Patterns and trends of behaviour in endangered species listing heavily biased towards imperiled vertebrates with high utility value to humans*

A replication analysis using data from United States, with climate change projections in trends to 2100.

Julia Kim

April 2, 2024

Abstract

The Endangered Species Act (ESA), reputed as the United States' strongest environmental law, empowers the federal government to protect listed species from extinction. This paper replicates a statistical analysis of the main determinants of actual government decisions about the ESA listing of endangered species. We apply secondary research to estimate how species listing will evolve over the 21st century due to climate change, regarded as the most serious and persistent threat to biodiversity. The results indicate that higher likelihoods of ESA listing are heavily biased towards vertebrate species and are associated with greater endangerment, greater evolutionary uniqueness, and higher utility value to humans, the latter as measured by Google Ngram frequencies. Unmitigated climate change would warrant the listing of an additional 690 species, with several thousand more becoming critically imperiled but remaining unlisted. These findings emphasise the roles of both scientific and anthropocentric characteristics in the listing decision, and highlight the importance of climate change mitigation efforts in sustainably listing and protecting biodiversity.

Table of contents

*Code and data are available at: https://github.com/julia-ve-kim/US_Climate_Change_Biodiversity; Minimal replication on Social Science Reproduction platform available at https://doi.org/10.48152/ssrp-tpay-ac71.

1 Introduction

Climate change poses a serious threat to biodiversity. As a consequence of economic growth and development, inhibited by concern of the environment, various species of wildlife in the United States, and elsewhere in the world, have been rendered extinct or are so depleted in numbers as to be in danger of extinction (Cornell Law School 2024). These species have integral ecological roles in the environment, and are of important aesthetic, historical, recreational and scientific value to people (Cornell Law School 2024). In response, the American Congress signed the 1973 Endangered Species Act, empowering the federal government to protect endemic species from extinction (U.S. Fish & Wildlife Service 2020). By listing a species as endangered, the government would open legal avenues for developmental projects to be delayed and millions of dollars to be incurred (Metrick and Weitzman 1996). In addition, the listed species would be eligible to receive funds for their recovery, in order to have their endangerment reduce to levels that would warrant their de-listing (Metrick and Weitzman 1996). It is clear that the costs of this type of environmental protection are very considerable, growing faster than any other expenditure of comparable size in the American economy (Metrick and Weitzman 1996). For these reasons, understanding the factors governing the decision to list a species under the ESA deserves serious attention. However, this subject has so far received little formal analysis.

In this paper, we investigate the main determinants of government decisions about the listing of endangered species, and apply secondary research to investigate how listing might evolve over the 21st century due to climate change. We follow a reproduction of a paper by Moore et al., "Noah's Ark in a Warming World: Climate Change, Biodiversity Loss, and Public Adaptation costs in the United States" (2022), published in the *Journal of the Association of Environmental and Resource Economists*. The study population consists of $n = 60,481$ species endemic to the United States. The estimand which we intend to replicate is the binary listing status of a species at the time of its assessment. The data are obtained from a number of open government and scientific sources, including NatureServe, Google Ngram, Integrated Taxonomic Information Service (ITIS), National Centre for Biotechnology

Information (NCBI) and the United States Fish & Wildlife Service (USFWS). As explanatory variables, we use proxies of both scientific and anthropocentric characteristics of species, including taxonomic classification, NatureServe conservation status, scientific and common name n-gram frequency, and evolutionary uniqueness. The anthropocentric characteristics, including taxonomic classification and n-gram frequency, are so named, because they are used in this analysis to estimate the species' utility value to humans.

Our paper successfully replicates three of their following research claims, that the probability of listing a species: (1) changes with its conservation status, decreasing from the most to least endangered, (2) increases with its evolutionary uniqueness and (3) is associated with its utility value to humans. In particular, imperiled vertebrates are more likely to be listed than imperiled non-vertebrates, and species more commonly evoked in scientific and popular literature are more likely to be listed than species less frequently written about. This shows that both scientific and anthropocentric characteristics play important roles in the government's decisions to list individual species. Due to climate change, a global warming of 5°C over the next 100 years would imply that the number of listed species would grow linearly by 690 species, with an additional 2,800 species becoming critically imperiled but remaining unlisted. Our reproduction is conducted using the open-source statistical programming language `R` (R Core Team 2023), with functionalities from `arrow` (Richardson et al. 2024), `car` (Fox and Weisberg 2019), `knitr` (Xie 2023), `kableExtra` (Zhu 2021), `here` (Müller 2020), `patchwork` (Pedersen 2024), `rstanarm` (Goodrich et al. 2024), `tidyverse` (Wickham et al. 2019) and `taxize` (Chamberlain and Szocs 2013).

The remainder of this paper is structured as follows: Section 2 introduces the data used for analysis, including visualisations of the variables of interest. Section 3 presents and validates the logistic regression model used to determine the relative importance of various species characteristics to USFWS' listing decision. Section 4 displays the interpretations of the model alongside other findings gained from analysis of the data. Section 5 addresses how these results are projected to evolve in the future under the effects of climate change. It also provides a discussion as to the implications of the findings, limitations of the paper, and next steps for future investigation.

2 Data

2.1 Source

The paper and data used for replication are obtained from "Noah's Arc in a Warming World: Climate Change, Biodiversity Loss, and Public Adaptation Costs in the United States" (Moore et al. 2022), published in the *Journal of the Association of Environmental and Resource Economists*. Their analysis sought, in part, to explore how the probability of a species being listed under the ESA was influenced by its conservation status, utility value to humans, and species uniqueness. Our replication seeks to address the validity of these three findings and to discuss the projected increases in species listing due to climate change.

2.2 Methodology

The original dataset is made publicly available by Moore et al. (2022). It composed of $n = 60,481$ observations, with each row representing a unique species and each column indicating a specific variable. As part of our reproduction, we removed variables not of interest to our analysis and clarified the names of variables to make them easier to work with. A further discussion as to the variables of actual interest, their source of data, measurement, and the methodology employed to clean them, follow.

Table 1: Definition of Conservation Status Ranks used by NatureServe.

Rank	Category	Definition
G1	**Critically Imperiled**	At very high risk of extinction, due to very restricted range, extreme rarity, very severe threats or other factors.
G2	**Imperiled**	At high risk of extinction due to restricted range, few occurrences, severe threats or other factors.
G3	**Vulnerable**	Rare or local, due to restricted range, few occurences, recent or widespread threats or other factors.
G4	**Apparently Secure**	Uncommon, but not rare, with some cause for long-term concern.
G5	**Secure**	Common, widespread and abundant.

2.2.1 Endangerment of Species

The taxon and conservation status of species were obtained from NatureServe, an authoritative source of biodiversity data throughout the United States (NatureServe 2024). The database is dynamic, being continuously updated by botanists, zoologists and ecologists, with inputs from scientists at heritage and conservation programs (NatureServe 2024). The assessment of species is revised on a periodic basis, for we use the most recent assessment conducted between 1985 and 2019.

There are nine taxa or classes of living things, of which there are five vertebrates species – amphibians, birds, fishes, mammals, reptiles – and four non-vertebrate species – plants, invertebrates, fungi and protists. As argued by Metrick and Weitzman (1996), a possible component of the utility value of species is the degree to which it is considered a "higher form of life." In many contexts, it seems that human beings value the existence of living things more in the degree to which they are related to them or can "identify" with them (Metrick and Weitzman 1996). This is particularly apparent in the case of "charismatic megafauna", large vertebrates with widespread popular appeal. Dividing living things into taxa allows us to test for the possible role of such a component of existence value in determining the listing likelihood of a species. To actually determine the taxonomic classification of a species, NatureServe employs trained experts, who make judgments on information from field surveys, taxonomic treatments and other scientific publications (NatureServe 2024).

Moreover, there are five NatureServe conservation status rankings assigned to extant species. NatureServe (2024) provides rigorous definitions of these rankings, which we summarise in Table 1. A standardised NatureServe protocol is used to assign conservation status ranking to species on the basis of ten factors, regrouped into three categories (Faber-Langendoen et al. 2009). Of these categories, the most important is rarity, composing population size and extent. The other two are the anthropogenic threats to the species and the long- and short-term trends in population size or range extent (Faber-Langendoen et al. 2009). Trained experts analyse data from primary literature and the field to assign scores for individual factors, which are then translated into one of five final conservation status ranks (Faber-Langendoen et al. 2009), as outlined in Table 1.

A breakdown of the conservation status by taxon is shown in Figure 1, which is a replication of Figure 1B in the paper by Moore et al. (2022). To improve the visualisation of the result, we have coloured each status ranking, according to the conventional colours employed by NatureServe (2024). Observe 60% of species have an unknown status, being almost entirely of the invertebrate (41%), plants (40%) or fungi (16%) categories. This distribution of missing data is informative, as it suggests a lack of scientific attention to these three non-vertebrate species that appear not to be highly prioritised in the ranking process (Moore et al. 2022). In terms of the data that are known, reptiles appear to have the greatest proportion of known secure species, whilst mammals have the greatest proportion of known critically imperiled species.

