Computational Methods in Ecology and Evolution:

Miniproject

Using different measures of biodiversity to investigate global drivers of local Avian biodiversity

March 8, 2019

Rachel Bates [1,2]

[1] Imperial College London, Department of Life Sciences

[2] r.bates18@imperial.ac.uk

Word Count: 3129

Abstract

Our planet's biodiversity is declining at an alarming rate. To stop it, we must know exactly where and why it is happening. Being able to predict biodiversity's responses to climate change and land-use change allows us to pinpoint the most impacted areas and predict biodiversity change into the future. If we can determine the state of a community by simply counting species rather than doing more time-intensive sampling, we can assess and protect sites faster. In this study, I investigate a) whether using measures of evenness reveal different predictors than using species richness as a biodiversity measure and b) what factors predict local Avian community biodiversity on a global scale. I show that species richness comes to the same conclusions as both Shannon's and Simpson's measures - that land-use intensity is the single most important predictive variable - without losing any significant information. This suggests that species richness is able to produce meaningful models with less sampling effort. It also shows that land-use intensity is a hugely impactful factor for global Avian biodiversity declines, highlighting the importance of conserving primary and secondary vegetation where possible.

1 Introduction

We are currently losing species at a rate comparable to previous mass extinctions (**Ceballos2015**) (**Ceballos2015**). The 2002 Convention on Biological Diversity contained the target to "to achieve by 2010 a significant reduction of the current rate of biodiversity loss at the global, regional and national level...to the benefit of all life on earth." (**Balmford2005**). In order to achieve this reduction in the rate of biodiversity loss the current rate needs to be known, and methods of tracking it going forward must be produced. Choosing a measure of biodiversity that contains the maximum amount of information about a system while also being easy to measure is vital to facilitate this tracking.

Measures of biodiversity tend to look at one of two aspects of a community: richness (i.e. the number of different species) and evenness of the number of individuals of those species. Richness is obviously much easier to measure. As long as a species is known to be present in an area, it can be recorded. This means that camera trapping, signs of activity and even environmental DNA analysis can be used to work out the species richness of a site (**Li2016**; **Rovero2014**). Calculating evenness requires knowledge of the abundance of each species. This often requires repeated sampling of a site (multiple quadrats, transects, traps etc.) to get an estimate of abundance for each species. However measures of evenness can identify situations where a community is dominated by one or two species.

Other ways to measure biodiversity include measuring genetic, taxonomic or functional diversity instead of species diversity, or relative biodiversity (for example the biodiversity intactness index (**Scholes2005**) which measures relative biodiversity between a human-impacted habitat and a pristine habitat that are otherwise the same). These all require more data, either of the species or the area in which they reside. This study aims to look at more basic measures of biodiversity as they are more widely used and easier to measure.

In this study I will be investigating 3 measures of biodiversity:

Species Richness:

This is simply calculated by counting the number of different species in an area (N). As such, this measure is easy to calculate and requires only presence/absence data for a site. It values each species the same amount, meaning rare and common species have the same weight. As such, species richness values rare species more highly than their relative contribution to the ecosystem. This may overestimate diversity where there is largely mono dominance with a few rare species, but also accounts for the fact that rare species often have important functional roles (**Bunker2013**; **Leitao2016**).

Simpson's Index of Biodiversity:

$$D = \sum \left(\frac{n_i}{N}\right)^2$$

n_i = no. of individuals of one species N = total no. of individuals

Simpson's Index (**SIMPSON1949**) is a measure of similarity and evenness. It is a value between 0 and 1 where values close to 0 indicate less similar (more diverse) communities. This index is highly impacted by dominant species, with rare species making little impact on the final measure (**Dejong1975**). It is also possible to convert Simpson's Index to a continuous measure of diversity with higher numbers indicating a more diverse community by taking the reciprocal of the index ($\frac{1}{D}$), known as the reciprocal Simpson's Index. This also fixes the issue of the model converging close to it's final value with only the first two or three most dominant species.

Shannon's Measure of Diversity:

$$H' = \sum \left(\frac{n_i}{N} \times \ln \frac{n_i}{N} \right)$$

n_i = no. of individuals of one species $\quad N$ = total no. of individuals

Shannon's measure (**Shannon1949**) is also a measure of evenness, but due to the inclusion of a *ln*, it is less affected by highly dominant species than Simpson's Index. The least diverse communities have a Shannon's measure of close to 0, while increasingly diverse communities have increasingly higher H' values. This is because Shannon measures the uncertainty that one individual from a community will be of a specific species.

