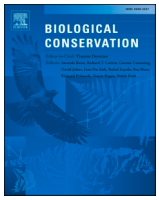
Biological Conservation 280 (2023) 109982

Contents lists available at ScienceDirect

Biological Conservation

journal homepage: www.elsevier.com/locate/biocon



Large biodiversity datasets conform to Benford's law: Implications for assessing sampling heterogeneity

Judit K. Szabo [a,b,*], Lucas Rodriguez Forti [a], Corey T. Callaghan [c]

[a] Instituto de Biologia, Universidade Federal da Bahia, Rua Barão de Jeremoabo, 668 - Campus de Ondina, CEP: 40170-115 Salvador, Bahia, Brazil
[b] College of Engineering, IT and Environment, Charles Darwin University, Casuarina, Northern Territory 0909, Australia
[c] Department of Wildlife Ecology and Conservation, Fort Lauderdale Research and Education Center, University of Florida, 33314-7719 Davie, FL, USA

ARTICLE INFO

Keywords:
Biodiversity data
Citizen science
Community science
First-digit frequency
Numeric data
Reliability
Species occurrences

ABSTRACT

Inadequate sampling can cause biased estimates of species diversity, as species occurrence generally follows a log-normal distribution with a long tail. Understanding this sampling bias is fundamental to inform biodiversity conservation actions. However, currently available tests to assess data quality, such as fitting species abundance distribution (SAD) models and rarefaction curves are computationally costly and can still lead to erroneous conclusions.

We evaluated Benford's law (first digit distribution) as a complementary method to assess data heterogeneity and survey coverage in large biodiversity datasets, including eBird data for 157 countries and three non-avian GBIF datasets. We also tested conformity to Benford's law of four simulated communities with different SAD models and four corrupted datasets with log-normal SAD. Finally, we evaluated the effect of including rare species in three datasets on the conformity to Benford's law and also compared Benford fit to the results of traditional methods to estimate survey completeness in seven datasets.

Species-rich datasets with a large number of observations tended to obtain a good fit. Benford conformity can be a simple and sensitive measure of sampling evenness, complementing traditional methods to assess quality data in large-scale studies. Benford's test can reflect species abundance heterogeneity, especially in log-normally distributed data, but was not ideal to evaluate surveys completeness, as its results diverged from those of traditional methods.

As the contribution of citizen science continues to increase in biodiversity monitoring, this fast and efficient method can play a critical role to assess the quality of datasets.

1. Introduction

With the continued decrease of biodiversity in space and time (Ceballos et al., 2017), tracking our global progress in curbing biodiversity loss is critical (Harrison et al., 2014; Magurran et al., 2010). However, the current funding for long-term ecological and conservation research is inadequate to monitor biodiversity at relevant scales (Bakker et al., 2010). The availability of data necessary to understand biodiversity status and trends has been recognised in the proposed Target 21 of the Kunming-Montreal Global Biodiversity Framework of the Convention on Biological Diversity (CBD, 2022). Nevertheless, citizen-science initiatives are a growing source of data used by research scientists, conservationists, and government agencies (Cooper et al., 2014; Theobald et al., 2015; Sullivan et al., 2017; Pocock et al., 2018; Chandler et al., 2017). With increasing taxonomic, geographic, and temporal coverage, citizen-collected data play a larger-than-ever role in informing biodiversity monitoring.

While the production and uptake of citizen science have been outstanding, particularly in countries with high gross domestic product (Meyer et al., 2015; Callaghan et al., 2021), other countries still need to increase data collection. With calls for international biodiversity monitoring and the generation of Essential Biodiversity Variables (Pereira et al., 2013; Jetz et al., 2019), nations are becoming more responsible for data collection and collation (Navarro et al., 2017). Nevertheless, quantifiable metrics need to be developed that can track the progress of a country towards closing biodiversity data gaps (Oliver et al., 2021).