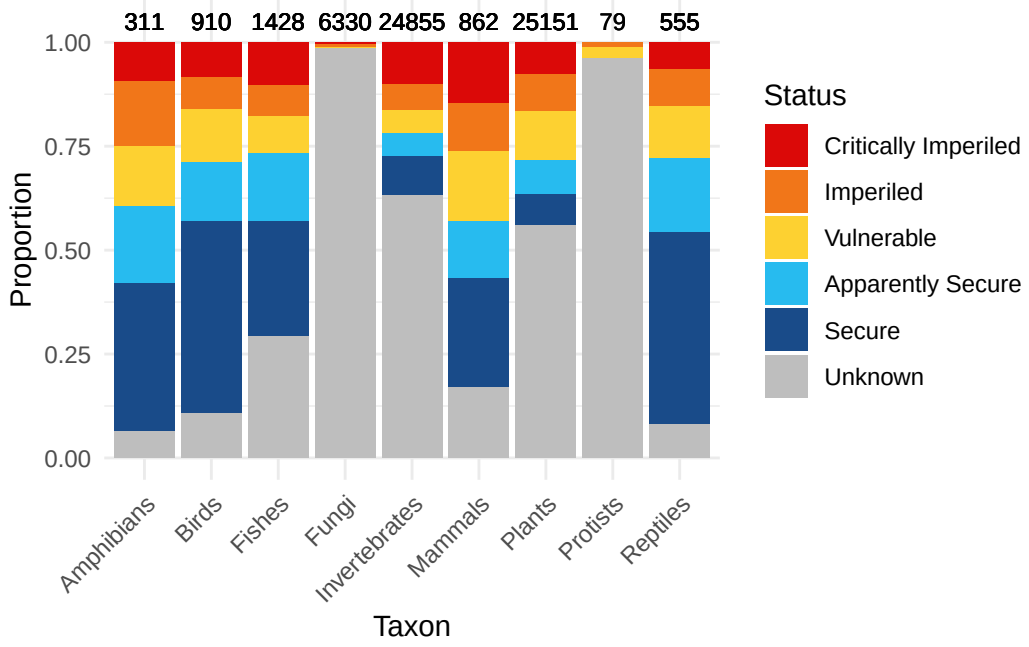

Figure 1: Distribution of assessed conservation status using NatureServe ranking system. Number above the bars gives the total number of species per taxon.

2.2.2 Species Uniqueness

Species uniqueness measures how evolutionarily isolated a species is on its family (or phylogenetic) tree (Redding et al. 2008). A highly unique species is expected to have split off long ago from its nearest living relative and, hence, to have evolved measurably different genetic features and functions compared to any other member in its phylogenetic tree (Redding et al. 2008). On a phylogenetic tree will generally appear members of the species' clade, which is to say the group of species that are descendants of a common ancestor (Moore et al. 2022). To illustrate, the Siberian tiger and North American brown bear are not regarded as evolutionarily unique: each has a number of closely-related species that are highly genetically similar and in little danger of going extinct (Metrick and Weitzman 1996). At the opposite extreme are the platypus and narwhal that are so evolutionarily unique as to each form a monotypic genus: they are the only genetically distinct representatives of their entire genus and are very distantly related to their nearest relatives in other genera (Metrick and Weitzman 1996). Note that a genus (or, in plural, genera) simply refers to a taxonomic classification of organisms between the levels of species and family.

The estimate of species uniqueness employed in this analysis is the number of species within a genus (or genus size), an idea to be accredited to Metrick and Weitzman (1996). As discussed previously, a large genus size should imply a smaller degree of species uniqueness, and vice-versa. A benefit to this measure is that, because genera were historically defined based on obvious similarities in physical traits, it may capture aspects of distinctiveness that are more visually evident to humans (Moore et al. 2022). Another benefit is that data on genus size are widely available for almost every species ($n = 55,406/64,583$) in our dataset. For our analysis, the data were obtained from two automated databases, the Integrated Taxonomic Information Service (ITIS) and the National Centre for Biotechnology Information (NCBI), using the `taxize` package in R (Chamberlain and Szocs 2013). Note ITIS and NCBI provide central databases that curate taxonomic information from submissions made by taxonomy services, specialists, and researchers working on primary literature in the field (ITIS et al. 2023; Schoch et al. 2020b). These experts collect and analyse morphological

and molecular data from thousands of species, and compare their values to references in relevant phylogenetic literature, in order to develop robust and consistent taxonomic classifications (Smith et al. 2006). Once taxonomic placement for each species is relatively certain, researchers may then approximate the number of species in each genera by counting (Smith et al. 2006).

In Figure 2, we plot the distribution of the logarithm of the genus size per taxon. A similar frequency distribution can be found among plants, fungi and invertebrates, characterised by a wide distribution with a median on the order of 100, and tails that extend to logged genus sizes of 1000. This suggests the presence of relatively more polytypic genera and less evolutionary uniqueness than species in amphibians, protists, reptiles, fish and mammals, whose frequency distributions are concentrated in the range of 1 to 100. A definite comparison is, however, inhibited by the substantial differences in the sizes of the datasets: as seen in the figure, there is very little data collected as to the genus size of some taxons (namely, protists and reptiles) and hundreds of times more data for other taxons (namely, invertebrates and plants).

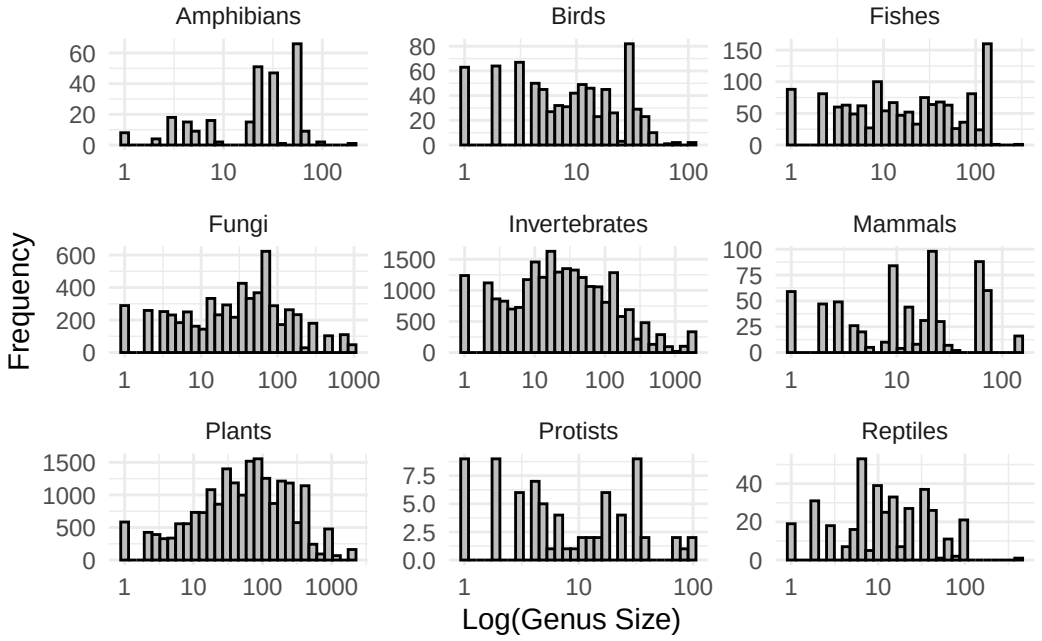

Figure 2: Distribution of the logarithm of genus size per taxon.

2.2.3 Google Ngrams for Species

An n-gram is defined as a sequence of n words in some particular order: a 2-gram, for instance, is a two-word sequence like "*Balaena mysticetus*", whilst a 3-gram is three-word sequence like "*Lithobates areolatus circulosus*." An on-line search engine, the Google Books Ngram Viewer uses Optical Character Recognition (OCR) to provide the frequency of such n-grams in a corpus of over 25 million digitised books published over the course of more than 500 years (Google 2020). These books are published in eight languages; in particular, the English corpus of 2019 consists over over 16 million of such books published between 1470 and 2019 (Michel et al. 2011).

As argued by Moore et al. (2022), the rich variety of ways in which a species may provide utility – through cultural significance, commercial value, scientific interest, and so forth – should likely directly influence the frequency with which the species is written about over time. This is consistent with intuition and broader literature in applied computing, with authors positing that elements most influential on a society and culture should naturally be reflected in the written word of those who are a

part of that society and culture (Knight and Tabrizi 2016; Aisopos et al. 2016). As a result, following the workflow of Moore et al. (2022), this analysis adopts common and scientific name n-grams as an additional estimate of the utility value of species to humans. Potential limitations of this measure are provided in Section 5.4.3 and Section 5.5.

To determine the n-gram frequencies for species names, Moore et al. (2022) performed case-insensitive searches in Google Books' English corpus of 2019 for all common and scientific names present in the NatureServe and ESA archives between 1800 and 2016. They assigned a frequency of 0 across all years whenever a name failed to return valid data. Here, the purpose of treating common and scientific name n-grams separately owes to the fact that common names pose a few challenges. In particular, a species may be designated colloquially by multiple common names or may lack a common name altogether (Cheese 2021). Conversely, a single common name may be used for multiple different species (e.g., a "millipede" referring to any of the 10,000 species in the anthropod class Diplopoda) or may have additional, unrelated meanings or uses (e.g., "British soldier" referring to an army serviceman of the United Kingdom or to the lichen *Cladonia cristatella*) (Cheese 2021). To account for such complicating factors, common and scientific names were thus treated separately.

As described by Moore et al. (2022), n-gram frequencies were then aggregated to the species level. More specifically, for unlisted species, common and scientific n-gram frequencies were taken to be their respective averages from 1950 to 2016. For listed species, they were calculated as their respective average frequencies from 1950 to 10 years *prior* to the date of the listing decision. This was done to minimise the bias to the n-gram values following any publicity generated by the listing decision (Moore et al. 2022). Owing to how small the raw n-gram frequencies were, all frequencies were then z-score normalised: this was accomplished by calculating the difference between each raw frequency and the corresponding n-gram mean and dividing the whole by the corresponding n-gram standard deviation. At last, to further account for the challenges posed by common names, as discussed previously, any species whose standardised common name n-gram had a frequency > 10 times its corresponding standardised scientific n-gram frequency were discarded in the dataset (Moore et al. 2022).