Birds are a good model species for several reasons. Firstly they are widely studied, being popular and easy to ID. We have lots of bird data, both contemporary and historical, for areas where we may be lacking data for other taxa. For example, birds were the first taxa to be included on the Red List Index due to having sufficient data (**Butchart2004**). Secondly, birds can be used as umbrella species due to being wide ranging species (**Suter2002**), are varied in their level of specialisations and are impacted by changes further down the food chain in flora in insect communities (**BURGHARDT2009**). This means that they can potentially act as indicators for the health of the community as a whole. Birds are undergoing large worldwide declines following the same pattern as global declines (**Pimm2006**).

Several studies have been carried out trying to discern traits that can predict either individual species threat/loss (**AmecayJuarez2014**; **Cardillo2004**; **Purvis2000**) or community biodiversity loss (**DePalma2016**). Using habitat level traits to predict biodiversity change is particularly powerful as these habitat level traits (change in climate, land-use etc.) can be predicted into the future.

This means that where patterns are present those traits can be used to predict biodiversity/species change into the future, which can in turn be used to test future scenarios and inform policy (**Newbold2015**).

This study looks at three aspects of global, abiotic factors that may influence biodiversity. Latitude represents geographical scale effects, accounting for habitat variation and climatic variables, and has been shown to correlate with biodiversity (**Gaston2000**). Diversity is larger in the tropics than outside of them, and so it is important to test for geographic variation. GDP is used as several human level factors have been shown to influence biodiversity including inequality of wealth and political instability (**HANSON2009**; **Mikkelson2007**). Finally land-use intensity is included because habitat loss and fragmentation are considered the leading causes of global biodiversity decline (**Dirzo2004**).

The aim of this study is firstly to determine whether biodiversity measures that include a measure of evenness show different patterns in explanatory variable, and so see if the extra data needed to calculate these measures is worth collecting. Secondly this study aims to determine what, if any, factors correlate with biodiversity in birds across the globe.

2 Methods

All species abundance data were collected from the 2016 public release version of the PREDICTS database (**Hudson2016**). It consists of x species records from y sites over z studies, and contains data on the geography of the records (mainly related to geographic position but also on the habitat of the area) as well as land-use intensity of each site.

Using Python (**VanRossum2016**) and the package pandas (**McKinney2010**) the data were subsetted to only include avian abundance data (i.e. presence/absence data were discarded). I also discarded any sites with fewer than 5 species for the evenness calculations. This resulted in 49742 records from 2456 sites. The species richness, reciprocal Simpson's index and Shannon's measure of diversity were calculated for each site (hereafter referred to as Richness, Simpson and Shannon). The reciprocal Simpson's index was used to enable linear models to be used with all three measures, as Simpson's index is bounded and is thus non-normal. Each site was given a weighting based on geographic realm to enable weighting in the linear models, and each land-use class was given a value. These were:

1. Primary vegetation
2. Secondary vegetation
3. Plantation forest
4. Agricultural land (both pasture and cropland)
5. Urban.

I also calculated the Gross Domestic Product (GDP) of the country in which the site was located, from the year of sampling (where sampling occurred over several years this was the midpoint). The GDP values were obtained from The World Bank Group (**TheWorldBankGroup2018**).

2.1 Modelling

All modelling was carried out in R v3.4.4 (**CoreTeam2019**). The factors that I included within the maximal model were Land-use class, Latitude (absolute values, i.e. measures of distance from the equator) and GDP at year of study. I did not include habitat patch area or years since fragmentation/conversion of the patch although these data were available for 482 and 358 sites respectively, because area data were largely only collected for primary vegetation sites (76%) and only 87 sites have data for both factors. This would sacrifice statistical power if the data were to be reduced to less than 4% of its size. I also included the original study as a random effect, so that variation due to different sampling techniques, recorders, time of year, study focus etc. would not mask other trends.

Both the Simpson and Richness measures were heavily right skewed (Simpson: range = 1.2 to 130.9, median = 9.8 and mean = 12.7; Richness: range = 5 to 474, median = 16, mean = 20.3) so these were log transformed for all further analyses. I checked for normality of the three diversity measures using quantile-quantile plots (via the ggqqplot() function from ggpubr (**Kassambara2018**)), and as the main portion of the points lay across the reference line normality was inferred. I did not carry out shapiro-wilkes tests or other normality tests as with very large sample sizes normality isn't required for parametric tests so a visual test was all I deemed necessary.

Linear mixed models were fitted using the R package lme4 (**Bates2015**). First a maximal model was used to test whether inclusion of study as a random effect had a significant impact on the results, then it was used to test if weighting by geographic realm had a significant impact on the results. This was done by fitting linear (mixed) models to both hypotheses and using an anova and the AIC values of the models to decide if there was a significant difference and which model provided a better fit.