* Corresponding author at: Instituto de Biologia, Universidade Federal da Bahia, Rua Barão de Jeremoabo, 668 - Campus de Ondina, CEP: 40170-115 Salvador, Bahia, Brazil.
E-mail address: judit.szabo@cdu.edu.au (J.K. Szabo).

https://doi.org/10.1016/j.biocon.2023.109982
Received 19 January 2022; Received in revised form 15 February 2023; Accepted 19 February 2023
Available online 26 February 2023

In spite of the increasing potential of citizen science data, concerns still remain surrounding data quality (Burgess et al., 2017). Indeed, in order to use citizen science data for research, conservation, and policy, we need to know when the data are 'reliable'. Data quality issues can originate from spatial, temporal, taxonomical and other biases (Szabo et al., 2012; Ward, 2014; Troudet et al., 2017), are inherent to the sampling methods used (Cox et al., 2017) and can result from incomplete sampling (Hortal et al., 2008; Beck and Schwanghart, 2010). In this work, we focus on the unevenness of the data (i.e., the heterogeneity of species' abundance), which can also result from incomplete coverage or biased sampling.

Finding adequate methods to estimate sample representativeness (including completeness) in species diversity studies is a major concern in ecology and consequently in conservation biology. Inaccurate measures of species richness and abundance originating from low detectability can lead to erroneous conclusions in biogeographical or macroecological research, impair the delimitation of priority areas for species protection and jeopardize other decision-making processes (Chao and Jost, 2012; Gotelli and Colwell, 2011; Roswell et al., 2021). Traditional statistical tools to check data quality are based on species accumulation curves, rarefaction curves, diversity estimation processes (including Hill numbers) and fit tests for species abundance distribution (SAD) models (Magurran, 2004; Lobo, 2008; Lobo et al., 2018; Chao et al., 2020). While these methods are widely accepted and provide confidence intervals and other parameters for the diversity estimate, they cannot always indicate whether the dataset is unbiased (Colwell and Coddington, 1994), as they often over-, or underestimate species richness (Melo et al., 2003). Furthermore, different diversity indices have been found to be sensitive to the sampling method used (Cox et al., 2017). Nevertheless, species richness and abundance patterns are also affected by sampling effort, as many species are naturally rare (Roswell et al., 2021; Magurran and Henderson, 2003), even at multiple spatial scales (Chiarucci et al., 2009; Warren et al., 2011). As an example, the Hill numbers paradigm (Hill, 1995; Chao et al., 2014) is a unified theory to approximate biodiversity, while accounting for abundance distribution (i.e., the relative abundance of species observed) to varying degrees based on the exponent q. Hill numbers have been widely used, as they help to overcome some of the shortcomings of biodiversity sampling (Chao et al., 2014). Nevertheless, their use requires homogeneous samples and therefore it is not recommended for unstructured datasets, such as citizen science data and museum collections.

The classic left-skewed pattern of species abundance distribution has also been used to check the reliability of datasets, as communities are usually composed of a few dominant and many rare species (Verberk et al., 2010; McGill et al., 2007a, 2007b). The classic rank-abundance analysis was originally used to evaluate empirical data to understand possible mechanisms of species structuring in natural communities (Magurran, 2004). This pattern is seen across various communities and is often explained by the fact that different species have different abilities to access limited resources, and their resource use is reflected in their abundance (Magurran, 2004). However, random datasets seem to generate similar results to what would be expected in communities structured by a niche competition process (Warren et al., 2011). In any case, the logarithmic curve shaped by the frequency distribution of species abundance apparently follows a power natural law in biodiversity datasets (Marquet et al., 2007). Along with log-normal, other SAD models have also been criticised for inaccurately representing rare species – as biodiversity researchers know, the "tail of rare species" is often longer than predicted by classical models.