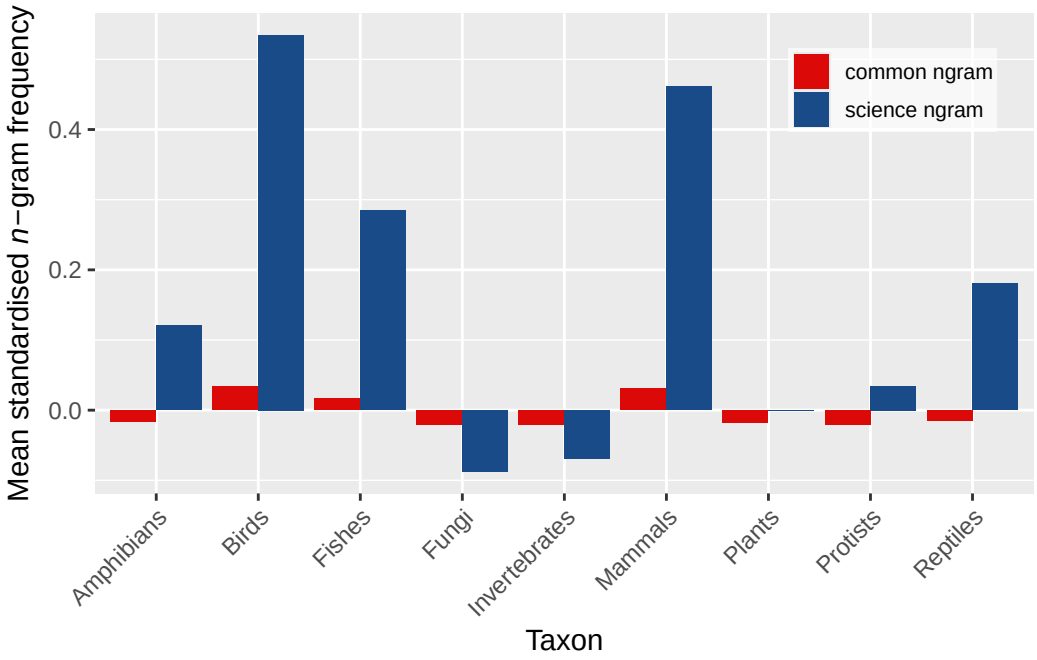

Figure 3: Distribution of the mean standardised common and scientific n-gram frequencies per taxon.

Table 2: Summary statistics table of the number of species listed under the ESA by taxon.

Taxon	Listed	Prop. (%)
Amphibians	13	4.2
Birds	61	6.7
Fishes	95	6.7
Fungi	0	0.0
Invertebrates	195	0.8
Mammals	69	8.0
Plants	606	2.4
Protists	0	0.0
Reptiles	26	4.7

A plot of the mean standardised common and scientific n-gram frequencies per taxon is shown in Figure 3. This figure shows that scientific name n-grams are typically correlated with their corresponding common name n-grams. For birds, fishes and mammals, both types of n-grams are well above the standardised mean, whilst for the non-vertebrates – comprising fungi, invertebrates and plants –, both types fall below the mean. Altogether, it appears that the utility value of a species is directly reflected in the frequency with which the species is written about *both* in colloquial and scientific settings, with vertebrate species exhibiting greater utility value than non-vertebrate species. That scientific name n-grams are much greater in magnitude than their corresponding common name n-gram may be especially driven by the preferred usage of scientific names in technical literature.

2.2.4 Listing of Species

The dependent variable of interest in this analysis is the `listed` variable, set to 1 if the species was listed for protection as of August 2020 under the ESA and to 0 otherwise. An excellent overview of the process of listing a species is described by Metrick and Weitzman (1996): it begins when the species is proposed by the USFWS as a "candidate" for protection. During its period of candidacy, USFWS collects data from internal and external sources to assess whether the species warrants listing. If sufficient scientific evidence exists, then the USFWS proceeds to submit a formal proposal under the Federal Register and makes requests for comments from the public. After this period of public surveying, USFWS comes a final decision, which officially determines whether the species is listed under the ESA.

Brief summary statistics of the total number and proportion (%) of listed species per taxon is presented in Table 2. The four non-vertebrates categories contain either no listed species (fungi, protists) or a relatively small proportion of listed species (invertebrates, plants). Instead, listing decisions are significantly biased towards the vertebrate categories, for which the average proportion is 7.3%– about three times as high as the highest proportion of 2.5% among the non-vertebrate categories.

3 Model

3.1 Model Set-up

We perform binary logistic regression to determine the relative importance of species characteristics in the listing decision. This type of regression model requires the logit link function, which takes a linear combination of the relevant covariate values and converts them to the scale of a probability between 0 and 1. In particular, for some coefficients $\beta_0, \beta_1, \beta_2, ..., \beta_U$, the logit link function is given by

$$\text{logit}(\pi_i) = \log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_U X_{iU}, \tag{1}$$

where π_i denotes the probability of success for the ith sampling unit and $X_{i1}, X_{i2}, ..., X_{iU}$ are the values of the U covariates of interests measured at the ith sampling unit (MacKenzie et al. 2018).

To apply this model to our dataset, we define, for each sampling unit i, Y_i as the listing variable, set to 1 if the species was listed as of August 2020 and to 0 otherwise. We define π_i as the corresponding probability of this event, that the species was listed as of August 2020. The $U = 5$ covariate values are X_{i1} denoting one of the nine taxonomic classes to which species i belongs, X_{i2} denoting its NatureServe conservation status, X_{i3} denoting its standardised scientific name n-gram score, X_{i4} denoting its standardised common name n-gram score and X_{i5} denoting the logarithm of its genus size. Our binary logistic regression model can then be expressed in terms of the logit link function as:

$$Y_i|\pi_i \sim \text{Bern}(\pi_i) \tag{2}$$

$$\text{logit}(\pi_i) = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i2} + \beta_4 X_{i4} + \beta_5 x_{i5} \tag{3}$$

$$\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5 \sim \text{Normal}(0, 2.5), \tag{4}$$

where the regression coefficients $\beta_1, \beta_2, ..., \beta_5$ determine the size of the respective covariates, and β_0 is the intercept term.

To implement this Bayesian model, we use the `rstanarm` package (Goodrich et al. 2024) in `R` (R Core Team 2023), with its default priors.

3.2 Model Justification

The logistic regression model is most appropriate, because the outcome of interest – whether or not a species was listed as of August 2020 – is binary and the covariates used are an admixture of categorical, discrete and continuous variables. Moreover, logistic regression is well suited to the primary interest of our analysis, which is to *predict* the probability of a species being listed when the values of its covariates are known, and to understand the relative importance of the covariates in making this prediction. Logistic regression is known to have a good predictive accuracy for many simple datasets, which proves often superior to a naïve Bayes (Jurafsky and Martin 2008). In part, this is due to logistic regression being less prone to overfitting, which occurs when a model attempts to fit too perfectly the details of the dataset on which it has been trained, including random noise (Jurafsky and Martin 2008). In tending to avoid this issue, logistic regression may be able to generalise well from the training set to the unseen data. What is more, the coefficients of logistic regression are especially interpretable for the purposes of prediction, with the magnitude of each giving an indication of the impact of the covariate on the log-odds of the outcome occurring and the sign of each giving its direction of association. We utilise this predictive power in Figure 7, where we graph the predicted probabilities of species being listed under the ESA, as implied by the regression coefficients.

Logistic regression is also advantageous in having relatively simple assumptions to be satisfied. It does not make the key assumptions of linearity, homoscedasticity and normality that are required to apply linear regression and other generalised linear models based on the ordinary least squares algorithm (Bandgar 2021). Nor does logistic regression require its covariates to possess equal variance as other similar linear classifiers do, such as discriminant function analysis (Press and Wilson 1978). We verify the assumptions that do underlie logistic regression in Section 3.3.

At last, logistic regression is simple to implement and computationally efficient, requiring few computational resources (Jurafsky and Martin 2008). This is suitable for our modestly-sized dataset ($n = 60,481$) and for other datasets in ecological literature, which often contain extensive amounts of information on various environmental variables.

3.3 Model Validation

To validate the use of a logistic regression model, we check its five main assumptions, as outlied by Stoltzfus (2011): dichotomy in the outcome, independence of observation errros, linearity in the logit for continuous variables, absence of multicollinearity and lack of strongly influential outliers.

3.3.1 Dichotomy in Outcome

A first assumption of logistic regression is that the response variable is binary or dichotomous, taking on only two possible outcomes. This assumption is evidently satisfied, because our response variable measures whether a species has been listed as of August 2020, assuming the value of 1 if true and 0 otherwise.

3.3.2 Independence of Observation Errors

A second assumption is the need for independence of error or residual terms. This means that errors associated with one observation are not correlated with errors of another (Schreiber-Gregory 2018). The presence of autocorrelation in error terms is undesirable, because it tends to inflate the significance results of coefficients, by underestimating their standard errors (Schreiber-Gregory 2018). Note that, if one's data includes repeated measures, or other correlated outcomes, then residuals would be likewise correlated, violating this assumption (Stoltzfus 2011). Thus, we have first checked that our data contains no duplicate measures, with each observation corresponding to a distinct species.

To check for the presence of autocorrelation in residuals, we implement a widely-used statistical test for this purpose, known as the Durbin-Watson (DW) Test (Schreiber-Gregory 2018). When the DW statistic is between 1.5 and 2.5, it is convention to assume an absence of (first-order) autocorrelation in the residuals (Schreiber-Gregory 2018). We run this test in `R`, using the `car` (Fox and Weisberg 2019) package. As indicated in Table 3, the test detects a small positive autocorrelation of 0.05. However, this autocorrelation is not deemed significant, because the corresponding DW statistic of 1.89 lies between 1.5 and 2.5. This lack of autocorrelation in the residuals is consistent with the second assumption of logistic regression.

Table 3: Results of the DW Test.

DW Statistic	Autocorrelation
1.89	0.05

3.3.3 Linearity

A third assumption is that a linear relationship must exist between continuous covariates and the logit of the outcome (Stoltzfus 2011). To assess this assumption, we make component-plus-residual (CR) plots of our three continuous covariates – the common and scientific name n-gram frequencies and the logarithm of the genus size – in Figure 4 and Figure 5. Note CR plots are a standard tool to effectively diagnose violations of the linearity assumption in logistic regression models (Nahhas 2024), that can be implemented in `R` using `car` package (Fox and Weisberg 2019). In all three cases, we restrict the y-axis to the range of 0 to 50, so as to better visualise the details of the fitted lines.

