub. In addition, the database that I compiled is publicly available and directly importable with the EcoData Retriever, as are the majority of the other datasets used.

Slide 38: After we finished running all the data through the analyses, we came up with the following results. The best fitting model for all datasets combined, as determined with AICc weight, was the logseries, and the worst fitting model was the Zipf distribution.

Slide 39: If we look at each dataset individually, we see that the logseries tends to perform well for all datasets, although the performance of the other models varies.

Slide 40: AICc weights range from zero to one, with zero being a poor fit, and one being the best fit. If we look at the frequency distribution of AICc weights for the Zipf distribution, we can see that it does have extremely poor fit overall.

Slide 41: If we look at the frequency distributions for all models combined, we can see that we get some separation between models, indicating that we may be able to determine which model of species abundance best characterizes the species abundance distribution. However, AICc weights only provide a relative measure of model fit. If we look at the frequency distribution for the likelihoods...

Slide 42: ... we can see that the frequency distributions for the likelihoods almost completely overlap, indicating that all models may provide equivalently good fits to the data.

Slide 43: We can examine this further by doing a one-to-one plot of the likelihoods. On the x-axis, we have the logseries, our best fitting model with AICc weights, and on the y, the Zipf distribution, our worst fitting model. However, all the points fall along the one-to-one line, indicating that both the logseries, the best fitting model, and the Zipf distribution, the worst fitting model, have equivalently good absolute fits to the data.

Slide 44: When we look at each model of species abundance relative to the logseries, we see that all points tend to fall along the one to one line, indicating that all models are fitting the data equivalently well.

Slide 45: From this, we can conclude that the existing models provide equivalently good fits to the empirical data. Models with fewer parameters perform better in AICc based model selection, and the logseries provides a good naive model for fitting species abundance distribution. The logseries has only a single parameter, is easy to fit to empirical data, and was the best fitting model overall.

Slide 46: Even though we used the largest dataset ever compiled to test species abundance distribution, it is still challenging to identify process. One possible approach to identify pattern-generating mechanisms might be to examine the scale dependence of a pattern, or subtle differences in how the model fits that data. Macroecology seeks to identify general ecological patterns and processes; it's possible that a more general approach to process, such as neutral vs. non-neutral processes could be a more productive approach.

Slide 47: For neutral processes, there have been many different formulations of neutral theory since its inception. However, they all share the following assumptions: species and individuals are ecologically and demographically equivalent, and stochastic variation in birth, death, and immigration processes

results in the observed variation in species abundance that generates the species abundance distribution.

Slide 48: Some examples of non-neutral processes include competition, niche differentiation, dispersal differences, and so forth. There are many different types of non-neutral processes, but they all suggest that differences in species abundance are due to differences among species, whereas neutral theory emphasizes species equivalence.

Slide 49: Early tests of neutral theory compared the fit of empirical species abundance distributions to the neutral prediction, but later tests suggested that species abundance comparisons alone were insufficient for a rigorous test of neutrality. However...

Slide 50: ...Connolly et al. 2014 simulated neutral communities through several different processes, and were able to identify a single of neutrality. On the x-axis is the median number of distinct abundance values in a community, which provides a measure of the ability of the community to detect a signal of neutrality. On the y axis, values closer to zero indicate that the community is best described by a neutral model, and values closer to one indicate that the community is best described by a non-neutral model. While the ability to detect a signal of neutrality increases with the number of distinct abundance values, the number of distinct abundance values isn't very high before a strong signal of neutrality can be detected.

Slide 51: Besides simulated communities, Connolly et al. went a step further and tested empirical marine communities, and were able to identify a strong signal of non-neutrality in these empirical marine systems, indicating that this may be a robust approach to detecting a signal of non-neutrality. However, this approach has not yet been tested in terrestrial systems.

Slide 52: To test this approach in terrestrial systems, we used the same data as with the previous species abundance model comparisons, and the same maximum likelihood model fitting approach.

Slide 53: Following Connolly et al 2014, we used the Poisson lognormal as our non-neutral model, and the negative binomial as the neutral model. The Poisson lognormal is a classic model of species abundance distributions, which makes it a good choice for the non-neutral model. While there are many different neutral models, all of them share the negative binomial distribution as the local community prediction.

Slide 54: Here we have a histogram of species abundance from a randomly selected community for each dataset. In magenta is the predicted form of the non-neutral model, and in cyan is the predicted form of the neutral model. As you can see, it is extremely difficult to distinguish between the two models; however, the results from Connolly et al. 2014 in marine systems suggest that it is possible.

Slide 55: Connolly et al. 2014 were able to clearly identify a signal of non-neutrality in marine systems. If terrestrial results are consistent with the marine results, we would expect the values to be closer to one on the y-axis...

Slide 56: ...however, the results for terrestrial systems are completely different from the Connolly et al. 2014 marine systems. We cannot distinguish between neutral or non-neutral models, looking at the combined datasets.

Slide 57: If we look at each site individually, we can see that some sites are clearly best characterized by a neutral model, some sites are best characterized by a non-neutral model, and some sites are unclear.

Slide 58: This demonstrates the importance of testing in multiple ecosystems. If we'd stopped with the results from Connolly et al. 2014 in marine systems, we would have concluded that neutral processes aren't important in empirical systems. Testing with our varied data revealed that it is very difficult to identify a clear winning model.

Slide 59: Using a data-intensive approach revealed some of the challenges of identifying process from species abundance distributions alone. Even using data for over 16,000 communities, for nine taxonomic groups, and over all the major biogeographic regions except Antarctica, we were still unable to identify a clear winning model or infer an overall mechanism. The results from Connolly et al. 2014

suggest that a broad model categorization may be a more productive approach, but there may still not be one single suite of processes that dominates in all systems.

Slide 60: In addition, there are also challenges in identifying mechanism among datasets. Using a data-intensive approach helps to remove uncertainty about non-biological pattern generating mechanisms. Because we used such a diverse set of data, with different spatial structuring and sampling intensities, we can say with a high degree of confidence that our results are due to biological, rather than non-biological, differences between terrestrial and marine systems. However, even with a great deal of data, it's still very challenging to identify mechanism.

Slide 61: All the code to fully reproduce these analyses and a description of workflow is available on GitHub. In addition, the database that I compiled for this research is publicly available on figshare.

Slide 62: With that, I'd like to thank the various funding sources that have supported me throughout my time at Utah State...

Slide 63: ...the Weecologists past, present, and future that I've gotten to talk science with, especially Xiao Xiao and Ken Locey, excellent collaborators, and fantastic labmates.

Slide 64: Dr. Thomas Price, and the USU Student Health Center, who provided excellent medical care as I went through the process of developing, and eventually being diagnosed with a chronic illness as a graduate student. My very supportive husband and family, who take care of me, so I can take care of my research. And, of course, the publicly available data, and the citizen scientists that make these large macroecological datasets possible.

Slide 65: This dissertation would not have been possible without disability accommodations. Ethan has been very supportive in allowing me to finish up my dissertation remotely, which made it easier to manage my condition and complete this research. Providing an accessibility statement as part of event announcements helps to remove some of the unintentional barriers to participation for chronically ill or disabled scientists. I wrote an accessibility statement generator to provide an educational tool to make it easier for organizers to integrate accessibility into event planning, at the link provided below.

Slide 66: And with that, I'll take any questions.