Interestingly, the shape of the logarithmic curve of rank-abundance graphs of communities is the same as the theoretical distribution of digits predicted by Benford's law. Benford's law states that the leftmost non-zero digit of any given series of numbers or a set of numbers measuring any given phenomenon, is not uniformly distributed, as most numbers start with the digit 1, followed by 2 and then 3 (Newcomb, 1881; Benford, 1938). Thousands of datasets have been found to conform to Benford's law (http://www.benfordonline.net/). This method is commonly used to check reliability of numeric data in many fields, to test specific frequency distribution patterns within a dataset, and to compare them to those expected by a specific first-digit distribution described by Benford's law. It is frequently used to check for data tampering, including detecting fraud in accounting (Nigrini, 2012), political election processes (Klimek et al., 2018), disease reporting (Sambridge and Jackson, 2020) and in academia (Horton et al., 2020).

Additionally, Benford's law has been used in natural resource management to check the credibility of reported harvest numbers, including fish or trophy hunting (Cerri, 2018), and illegal deforestation (Perazzoni et al., 2020). Besides being used to detect cases of data tampering, natural biological datasets have also been found to conform to Benford's law. The number of cells in cyanobacterium colonies (Costas et al., 2008), pollen counts (Docampo et al., 2009), genome size (Friar et al., 2012) and the number of angiosperm taxa (Campos et al., 2016) all obey Benford's law to some extent. Broadly, Benford's law and other digit-based tests have been suggested as a simple initial screening step for large and complex ecological datasets (Docampo et al., 2009; Özkundakci and Pingram, 2019). This is particularly important, given that ecological datasets, boosted by the recent increase in citizen-science data, continue to grow in size and complexity (Michener, 2006).

Here we use Benford's law to evaluate data quality based on sampling heterogeneity (abundance heterogeneity among species) in large biodiversity datasets of different spatio-temporal scales, some of them contributed by citizen scientists. We also tested conformity to Benford's law as a measure of survey coverage, i.e., how completely the community has been sampled. We assumed that the results of Benford's test would be similar to traditional methods with regard to data quality in community ecology, because the digit distribution model predicted by Benford's law has a similar left-skewed shape (log-normal model) as the patterns seen in species abundance in communities. Finally, we discuss the implications of using this method to evaluate large-scale biodiversity datasets.

2. Methods

2.1. Testing the conformity to Benford's law on bird data

Mean Absolute Deviation (MAD) measures conformity to Benford's law, without considering the number of records. It is defined as the mean of the absolute value of the difference between the frequency of each first digit within the sample, and the frequency as determined by Benford's law. The higher the MAD, the larger the average difference between the actual and expected proportions, with a value above 0.015 categorised as non-conformity (Nigrini, 2012). Mantissa Arc test is another test of the probability of the data fitting Benford's distribution. The null hypothesis is that the data is uniformly distributed, and the degrees of freedom is 2 (Nigrini, 2012). We carried out goodness-of-fit testing and calculated MAD using the 'benford.analysis' package (Cinelli, 2014) in the R environment using version 4.0.3 (R Core Development Team, 2020).

We explored the conformity to Benford's law on different spatial and temporal subsets of bird data. First, we downloaded the eBird basic global dataset including observations for 2000–2020 (ebd_version_-May2020). With over 900 million bird observations, eBird is one of the most successful global citizen science projects (Sullivan et al., 2014). Compared to other taxa, bird atlases and other semi-structured and unstructured bird datasets are known to be relatively complete and of good quality at regional (Szabo et al., 2012; Troudet et al., 2017), national (Troia and McManamay, 2016) and global scales (Oliver et al., 2021; La Sorte and Somveille, 2019). Besides testing the full dataset comparing different countries, we also used the subset for the United States to compare Benford fit of regions at a subnational level and yearly. These datasets allowed us to compare the results among datasets with variable species richness that were collected within different