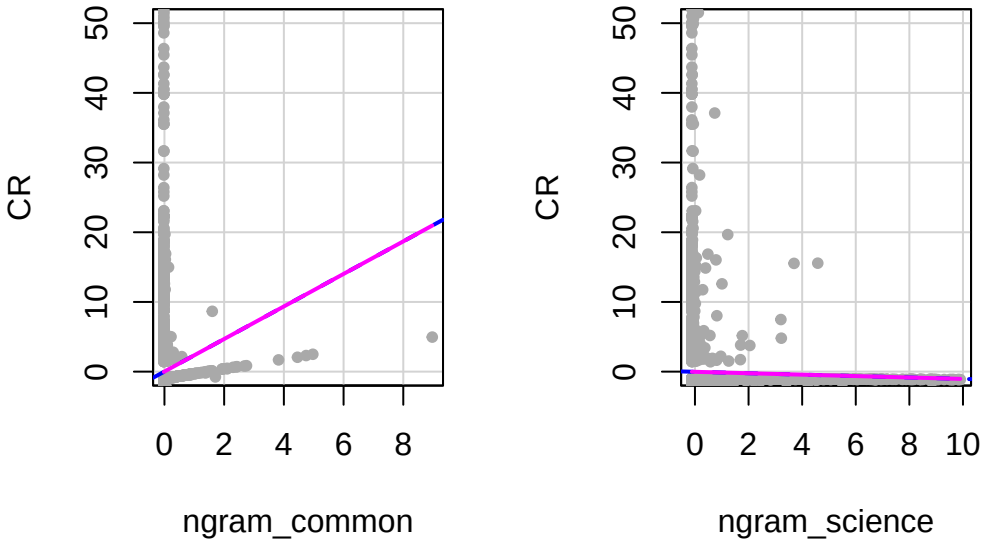

Figure 4: CR plots for common and scientific n-gram frequencies.

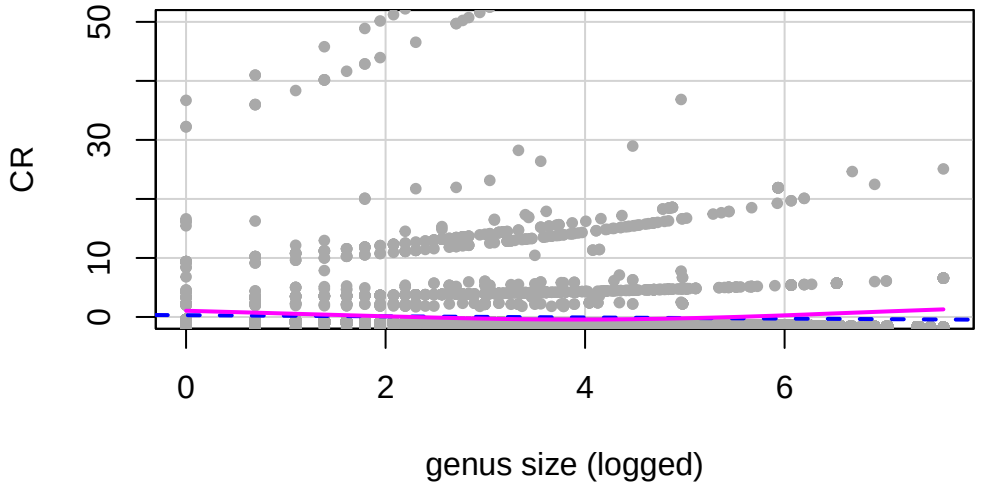

Figure 5: CR plot for the logarithm of the genus size.

In each panel above, the raw predictor values of a continuous covariate X are plotted on the horizontal axis. The residuals and contribution of X, adjusted for all the other predictors in the model, are plotted on the vertical axis. The dashed blue line illustrates the relationship between the continuous covariate and outcome under the assumption of linearity, whilst the solid line is a smoother that relaxes the linearity assumption (Nahhas 2024). To diagnose non-linearity, it suffices to observe whether or not the solid line falls exactly on top of the dashed line: a perfect coincidence of these lines signifies that the linearity assumption is perfectly met (Nahhas 2024). At the scale of our plot, there is an excellent coincidence between the solid and dashed lines in both panels Figure 4. There is a slight curvature to the solid line in Figure 5, which can be ignored – the deviation is very small and smoothers are known to be highly influenced by local fluctuations (Nahhas 2024). Having found no strong non-linearity, we may conclude that the third assumption of logistic regression is met.

3.3.4 Multicollinearity

A fourth assumption is the absence of mutlicollinearity, or high intercorrelations, among different covariates. In general, a model with highly correlated covariates will produce regression coefficients that suffer from large standard errors (Stoltzfus 2011). This leads to instability in the model. To assess the degree of multicollinearity for each covariate, we make use of the `vif` function in `R`'s `car` package (Fox and Weisberg 2019). In a generalised linear model, when two or more covariates are highly correlated, the `vif` function can quantify how much variance in an estimated regression coefficient results from this correlation (Repala 2023). The score that is assigned is called a generalised variance inflation factor (GVIF). In applying the `vif` function to our model, the following GVIFs in Table 4 are produced.

Table 4: VIF values for predictor variables.

	taxon	status	common ngram	science ngram	genus size (logged)
GVIF	1.2	1.2	1.5	1.6	1.1

It is convention that a GVIF exceeding 4 warrant further investigation, whilst GVIFs exceeding 10 indicate serious evidence of multicollinearity requiring correction (Pardoe, Simon, and Young 2018a). For our model, no significant multicollinearity was observed – all variables have GVIFs well below 4 – consistent with the fourth assumption required to apply a logistic regression model.

3.3.5 Influential Values

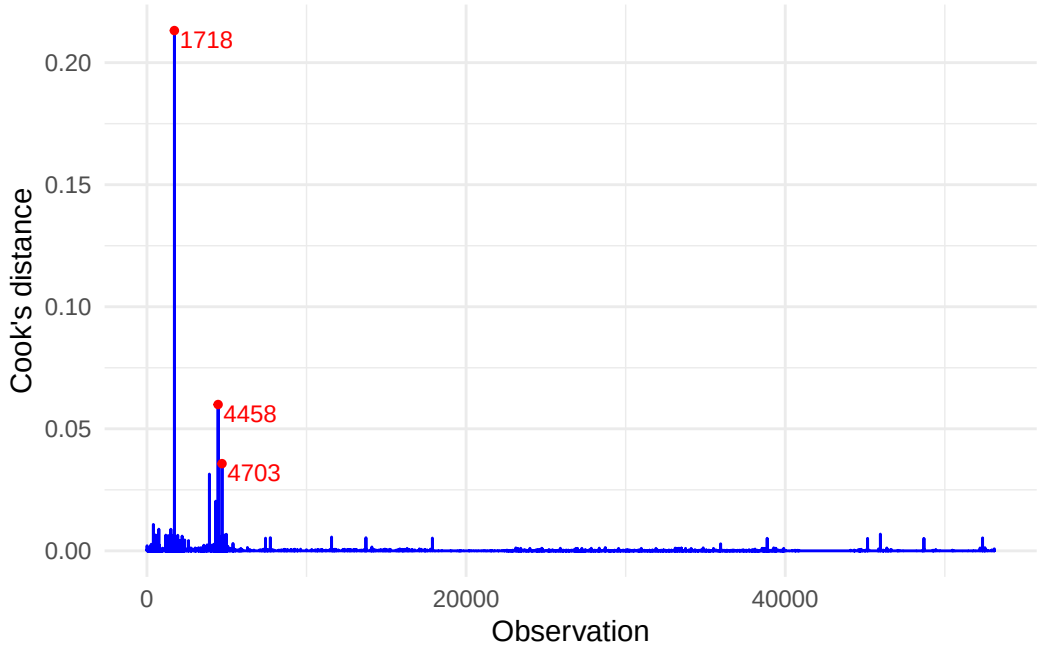

Figure 6: Cook's distance values for each observation in the dataset.

A final assumption is a lack of strongly influential outliers, observed when an observation's predicted outcome substantially differs to its actual outcome (Stoltzfus 2011). To detect the presence of outliers, which may affect the quality of the logistic regression model, we use Cook's distance, an estimate of

Table 5: Model summary of listing decision regression. SE denotes standard error.

Variable	Estimate	SE
(Intercept)	0.28	0.16
Taxon (dropped = Mammals)		
Amphibians	-0.56	0.35
Birds	0.53	0.23
Fishes	0.17	0.21
Fungi	-6.89	2.93
Invertebrates	-2.32	0.17
Plants	-0.82	0.16
Protists	-42.67	37.1
Reptiles	0.11	0.3
Conservation status (dropped = Critically imperiled)		
Imperiled	-1.37	0.08
Vulnerable	-3.01	0.14
Apparently Secure	-4.62	0.3
Secure	-7.24	0.78
Unknown	-6.02	0.29
Other covariates		
Common ngram	0.65	0.31
Science ngram	0.02	0.12
Genus size (logged)	-0.17	0.02
Observations	53,104	

the influence of a data point. The Cook's distance of a data point i, denoted D_i, quantifies how much the regression is changed when that data point is removed: in particular, a large Cook's distance signifies the point strongly influences the fitted values and vice-versa (Pardoe, Simon, and Young 2018b). It is convention that if $D_i > 0.5$ or if D_i is substantially larger than others, then the ith data point *may* be influential and warrant closer investigation (Pardoe, Simon, and Young 2018b). The distribution of Cook's distances for all observations in the dataset is shown in Figure 6, with the three observations with the highest Cook's distances labelled in red. It appears that, whilst no observation exhibits $D_i > 0.5$, there are a few observations (especially, the 1718th) whose D_i are significantly larger than the rest.

To check whether these three observations are indeed influential, we run the logistic regression model without them, the results of which are included in Table 6 in Section A.2 in the Appendix. As noted therein, changes to the estimated regression coefficients are typically on the order of 10^{-2}, well-within one standard error, and so are not significant. Given their small Cook's distances (< 0.5) and negligible weighting on the regression coefficients, we may conclude that no significant influential outliers likely exist, consistent with the fifth assumption required to apply a logistic regression model.

4 Results

Table 5 shows the coefficients of the predictor variables of the logistic regression model, defined in Section 3.1.