Linear mixed models were then fitted to every combination of explanatory variables for each diversity metric. Akaike information criterion (AIC) values (**Akaike1974**) - a measure of goodness of fit - were calculated for every model, and then converted to relative AIC values based on the smallest AIC. The R package xtable (**Dahl2018**) was used to format the AIC tables for LaTeX.

2.2 Computing Languages

This project used 3 main coding languages to carry out the data manipulation, analyses and to construct a reproducible workflow.

Python: Used for initial data manipulation, calculation of diversity metrics and creating the final subsetted dataframe. Python was used as it can handle large files more efficiently than R (the initial database was 2.7GB) and can apply functions to these large dataframes efficiently by use of the pandas package.

R: Used for fitting the linear mixed models, calculating AIC values and plotting the data. R was chosen as the lme4 package is useful for fitting mixed models, and subsequently the inbuilt AIC() function can be used directly on the mixed model. The package GGPlot2 also makes creating attractive plots very simple.

Bash: Used to glue the workflow together so that the project becomes fully reproducible. Using bash means that the LaTeX file could be compiled with references directly as opposed to using subprocess modules in Python.

3 Results

The final database contained data for 2456 sites from 49 studies, a total of 49742 bird species records.

Including species richness as a random effect accounted for a significant amount of variation (ANOVA, p<0.005) and produces a better fitting model (AIC = 3918 with study and 5831 without using the maximal species richness model). When weighting by geographic realm, the AIC values produced were different enough that the uneven geographical distribution of sites must be producing biases in the results (AIC=3000 for non-weighted model and AIC=3918 for weighted model, again using the maximal species richness model). Geographic weighting was therefore included in all models despite not being significantly different to the unweighted model (ANOVA, p=1).

AIC values were calculated for each possible combination of variables (GDP at year of sampling, Latitude of site and Land-use intensity) for all three measures of diversity (Table 1).

Table 1: Relative AIC values for all variations of the linear mixed models for each diversity metric. All values are relvative to the smallest AIC value (Richness, Latitude + Land-use intensity) for which AIC=3956. Values in brackets are the AIC relative to the smallest for that diversity metric. Values with a * are the best fitting models for that diversity metric.

Δ AIC	Richness	Simpson	Shannon
Null Model	121	687(141)	350(142)
GDP	125	692(146)	355(147)
Land-use	0*	546(0)*	208(0)*
Latitude	129	694(148)	358(150)
GDP + Land-use	5	551(5)	213(5)
GDP + Latitude	135	701(155)	364(156)
Latitude + Land-use	8	553(7)	216(8)
Maximal Model	14	560(14)	222(14)

From these AIC values it can be seen that land-use intensity is the main explanatory variable for biodiversity. Models containing land-use intensity as a variable are the best fitting for all three measures, and for all metrics the best fitting model contains only land-use intensity - although the fit is very similar to both GDP + land-use and latitude + land-use models. We can also see that, in general, species richness fits the data the best followed by Shannon then Simpson. The general trend is of decreasing biodiversity as land-use intensity increases (Figure 1). Models not containing land-use intensity as a variable tend to fit less well for the evenness measures than for species richness (Table 1).

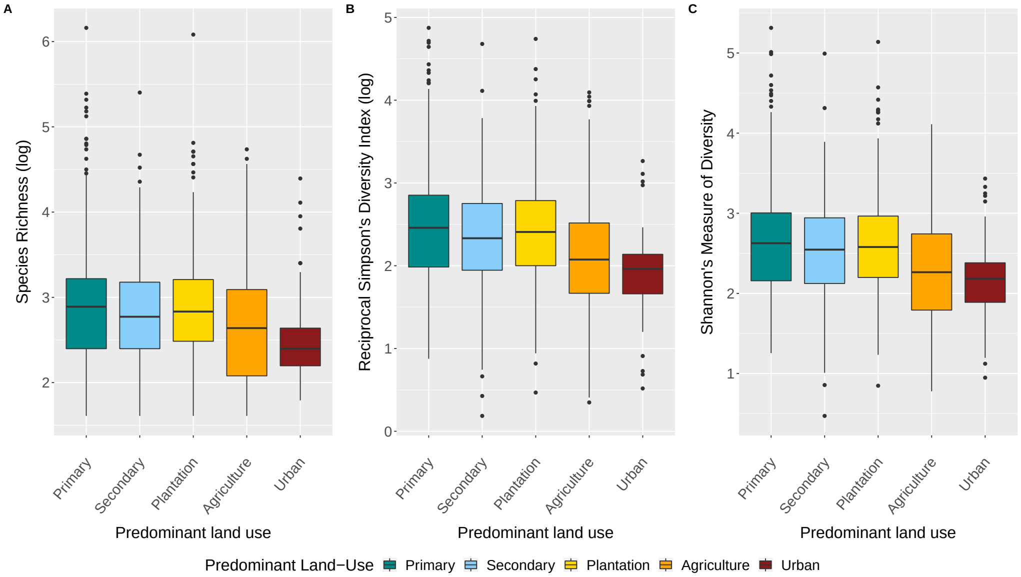

Figure 1: Site biodiversity across different land-use classes measured by A) species richness, B) reciprocal Simpson's Index and C) Shannon's measure of diversity.