geographical limits and with variable sampling efforts. For each dataset, we removed species with fewer than 100 observations, since the standard procedure of fitting Benford's law recommends the using numbers with at least three digits (Nigrini, 2012). While complex socio-political factors are known to influence sampling at the global level (Leong et al., 2018; Zizka et al., 2021), we assumed similar random biases at the country level, resulting in low directional bias with regard to methods and other sampling biases among countries. Next, we selected eBird data from the United States of America, the country with the highest number of observations in the eBird dataset, to check if the patterns were similar at a finer, non-geopolitical spatial scale by aggregating observations based on North American Bird Conservation Regions (Pavlacky et al., 2017). We used the same dataset to subsample data temporally, to test the effect of cumulative effort (and therefore increasing sample size) on Benford conformity. For these tests, we filtered data to the best quality lists, including only complete checklists from surveys with 5–240 min in duration and under 5 km of distance travelled. In summary, we collated the number of observations per bird species 1) for each country, 2) each Bird Conservation Region and 3) yearly for the USA. To evaluate the effect of small sample sizes, we also checked a very small community the Birds of Joshua Tree National Park from iNaturalist dataset with 564 observations of 103 bird species https://www.inaturalist.org/observations?project_id=4786.

While Aves are known to be oversampled in GBIF with relatively high completeness, Amphibia and Plantae have lower coverage, but are still relatively well sampled and Arachnida are under-represented (Troudet et al., 2017). Therefore, we selected additional case studies from these groups (i.e., anurans, plants, and spiders), using regional datasets that we assumed to have lower species diversity and completeness than birds. Besides citizen-collected data, these datasets also include species occurrence data originating from other sources, such as museum collections and observations from scientific expeditions (https://www.gbif.org/what-is-gbif). Therefore, we downloaded three relatively large datasets from GBIF on 2 November 2021: (1) The plants of the Parisian Basin (Flore du Bassin Parisien; thereafter 'plants'; https://doi.org/10.15468/dl.2da96q) with 7.9 million observations and 5275 species, (2) the anurans of the Southern Hemisphere (frogs; https://doi.org/10.15468/dl.g48yd6) with 1.1 million observations and 3281 species, and (3) the spiders of the Southern Hemisphere (spiders; https://doi.org/10.15468/dl.6wfshp) with 516,089 observations and 9902 species. The frog and spider datasets were downloaded from the general database using GBIF filters (https://www.gbif.org/occurrence/search).

2.2. Comparing the conformity to Benford's law for datasets with different species abundance distribution

In order to assess conformity to Benford's law as a measure of sampling heterogeneity (uneven sampling of the abundance of different species in the dataset), we compared the fit to Benford's law among simulated communities based on four species abundance distribution models: log-normal, log-series, Poisson log-normal and MacArthur's broken stick for Simulations 1 to 4, respectively. We simulated these communities using the same number of individuals ($n = 7{,}839{,}439$), and a bit over 2000 species using the sim_sad function of the 'mobsim' package (May et al., 2018). We also created four biased communities based on the Costa Rica eBird dataset, manipulating the abundance of species in different ways. We separated species into categories of abundance and treated very common, common and rare species differently creating somewhat realistic scenarios (Szabo et al., 2012; Tulloch and Szabo, 2012). For the Biased 1 dataset, we decreased the number of observations for the 20 most common species to half (i.e., observers ignoring common species, for instance in the case of introduced species), increased the number of observations 50 times for the 50 least common species (i.e., observers preferentially recording rare species), and for the rest of the community, we randomly added or subtracted 1–20 % from the number of observations (to represent detection errors or misidentifications, including both false negatives and false positives). For Biased 2, we doubled the number of observations for the 20 most common species (i.e., observers preferentially recording common species compared to their real abundance), increased the number of observations 10 times for the 50 least common species (i.e., observers having some preference for rare species), and for the rest of the community, we randomly added or subtracted 1–30 % from the number of observations, to represent a somewhat higher rate of error. For Biased 3, we doubled the number of observations for the 20 most common species, eliminated the 50 least common species (i.e., the birds having extremely low detection rates or the observer lacking the knowledge to identify and therefore missing the species), and for the rest of the community, we randomly added or subtracted 1–30 % from the number of observations, and for Biased 4, we decreased the number of observations by half for the 20 most common species, eliminated the 50 rarest species and added or subtracted 1–30 % for in-between species.