The estimates of the regression coefficients are not precisely the same as those provided by Moore et al. (2022). However, most agree within one standard error (SE), and preserve the same relative magnitude and sign as those obtained originally. There may be two main reasons for this discrepancy:

first, to run their regression, Moore et al. (2022) employed the `glm` function from `stats` in `R` (R Core Team 2023), rather than `stan_glm` from `rstanarm`. Accordingly, their logistic regression model was fit using maximum likelihood estimation, rather than by full Bayesian estimation by means of Markov chain Monte Carlo algorithms. Second, Moore et al. (2022) included species with NatureServe conservation statuses of "Extinct" or "Probably Extinct", whereas our analysis omits them. Indeed, we are only interested in the listing likelihood of *extant* species as opposed to those that are regarded with significant probability to be already extinct.

In spite of the discrepancies, our results sufficiently agree with those of Moore et al. (2022) as to point to the same conclusions. We discuss these findings below from a probabilistic perspective. Useful to this discussion is that, to a first approximation, logit coefficients can be converted into probabilities by multiplying them by $p(1-p)$, where p is the mean of the dependent variable (Metrick and Weitzman 1996)– in our case, we have computed $p \approx 0.02$ and $p(1-p) \approx 0.02$. Our findings follow:

(1) Regression coefficients and, thus, the likelihoods of listing, do not differ significantly among vertebrate species, including mammals, amphibians, birds, fish and reptiles. Overall, the non-vertebrates, including invertebrates, fungi, protists and plants, exhibit much smaller likelihoods of listing. To a first order approximation, *relative to mammals*, the listing probabilities of amphibians, birds, fish, fungi, invertebrates, plants, protists and reptiles differ by $-1, 1, 0.3, -14, -4, -2, -85$ and -0.2, percentage points, respectively, all other variables being equal. Overall, the taxa with the greatest and least listing likelihoods are birds and protists, respectively, all other variables being equal.

(2) The NatureServe conservation status has the expected influence on listing. The negative coefficients in Table 5, decreasing monotonically from the most (status = G1, critically imperiled) to the least (status = G5, secure) endangered, imply that less severe degrees of endangerment result in lower likelihoods of listing. A translation of the logit coefficients into probability terms implies that a one unit increase in conservation status on average results in a 2 percentage point rise in the likelihood of listing. In addition, species with unknown conservation status are less likely to be listed.

(3) The notability of a species, as measured by n-gram frequency, has a positive influence on listing. Similar translations of coefficients into probabilities show that a unit increase in mean standardised common and scientific n-gram frequencies respectively yield 0.04 and 1 percentage point rises in the likelihood of listing.

(4) Higher species uniqueness, as measured by a smallness in genus size, is associated with an increased likelihood of listing. In particular, decreasing the number of species in a genus by 10% increases the listing likelihood by 0.4 percentage points, all other variables being equal. As shown in Figure 2, this effect is largely driven by plants and invertebrates, which have larger and more variable distributions of genus size.

Moore et al. (2022) do not include an estimate $\hat{\beta}_0$ of the intercept coefficient in their summary model. For our analysis, Table 5 shows $\hat{\beta}_0 = 0.28$, which implies that the probability of listing a critically imperiled mammalian species in a monotypic genus, with a common and scientific name n-gram frequency of zero, is roughly 56.

To visualise these relationships, Figure 7 converts the logit coefficients in Table 5 into listing probabilities for each combination of taxon and NatureServe conservation status. Probabilities are estimated at the median values of the remaining model covariates, *i.e.*, n-gram values and genus sizes. Replicating the results in Moore et al. (2022), Figure 7 shows that listing probability steeply declines with improvements in the assessed conservation status of a species. For any fixed conservation status, it shows strong preferences in the listing decision for vertebrate species, notably birds, over non-vertebrates, as discussed previously.

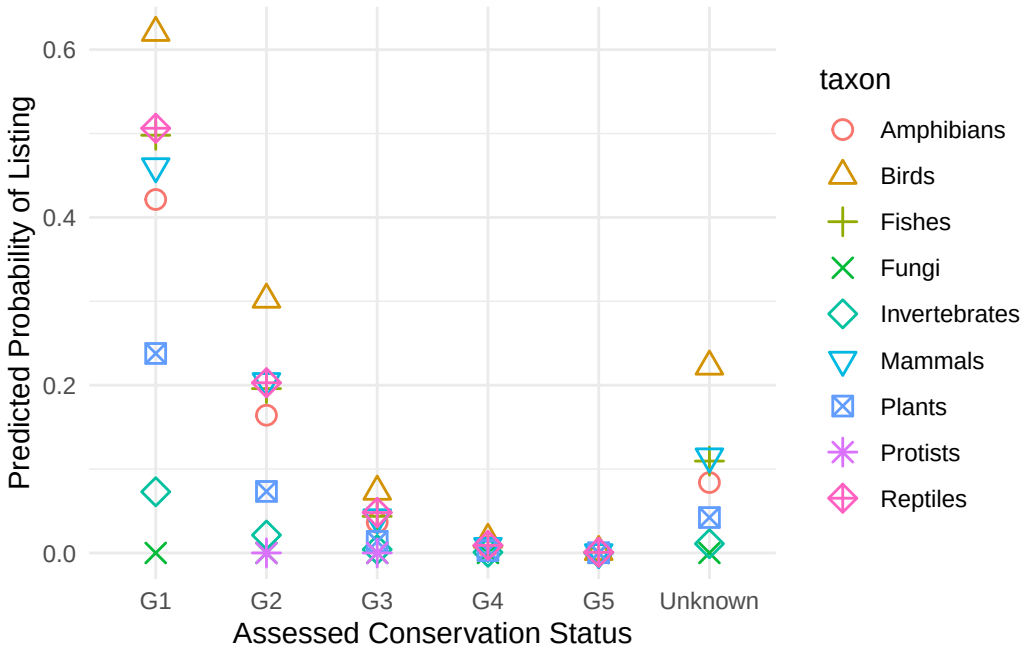

Figure 7: Probability of species appearing on the ESA list at the time of conservation assessment, as implied by cofficients in Table 5, for each taxon and NatureServe conservation status.

To extend the results of Moore et al. (2022), we also make predictions of the listing probabilities of species, as implied by the regression coefficients in Table 5, based on their scientific and common n-gram values and logged genus sizes. These three covariates assume simulated values between their actual minimum and maximum bounds, whilst the taxon and conservation status are set to their reference levels of "Mammal" and "Critically imperiled", respectively. Figure 8 shows how the predicted listing probability evolves smoothly with these three covariates, and displays the actual listing status of species in the dataset as points at binary values of 0 or 1. The general trend in all cases is that the listing probability rises with the notability of a species and its genus size, as discussed previously.

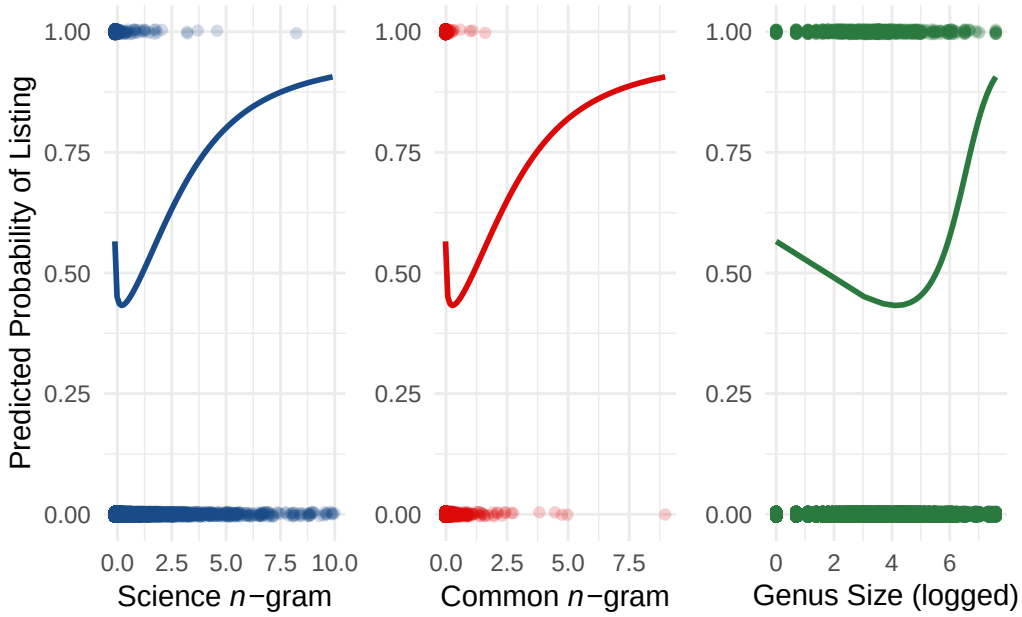

Figure 8: Probability of species appearing on the ESA list at the time of conservation assessment, as implied by cofficients in Table 5, for science and common ngrams as well as (logged) genus size.

5 Discussion

5.1 Findings

Since listing a species is the crucial first step in its protection, Moore et al. (2022) seek to gain insights into the determinants of the government's decision to list the species under the ESA. Using a well-validated binary logistic regression model, our paper has successfully replicated and extended three of their major findings:

(1) The probability of listing appreciably changes with conservation status, decreasing monotonically from most (G1) to least endangered (G5) by an average of 2 percentage points per one unit decrease in status.

(2) The probability of listing is greater for species with higher evolutionary uniqueness, which belong to smaller genera. Evolutionary uniqueness has less influence than conservation status on the listing likelihood, with a 10% decrease in genus size increasing the listing likelihood by 0.4 percentage points.

(3) Some evidence points to the probability of listing a species to be associated with its utility value to humans. In particular, imperiled vertebrates, towards which the general public tends to associate more charisma and value (Kellert 1993), are much more likely to be listed than their non-vertebrate counterparts. In addition, a species more frequently evoked in popular and scientific literature over the last 200 years, as measured by its mean standardised n-gram frequencies, are more likely to be listed than those less frequently written about.