4 Discussion

This study aimed to look at whether diversity measures that incorporated a measure of evenness can give us different outcomes from predictive models than species richness alone. It also aimed to discern what variables can predict biodiversity on a global scale.

4.1 Geographic Bias

Geographic bias has been shown to exist in many studies and papers, as most data on biodiversity is collected where scientists live and work - primarily temperate regions such as North America and Western Europe (**Trimble2012**) - although threatend species have recieved more attention in recnet years (**Roberts2016**). It has also been shown that these biases can influence attempts to correlate biodiversity change with other factors as there can be interactive effects (**DePalma2016**) such as Europe with has been largely without primary forest for the last few hundred years in comparison to Central and South America for which any habitat destruction is largely more recent.

I decided to follow a similar strategy to that of the weighted Living Planet Index (**Hudson2016**) and weight each site so that the contribution of each realm to the model was equal. This reduced the goodness of the fit of the model, but mean that predictions made from the model are able to be applied globally.

Future studies could look at modelling each realm individually, which would enable any differences between them to be more easily identified. It would also mean that the predictions for each realm would be more precise. For this do be achieved, more data needs to be acquired on the under-represented realms. For example in these data 1166 sites were located in the Afrotropics, compared to only 95 in Australasia.

4.2 Measures of Diversity

Species richness produced the best fitting models of all three measures. However all three measures showed the same pattern of fit given the same combinations of variables (i.e. models containing land-use intensity produced the best fit). This suggests that there is little benefit to using measures of evenness instead of species richness. Species richness contains less information about the community, but from these models it appears that the predictive variables correlate mainly to species richness rather than species composition. This is interesting, as several studies have shown that dominance of some species increases and specialist species decline with increasing habitat degradation (**Devictor2008**). What this study may be capturing, therefore, is the influx of generalist species and invasive species to impacted areas (**With2004**). In this case, looking more closely at the species identities of a community to produce a measure of biodiversity could produce different or varying strengths of predictive effects. It would also mean that invasive species presence could be identified, a leading cause of Avian biodiversity loss (**Clavero2009**) For example the Biodiversity Intactness Index (BII) looks at species composition in pristine and impacted sites, and doesn't include any species that are only present in the impacted site. Future studies could look into this, and compare species richness to measures that take community composition into account, as well as repeating this study for other taxa to see if the same trends exist.

4.3 Predictive Variables

Land-use change in the form of habitat loss and fragmentation has been identified as the leading cause of biodiversity and species loss both globally and in birds (**Gaston2003**; **Dirzo2004**). This study confirms that finding, models that included land-use intensity more accurately fit the data than those that didn't. The best model fit land-use alone, but the fit was only minorly worse when including GDP, latitude, or both. This shows that including GDP and latitude does not explain much meaningful variation, and so they are not significant predictive variables of local biodiversity.

Three factors that are closely related to land-use intensity are the amount of time since any land-use change, the size of the area and distance to the nearest disturbed or undisturbed habitat. Edge effects, extinction debt and spillover effects are all factors that can be hugely impactful on a community that has been impacted by land-use change (**Banks-Leite2010**; **Ford2009**; **Robinson1995**). These factors were measured for some data in this database, but unfortunately not enough to include in the model. Future studies could focus on obtaining these data, as including them in models could show that land-use is not in fact the most important determining factor. Also expanding this to non-Avian species may also produce different results, birds are a highly mobile taxa and so issues such as fragmentation tend to impact them less than others such as reptiles (**Keinath2017**; **Villard1994**).

Conclusion

This study has shown that both species richness and measures of evenness are useful measures of biodiversity. Species richness provides an easy way to assess a community, and can produce predictions that are more precise than other measures. It may miss some patterns that measures of evenness are able to capture, but the difference does not seem significant enough to warrant the extra effort needed to measure evenness. Additionally using only species richness data allows the use of presence/absence data, and so increases the pool of data available for meta-analyses.

Land-use intensity was shown to be the single most important factor in predicting differences in biodiversity. This is consistent with other studies and the view that habitat loss and fragmentation are the largest threats to this planet's biodiversity at this time. Future studies should focus on determining if there are geographic differences in effect sizes of predictive variables and the predictive variables themselves. It would also be useful to investigate other measures of biodiversity, such as looking more closely as species composition.