2.3. The effect of including rare species

For all general calculations in this study, we removed species with fewer than 100 observations for each unit following the standard procedure of fitting Benford's law (Nigrini, 2012). However, rare species, such as habitat specialists, are important in real communities and together make up a large percentage of individuals (Verberk et al., 2010). We tested the effect of removing rare and extremely rare species on Benford fit using eBird datasets from the three representative countries (Costa Rica, Brazil and Thailand) with different Benford conformities. We tested Benford conformity and fit to unimodal gambin distribution including a) only rare (11–99 observations) and b) rare as well as very rare (1–9 observations) species in the subset. The gambin model is considered suitable for species abundance distributions and performs better than preferred models for real communities with rare species (Ugland et al., 2007). We used the "fit_abundance" function in the "gambin" package to estimate statistical parameters and to test the fit using the maximum likelihood method (Matthews et al., 2014). The test estimates α, a parameter that summarises the shape of the species abundance distribution in a single number. High α values describe a community with many abundant species and lower values indicate the presence of many rare species (Ugland et al., 2007).

2.4. Assessing conformity to Benford's law as a measure of coverage

Besides testing the applicability of Benford's Law as a measure of sampling heterogeneity in large biodiversity datasets, we performed a series of tests of its applicability for assessing biodiversity survey completeness. We first tested the sensitivity of MAD to the number of species and the number of eBird observations in each country, relating these two factors. Next, given the strong correlation between these variables ($r = 0.89$, $p < 0.01$, $n = 157$), we tested the effect of the number of observations on the natural log-transformed MAD, fitting a linear regression model using the "lm" function in R. To benchmark Benford conformity, we used an assessment of eBird survey completeness (La Sorte and Somveille, 2019). This approach estimates survey completeness by modelling the relationship between the number of species and sampling effort to develop a species accumulation curve describing the relationship between the accumulated number of species and survey effort (for full details see Lobo et al., 2018; La Sorte and Somveille, 2019). We used the equal-area hexagonal cells (49,811 km^2) from La Sorte and Somveille (2019) and averaged the values to derive an average completeness score based on cells belonging to each country and across months. We rounded these values to the nearest integer between 0 and 100 to achieve survey completeness for each country. We then fit a linear model to test for the relationship between MAD and survey completeness, where MAD was the log-transformed response variable and completeness was the predictor variable.

To relate bird survey coverage to sampling heterogeneity where sampling occurred for all terrestrial vertebrates, we used a linear regression between MAD scores of national eBird datasets and country-level Species Sampling Effectiveness Index values (Oliver et al., 2021). In this test we included 131 countries, with values in both datasets.

Based on the MAD scores, we selected a country eBird dataset from each of the close conformity, marginally acceptable conformity and non-conformity categories. For Costa Rica, which had high avian species diversity and high number of observations and was the only country with close conformity, we selected random subsets of 20, 40, 60 and 80 % of the total number of observations and calculated the Benford fit for these subsets along with chi-squared difference and summation difference. For these subsets and for Brazil and Thailand, two countries of lower completeness, we tested how the conformity to Benford's law correlated with the results of traditional measures of survey completeness, including rank-abundance graphs and rarefaction curves. We constructed rarefaction curves using the 'vegan' package and *specaccum* function based on 100 permutations (Oksanen et al., 2020). Similarly, we calculated these indices for the plant, frog and spider datasets from GBIF.