All findings are expected, with the first two consistent with the criteria used by the USFWS when assessing a candidate species for listing. In the USFWS priority system, the actual level of endangerment of the species in question is the highest criterion for assessing listing, followed by its evolutionary uniqueness (U.S. Fish & Wildlife Service 2020). Other things equal, a higher likelihood of listing should thus be appreciably associated with species in greater endangerment and, secondly, with smaller genus

sizes, consistent with the findings in (1) and (2). The USFWS places such priority on species uniqueness, because more biodiversity is generally preserved when one saves a species that is one of the few members of its genus than one of many closely related species. Phrased differently, the extinction of the sole member of a monotypic genus produces a generally greater environmental impact than the extinction of one of the many related members of a polytypic genus (Vidal et al. 2024).

The USFWS claims to give no preference to popular species or so-called "higher-forms of life" (U.S. Fish & Wildlife Service 2020), which seems, at first glance, contradictory to our third finding. However, we find that upholding this claim is unfeasible, for the USFWS has also acknowledged the importance of political and economic influences, as well as the role of public opinion, in shaping its listing activities (Lieben 1997). In addition, a listing decision is made only when sufficient evidence has been gathered from scientific research, which has itself been observed to be systematically biased towards vertebrates (Fazey, Fischer, and Lindenmayer 2005; Titley, Snaddon, and Turner 2017). A number of factors have been proposed to explain this phenomenon, perhaps the most interesting of which is that researchers, their funding bodies or the media simply find studying charismatic vertebrates more appealing, owing to their impression as being a "higher form of life" with which human beings are more closely related or can more easily 'identify' (Metrick and Weitzman 1996). The taxonomic weighting towards vertebrates and under-representation of invertebrates in literature, as reflected in n-gram frequencies, have implications for the awareness of the natural world in scientific and popular communities (Titley, Snaddon, and Turner 2017). With a lack of awareness of the conservation needs of non-vertebrates in the scientific and public conscience, it is expected that they are less likely to be listed by the USFWS, consistent with the finding in (3).

5.2 Projected Effects of Climate Change on Listing

Making predictions of the response of biodiversity to global climate change has become a highly active area of research. They help to alert scientists and policy makers to potential risks and support the development of mitigation strategies to reduce undesirable impacts (Bellard et al. 2012). Though climate change models remain insufficiently well-developed, the best evidence suggests climate change will represent the greatest threat to global biodiversity over the next few decades, surpassing habitat destruction (Leadley et al. 2010).

An important model of the extinction rate of species from climate change was created by Urban (2015). In a Bayesian meta-analysis, Urban (2015) estimates how the extinction rate of species would vary depending on future mean global temperature increases and taxonomic groups. The Representative Concentration Pathway (RCP) 8.5 constitutes one of Urban (2015)'s model of future mean global temperatures, characterised by high emissions of greenhouse gases and a total global warming of 5°C by 2100. Under this trajectory, the study shows an estimated linear effect of warming on the extinction rate of species across all taxonomic groups. Overall, Urban (2015) reports that 7.9% of species are predicted to become extinct from climate change.

Perhaps one of the only models of the impact of climate change on listing was created by Moore et al. (2022). They combine Urban (2015)'s model effect of global warming on extinction risk, as described previously, with the likelihood of listing conditional on conservation status. Due to the estimated linear effect of warming on extinction risk, they show a linear increase in the number of critically imperiled species across all taxonomic groups. Of these critically imperiled species, they estimate that only a proportion ($< 20\%$) would be listed, with the precise proportion per taxon depending on its absolute prevalence and differential probability of listing. Overall, under RCP 8.5, Moore et al. (2022) project that approximately 690 new species would be listed over the next 100 years, with an additional 2,800 becoming critically imperiled by remaining unlisted.

5.3 Ethical Implications

A main ethical implication is the question of whether it is ethically appropriate to assign species a utility value to humans. By utility value to humans is meant one of these three broad classes, originally defined by Metrick and Weitzman (1996): The first is *commercial* value, defined in the uses of a species as food, medicine, clothing or entertainment. The second is *existence* or *aesthetic* value, representing the simple pleasure people derive from knowing the species exists in the wild, irrespective of whether they can observe it directly. The third is *contributory* value, whereby the species renders ecological services integral to the broader ecosystem and thus indirectly to humans. However, utility values are criticised as anthropocentric, measuring the worth of living things according to the services they provide to human societies (Piccolo et al. 2022).

This analysis uses two proxies for the utility value of a species to humans, being its taxonomic classification and the frequency the species' name has been evoked in English literature over the last 200 years. The evidence in Section 4 indicates that both covariates exert important influence in predicting the listing likelihood of a species. This suggests that more attention is paid to species in the degrees to which they resemble us in size or characteristics and to which they provide utility to us. An ethically provocative interpretation is to regard USFWS' preservation policy as an expansion of rights and obligations appreciably *biased* towards species proven to be useful to human societies. Though more evidence is required to support this interpretation, it provides a reasonable explanation of part of our results. Consistent with Lieben (1997), it implies that the listing activities of the USFWS are subject to the influence of political and popular influences, and not merely to scientific evidence.

5.4 Accounting for Bias

5.4.1 Endangerment of Species

To perform NatureServe assessments, experienced scientists consider information on factors relevant to the persistence of a species. Each factor has a quantitative threshold, but the weight given to each factor is subject to the assessor's judgment (Regan, Master, and Hammerson 2004). What is worse, there is no published documentation to date of the process of weighting and combining each factor to determine an overall NatureServe rank (Regan, Master, and Hammerson 2004). As such, determining an overall threat category is in part subjective, which may lead to biases or inconsistencies in the classification of a species' conservation status. Following Hayward et al. (2015), inconsistencies in applying the NatureServe criteria can be separated into two broad classes: subconscious and conscious bias. The first arises when the NatureServe guidelines are unclear or ambiguous, leading to assessors making differential interpretations depending on their level of experience (Hayward et al. 2015). The second occurs when experts make decisions based on personal values or agendas– particularly concerning, as it may introduce politicisation into NatureServe assessments (Hayward et al. 2015). To quantify the amount of error present, Regan, Master and Hammerson (2004) captured the assumptions, reasoning and logical ordering of conservation experts, mapping this partially subjective process onto a set of explicit and objective decision rules. They denoted this algorithm, that formalised expert knowledge in a transparent and repeatable way, the NatureServe method. Regan et al. (2004) discovered that 77% of species' assessments using the NatureServe method matched the more subjective assessments done by NatureServe staff, with no rank varying by more than one rank level.

In the context of this analysis, little can be done to account for the possible bias in NatureServe assessment data. All other credible sources of conservation rank assessment data, such as the International Union for Conservation of Nature (IUCN) Red List, rely on the subjective judgment of its assessors (Hayward et al. 2015). Comparisons between NatureServe, IUCN and other widely-used classification protocols have been made, with results finding the correspondence between methods to be unpredictable, with large variation among assessors (Regan et al. 2005). The analysis is thus limited by the subjective procedure used to collect the NatureServe assessment data. The large match

rate found by Regan et al. (2004) is however suggestive of the fact that subjectivity is much less significant of a factor in NatureServe assessments than the objective expertise of its assessors.

5.4.2 Species Uniqueness

The NCBI and ITIS databases, from which we obtained genus size data, are also neither error-free nor complete. To minimise the possibility of classification and processing errors, the organisations employ expert curators, who select and verify high-quality data, and taxonomists, who investigate and correct parts of the classification (Schoch et al. 2020a). Any disagreements are at once reported for review by a separate taxonomist, resulting in a high-quality database covering the major lineages of species (Schoch et al. 2020a).

5.4.3 Google Ngrams for Species

The Google Ngram frequency data are also subject to biases and errors, of which we discuss four possible sources: truncation bias, Optical Character Recognition (OCR) errors, skewed representativeness and metadata errors. We begin by arguing that the first two sources of errors are not significant to our analysis: As to the first, Moore et al. (2022) reported occasions when the name of a species failed to return valid data between the years 1800 and 2016. In these cases, they simply set its corresponding n-gram frequencies to zero across all years. The null data was due to Google Books' tendency to *truncate* data for any word or phrase occurring in fewer than 40 books (Moore et al. 2022). Our analysis uses n-grams merely as a rough proxy of the relative popularity of species, so we do not expect truncating data below a low threshold to significantly influence the interpretation of our results. As to the second, Google Books is known to contain optical character recognition (OCR) errors, mostly owing to the poor print quality of ancient books (Solovyev, Bochkarev, and Akhtyamova 2020). However, with recent improvements, the recognition error rate is now so negligibly low that it is not thought to seriously affect the results of statistical research (Solovyev, Bochkarev, and Akhtyamova 2020).

The two sources of error that may be significant to our analysis follow: in corpus design, a *representative* corpus contains balanced proportion of texts of different genres and types across different time periods (Solovyev, Bochkarev, and Akhtyamova 2020). Critics have argued that Google Books' English corpus is not a representative corpus, being composed of a sample of publications heavily biased towards scientific rather than colloquial works (Pechenick, Danforth, and Dodds 2015). This is reflected in the substantial difference between the mean scientific and common name n-gram frequencies shown in Figure 3. As such, the common n-gram measure is biased in that it vastly under-represents the popularity of species in colloquial works. Finally, Google Books suffers from metadata errors at a high rate of 37%, characterised by a flawed determination of title (15%), author (24%), publication date (20%), and so forth (James and Weiss 2012). Since our n-grams were calculated as the average of frequencies from 1950 to 2019 (or to 10 years prior to a listing decision), it may be the case that some 20% of frequencies were in fact not calculated within this period. Such errors may be important to our analysis, because we are concerned with estimating the utility value of species in modern society and culture.

An important step performed to minimise bias in n-gram frequencies is described in Section 2.2.3. To summarise, the n-grams of any single listed species were averaged from 1950 to 10 years *prior* to the date of its listing decision. This helped to minimise bias to the n-gram values following any publicity generated by the listing decision (Moore et al. 2022).

5.4.4 Listing of Species

Sources of bias in the ESA listing data are discussed in Section 5.1 and Section 5.3. To summarise, the ESA data are systematically skewed to vertebrate species, reflecting the bias of scientific and popular communities to charismatic species with which they can more closely identify (Metrick and Weitzman

1996). Moreover, USFWS' listing activities are swayed by political and economic influences, as well as public opinion, rather than merely by impartial analysis on scientific facts (Lieben 1997). Like the NatureServe assessment data, appreciable conscious bias is present in the ESA listing data, where political commentators, direct agency admissions and judicial rulings intervene to make decisions based on their own personal values or agendas (Lieben 1997).

5.5 Limitations

We are concerned with the degree to which genus size is a good measure of species uniqueness. To investigate this question, we will describe a second measure of species uniqueness, known as evolutionary distinctiveness (ED), which has received the most widespread use of all uniqueness measures (Gumbs et al. 2018). ED describes the species' relative contribution to the total evolutionary history of its clade (Moore et al. 2022). Known to be more accurate proxy than genus size, ED is computed by first assigning a value to each branch in the species' phylogenetic tree, equal to the branch's length (measured in millions of years) divided by the number of species subtending it (Isaac et al. 2012). The ED of a species is then the sum of these values for all branches from which the species is descended (Isaac et al. 2012). Moore et al. (2022) obtained published ED scores for mammals, amphibians, birds and reptiles from the Evolutionary Distinctive Globally Endangered (EDGE) of Existence program. (As of August 2020, no data were yet available for fish, fungi, invertebrates or protists, and too few data ($n = 319/19,092$) were yet known for the plants in our study. This resulted in our decision to omit any consideration of ED scores in this analysis). Moore et al. (2022) then plotted ED scores against genera size for species, for which ED scores were available. Although they noted the relationship tended in the expected direction, whereby species in larger genera tend to be less evolutionarily distinct, they established that only a very small proportion of the variance ($R^2 = 3.6$) in the ED variable could be explained by the genus size variable (Moore et al. 2022). This shows that genus size is only a very weak measure of the ED of a species. Taking ED as the superior measure for species uniqueness, it follows that genus size is a limited measure of species uniqueness.

An evident limitation of the Google Ngram measure is that it captures only material published in digitised books, excluding popular media such as magazines, websites or newspapers. It cannot incorporate the frequency with which the species name is evoked in oral discourse, videos or podcasts, though these are important and rich media for determining the elements influential on a society and culture. Given the over-saturation of scientific texts in Google Books' English corpus (Solovyev, Bochkarev, and Akhtyamova 2020), the common name n-gram measure misses many written media platforms where non-technical writers may freely communicate on species and ecological issues of interest. Moreover, the English corpus exclusively includes material written in English, thereby failing to account for documentation of the species in other languages such as Spanish, the second most commonly spoken language in the United States. This further limits the n-gram measure, as solely assessing the utility value of species to an English-speaking public.

Although the USFWS mandates the listing all species at risk of extinction, capacity and budgetary constraints often slow down the the listing process, leading to a backlog of species awaiting assessment (Alexander 2010). In some cases, USFWS will designate a species as "warranted but precluded" to acknowledge that listing the species is necessary but of lesser priority than other species in need of greater protection (Alexander 2010). The binary nature of the ESA listing variable is limited to the degree that it treats a species in "warranted but precluded" category in the same as the "unlisted" category, though the distinction between the two may be of importance. Moreover, the NatureServe assessment and ESA listing data are a snapshot of a point in time, rendering it impossible to analyse changes in listing over time. As all datasets are dated before August 2020, they may not accurately reflect current endangerment and listing statuses of the species.

5.6 Future Research

According to the Endangered Species Act (ESA) of 1973, the United States has declared a strong commitment to conserving biodiversity; but how do they spend their limited resources on this commitment? We investigated this question by studying actual decisions made by the U.S. government about which species to list and protect under the ESA. It remains to analyse the decisions of the U.S. government about how much to spend on them. At the time of the analysis, an inspection of the expenditures for the conservation of ESA-listed species show an order-of-magnitude difference between vertebrates and non-vertebrates (Moore et al. 2022). Among vertebrate species, expenditures are significantly biased towards salmonid species, notable in having large utility (especially, commercial) value to humans (Moore et al. 2022). An insightful statistical analysis for future study might be to regress USFWS' spending on a particular species in a particular year on a set of variables of interest. These variables might replicate those in our listing analysis – taxonomic class, conservation status, genus size and notability (as measured by its n-gram frequencies) – or others – such as physical size, spatial extent and number of years since listing. Three other highly interesting covariates to be considered include the species' taxonomic rarity, recovery potential and an indicator capturing whether conservation of the species conflicts with economic development. Here, rarity denotes a measure of the restriction of a species in numbers or in area, whilst potential denotes the ease of improving a species condition (Metrick and Weitzman 1996). The data for all three are made publicly available by the USFWS. Parallel to our analysis, the results of regression would reveal which scientific or anthropocentric species characteristics are prioritised in the spending decision.

In addition, future studies may use modelling to predict the change in ESA expenditures over the 21st century due to climate change. A comparative analysis to the total costs of climate change or to the discretionary costs of the U.S. government would be insightful in revealing how large these costs are. Further, estimates can be made to predict the magnitude of costs directly arising from the listing of an additional 690 species (see Section 5.2) due to climate change.

A Appendix

A.1 Supplemental Datasheet

A supplemental datasheet for the dataset used in this analysis is included with this paper. Please refer to this link in the GitHub Repository.

A.2 Additional Model Details

As described in Section 3.3.5, we run a second logistic regression model in `rstanarm` (Goodrich et al. 2024), with three points with highest Cook distance values removed from the dataset. The results of the regression are included in Table 6. A quick comparison to the results of the original regression in Table 5 show that any discrepancies in the estimates are not significant, being typically on the order of 10^{-2}, which is well-within all standard errors. This evidence points to those three high Cook distance points as not being significantly influential.

References

Aisopos, Fotis, Dimitrios Tzannetos, John Violos, and Theodora Varvarigou. 2016. "Using n-Gram Graphs for Sentiment Analysis: An Extended Study on Twitter." In *2016 IEEE Second International Conference on Big Data Computing Service and Applications (BigDataService)*, 44–51. https://doi.org/10.1109/BigDataService.2016.13.

Alexander, Kristina. 2010. "Warranted but Precluded: What That Means Under the Endangered Species Act (ESA)." Washington, DC: Congressional Research Service.

Table 6: Model summary of listing decision regression run without three points with the highest Cook distance values.

Variable	**Estimate**	**SE**
(Intercept)	0.28	0.17
Taxon (dropped = Mammals)		
Amphibians	-0.57	0.36
Birds	0.53	0.24
Fishes	0.17	0.21
Fungi	-6.62	2.64
Invertebrates	-2.31	0.17
Plants	-0.82	0.17
Protists	-42.07	38.8
Reptiles	0.11	0.28
Conservation status (dropped = Critically imperiled)		
Imperiled	-1.37	0.08
Vulnerable	-3.01	0.14
Apparently Secure	-4.61	0.3
Secure	-7.24	0.78
Unknown	-6.03	0.29
Other covariates		
Common ngram	0.65	0.31
Science ngram	0.02	0.13
Genus size (logged)	-0.17	0.02
Observations	53,101	

Bandgar, Swapnil. 2021. "Logistic Regression." *Medium*.

Bellard, Céline, Cleo Bertelsmeier, Paul Leadley, Wilfried Thuiller, and Franck Courchamp. 2012. "Impacts of Climate Change on the Future of Biodiversity." *Ecology Letters* 15: 365–77. https://doi.org/10.1111/j.1461-0248.2011.01736.x.

Chamberlain, Scott, and Eduard Szocs. 2013. "Taxize - Taxonomic Search and Retrieval in r." *F1000Research* 2: 191. https://f1000research.com/articles/2-191/v2.

Cheese, Tyler. 2021. "What's in a Name? Problematic Names in the World of Wildlife." *Canadian Geographic*. Canadian Geographic. https://canadiangeographic.ca/articles/whats-in-a-name-problematic-names-in-the-world-of-wildlife/.

Cornell Law School. 2024. "16 U.S. Code § 1531 - Congressional Findings and Declaration of Purposes and Policy." https://www.law.cornell.edu/uscode/text/16/1531.

Faber-Langendoen, Don, Larry Master, Jennifer Nichols, Kristin Snow, Adele Tomaino, Roxanne Bittman, Geoffrey Hammerson, Bonnie Heidel, Leah Ramsay, and Bruce Young. 2009. "NatureServe Conservation Status Assessments: Methodology for Assigning Ranks." Arlington, VA: NatureServe.

Fazey, I, J Fischer, and DB Lindenmayer. 2005. "What Do Conservation Biologists Publish?" *Biological Conservation* 124: 63–73.

Fox, John, and Sanford Weisberg. 2019. *An R Companion to Applied Regression*. Third. Thousand Oaks CA: Sage. https://socialsciences.mcmaster.ca/jfox/Books/Companion/.

Goodrich, Ben, Jonah Gabry, Imad Ali, and Solomon Brilleman. 2024. *Rstanarm: Bayesian Applied Regression Modeling via Stan*. https://mc-stan.org/rstanarm.

Google. 2020. "Google Books Ngram Viewer." https://storage.googleapis.com/books/ngrams/books/datasetsv3.html.

Gumbs, Rikki, Claudia L. Gray, Oliver R. Wearn, and Nisha R. Owen. 2018. "Tetrapods on the EDGE: Overcoming Data Limitations to Identify Phylogenetic Conservation Priorities." *PLoS ONE* 13 (4): 1–19.

Hayward, Matt W., Matthew F. Child, Graham I. H. Kerley, Peter A. Lindsey, Michael J. Somers, and Bruce Burns. 2015. "Ambiguity in Guideline Definitions Introduces Assessor Bias and Influences Consistency in IUCN Red List Status Assessments." *Frontiers in Ecology and Evolution* 3: 87. https://doi.org/10.3389/fevo.2015.00087.

Isaac, Nick J. B., David W. Redding, Helen M. Meredith, and Kamran Safi. 2012. "Phylogenetically-Informed Priorities for Amphibian Conservation." *PLoS ONE* 7 (8): 1–8.

ITIS, S. Alexander, A. Hodson, D. Mitchell, D. Nicolson, T. Orrell, and D. Perez-Gelabert. 2023. "The Integrated Taxonomic Information System." In *Catalogue of Life Checklist (Version 2023-10-31)*, edited by O. Bánki, Y. Roskov, M. Döring, G. Ower, D. R. Hernández Robles, C. A. Plata Corredor, T. Stjernegaard Jeppesen, et al. ITIS. https://doi.org/10.48580/dfgnm-4ky.

James, Ryan, and Andrew Weiss. 2012. "An Assessment of Google Books' Metadata." *Journal of Library Metadata* 12: 15–22. https://doi.org/10.1080/19386389.2012.652566.

Jurafsky, Daniel, and James Martin. 2008. *Speech and Language Processing*. 2nd ed. Pearson.

Kellert, Stephen R. 1993. "Values and Perceptions of Invertebrates." *Conservation Biology* 7 (4): 845–55. http://www.jstor.org/stable/2386816.

Knight, Gregory P., and Nasseh Tabrizi. 2016. "Using n-Grams to Identify Time Periods of Cultural Influence." *J. Comput. Cult. Herit.* 9 (3): 15:1–19. https://doi.org/10.1145/2940332.

Leadley, Paul, Henrique M Pereira, Rob Alkemade, Juan F Fernandez-Manjarres, Vania Proenca, Jörn P W Scharlemann, et al. 2010. "Biodiversity Scenarios: Projections of 21st Century Change in Biodiversity and Associated Ecosystem Services." In *Secretariat of the Convention on Biological Diversity*, 1–132. Technical Series 50. Published by the Secretariat of the Convention on Biological Diversity, Montreal.

Lieben, Ivan J. 1997. "POLITICAL INFLUENCES ON USFWS LISTING DECISIONS UNDER THE ESA: TIME TO RETHINK PRIORITIES." *Environmental Law* 27 (4): 1323–71. http://www.jstor.org/stable/43266869.

MacKenzie, D. I., J. D. Nichols, J. A. Royle, K. H. Pollock, L. L. Bailey, and J. E. Hines. 2018. *Occupancy Estimation and Modeling: Inferring Patterns and Dynamics of Species Occurrence*. Elsevier.

Metrick, Andrew, and Martin Weitzman. 1996. "Patterns of Behavior in Endangered Species Preservation." *Land Economics* 72 (1): 1–16. https://EconPapers.repec.org/RePEc:uwp:landec:v:72:y:1996:i:1:p:1-16.

Michel, Jean-Baptiste, Yuan Kui Shen, Aviva Presser Aiden, Adrian Veres, Matthew K Gray, Joseph Porter Pickett, Dale Hoiberg, et al. 2011. "Quantitative Analysis of Culture Using Millions of Digitized Books." *Science* 331 (6014): 176–82. https://doi.org/10.1126/science.1199644.

Moore, Frances C, Arianna Stokes, Marc N Conte, and Xiaoli Dong. 2022. "Noah's Ark in a Warming World: Climate Change, Biodiversity Loss, and Public Adaptation Costs in the United States." *Journal of the Association of Environmental and Resource Economists* 9 (5): 981–1015.

Müller, Kirill. 2020. *Here: A Simpler Way to Find Your Files*. https://here.r-lib.org/.

Nahhas, Ramzi W. 2024. *Introduction to Regression Methods for Public Health Using r*. https://www.bookdown.org/rwnahhas/RMPH/blr-linearity.html.

NatureServe. 2024. "Unlocking the Power of Science to Guide Biodiversity Conservation." https://www.natureserve.org/.

Pardoe, I., L. Simon, and D. Young. 2018a. "10.7 - Detecting Multicollinearity Using Variance Inflation Factors." STAT 462; University of Pennsylvania. https://online.stat.psu.edu/stat462/node/173/.

———. 2018b. "9.5 - Identifying Influential Data Points." STAT 462; University of Pennsylvania. https://online.stat.psu.edu/stat462/node/173/.

Pechenick, Erez A, Christopher M Danforth, and Peter Sheridan Dodds. 2015. "Characterizing the Google Books Corpus: Strong Limits to Inferences of Socio-Cultural and Linguistic Evolution." *PloS One* 10 (10): e0137041. https://doi.org/10.1371/journal.pone.0137041.

Pedersen, Thomas Lin. 2024. *Patchwork: The Composer of Plots*. https://CRAN.R-project.org/package=patchwork.

Piccolo, John J, Bron Taylor, Haydn Washington, Helen Kopnina, Joe Gray, Heather Alberro, and Ewa Orlikowska. 2022. "Nature's Contributions to People and Peoples' Moral Obligations to Nature." *Biological Conservation* 270: 109572. https://doi.org/10.1016/j.biocon.2022.109572.

Press, S James, and Sandra Wilson. 1978. "Choosing Between Logistic Regression and Discriminant Analysis." *Journal of the American Statistical Association* 73 (364): 699–705.

R Core Team. 2023. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Redding, David W, Klaas Hartmann, Akihiro Mimoto, Dragan Bokal, Maud Devos, and Arne Ø Mooers. 2008. "Evolutionarily Distinctive Species Often Capture More Phylogenetic Diversity Than Expected." *Journal of Theoretical Biology* 251 (4): 606–15. https://doi.org/10.1016/j.jtbi.2007.12.006.

Regan, Tracey J., Mark A. Burgman, Michael A. McCarthy, Lawrence L. Master, David A. Keith, Georgina M. Mace, and Sandy J. Andelman. 2005. "The Consistency of Extinction Risk Classification Protocols." *Conservation Biology* 19 (6): 1969–77. https://doi.org/10.1111/j.1523-1739.2005.00235.x.

Regan, Tracey J., Lawrence L. Master, and Geoffrey A. Hammerson. 2004. "Capturing Expert Knowledge for Threatened Species Assessments: A Case Study Using NatureServe Conservation Status Ranks." *Acta Oecologica* 26 (2): 95–107. https://doi.org/10.1016/j.actao.2004.03.013.

Repala, Satya. 2023. "Tackling Multicollinearity: Understanding Variance Inflation Factor (VIF) and Mitigation Techniques."

Richardson, Neal, Ian Cook, Nic Crane, Dewey Dunnington, Romain François, Jonathan Keane, Dragoș Moldovan-Grünfeld, Jeroen Ooms, Jacob Wujciak-Jens, and Apache Arrow. 2024. *Arrow: Integration to 'Apache' 'Arrow'*. https://CRAN.R-project.org/package=arrow.

Schoch, Conrad L, Stacy Ciufo, Michael Domrachev, Carol L Hotton, Sujatha Kannan, Rada Kho-

vanskaya, Detlef Leipe, et al. 2020a. "NCBI Taxonomy: A Comprehensive Update on Curation, Resources and Tools." *Database : The Journal of Biological Databases and Curation* 2020: baaa062. https://doi.org/10.1093/database/baaa062.

Schoch, Conrad L, Stacy Ciufo, Michael Domrachev, Carol L Hotton, Sujatha Kannan, Rita Khovanskaya, Detlef Leipe, et al. 2020b. "NCBI Taxonomy: A Comprehensive Update on Curation, Resources and Tools." *Database: The Journal of Biological Databases and Curation* 2020. https://doi.org/10.1093/database/baaa062.

Schreiber-Gregory, D. 2018. "Logistic and Linear Regression Assumptions." Uniformed Services University of the Health Sciences. https://www.lexjansen.com/wuss/2018/130_Final_Paper_PDF.pdf.

Smith, Alan R., Kathleen M. Pryer, Eric Schuettpelz, Petra Korall, Harald Schneider, and Paul G. Wolf. 2006. "A Classification for Extant Ferns." *Taxon* 55 (3): 705–31. https://doi.org/10.2307/25065646.

Solovyev, V. D., V. V. Bochkarev, and S. S. Akhtyamova. 2020. "Google Books Ngram: Problems of Representativeness and Data Reliability." In *Data Analytics and Management in Data Intensive Domains. DAMDID/RCDL 2019*, edited by A. Elizarov, B. Novikov, and S. Stupnikov. Vol. 1223. Communications in Computer and Information Science. Springer, Cham. https://doi.org/10.1007/978-3-030-51913-1_10.

Stoltzfus, J. C. 2011. "Logistic Regression: A Brief Primer." *Academic Emergency Medicine* 18: 1099–1104. https://doi.org/10.1111/j.1553-2712.2011.01185.x.

Titley, Megan A, Jake L Snaddon, and Edgar C Turner. 2017. "Scientific Research on Animal Biodiversity Is Systematically Biased Towards Vertebrates and Temperate Regions." *PLOS ONE* 12 (12): e0189577. https://doi.org/10.1371/journal.pone.0189577.

Urban, Mark C. 2015. "Accelerating Extinction Risk from Climate Change." *Science* 348 (6234): 571–73.

U.S. Fish & Wildlife Service. 2020. *Federal and State Endangered Species Expenditures Fiscal Year 2020*. Washington, DC: U.S. Department of the Interior, Fish and Wildlife Service.

Vidal, M. A., N. Henríquez, C. Torres-Díaz, G. Collado, and I. S. Acuña-Rodríguez. 2024. "Identifying Strategies for Effective Biodiversity Preservation and Species Status of Chilean Amphibians." *Biology* 13: 169. https://doi.org/10.3390/biology13030169.

Weitzman, Martin L. 1992. "On Diversity*." *Quarterly Journal of Economics* 107 (2): 363–405.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.

Xie, Yihui. 2023. *knitr: A General-Purpose Package for Dynamic Report Generation in R*. https://yihui.org/knitr/.

Zhu, Hao. 2021. *kableExtra: Construct Complex Table with 'Kable' and Pipe Syntax*. https://CRAN.R-project.org/package=kableExtra.