3. Results

3.1. Conformity of avian and non-avian datasets to Benford's law

Out of 253 countries in the eBird dataset, 157 had over 25 species with over 100 observations. Among these, almost two-third of the countries (93) did not conform to Benford's law (see Table 1 for selected countries, Table S1 for all countries). Generally, the total number of observations and the total number of species observed in the same country increased conformity (with the number of observations positively related to the total number of species per country; $r^2 = 0.36$, f-value = 91.96, $p < 0.01$, $n = 157$, Fig. 1). The level of conformity varied among countries, those with a higher number of observations in general had a better fit to Benford's law based on the first digits (Fig. 2). Results were similar for USA Bird Conservation Regions (Fig. S1). The small avian dataset from iNaturalist ($N1_{JoshuaTreeNP} = 103$ and $N2_{JoshuaTreeNP} = 22$) had non-conformity (with $MAD_{JoshuaTreeNP} = 0.0585$), with Mantissa mean and variance values of 0.252 ± 0.082.

The GBIF datasets for spiders (the number of observations used for first and second-order tests: $N1_{spiders} = 477$ and $N2_{spiders} = 273$) and frogs ($N1_{frogs} = 789$, $N2_{frogs} = 492$) both had non-conformity (with $MAD_{spiders} = 0.0617$ and $MAD_{frogs} = 0.0463$), with Mantissa mean and variance values of 0.323 ± 0.063 and 0.371 ± 0.074, respectively. The plant dataset ($N1_{plants} = 2071$, $N2_{plants} = 1498$) reached marginally acceptable conformity ($MAD_{plants} = 0.0139$), with a mean ± variance Mantissa value of 0.456 ± 0.085. The Pearson's χ^2 statistics were all significant at $p < 0.0001$ ($\chi^2_{spiders} = 188.78$, $\chi^2_{frogs} = 160.32$ and $\chi^2_{plants} = 51.289$). Similarly, each of the three datasets had a significant ($p < 0.0001$) Mantissa arc test ($L2_{spiders} = 0.1326$, $L2_{frogs} = 0.0673$, $L2_{plants} = 0.0067$).

3.2. Benford fit as a measure of sampling heterogeneity of simulated and corrupted community datasets

Among the four simulated communities, we found close conformity to Benford's law only for Simulated 1, which used the log-normal model for species abundance distribution (Table 2). Simulated 4, the community created using MacArthur's broken stick model reached marginally acceptable conformity. The four biased datasets, which were based on log-normally distributed eBird data from Costa Rica, also maintained a good fit to Benford even after relatively high levels of data manipulation, maintaining close conformity in Biased 1–3 and acceptable conformity in Biased 4 (Table 2).

3.3. Fitting Benford's law and the gambin model to evaluate datasets including rare and extremely rare species

Including rare and extremely rare species increased conformity to Benford's law for the eBird datasets for Brazil and Thailand (Table 3). In general, AIC_c values showed better gambin model fit for datasets from Brazil, while the Thailand dataset with only >100 observations species had the lower AIC_c value for the gambin model (Table 3).

3.4. Benford fit to assess coverage in species-rich datasets

Increased sampling, reflected by cumulatively increasing number of observations each year within the USA, steadily decreased MAD, suggesting closer conformity to Benford's law through time (Fig. 3a). Similarly, countries with high coverage in general had lower MAD scores (linear regression of MAD on survey completeness, $r^2 = 0.23$, t-value = −6.49, $p < 0.001$, $n = 140$, Fig. 3b).

The country-level correlation between MAD scores for birds and the Species Sampling Effectiveness Index for all vertebrates was poor ($r^2 = 0.0901$, $n = 131$). Some countries with non-conformity to Benford's law had low, while others had high Species Sampling Effectiveness Index (vales ranging from 0.661 to 1), while the country with close conformity had an index of (only) 0.851.

The variable levels of conformity to Benford's law among the eBird datasets from Costa Rica, Brazil and Thailand were reflected in the differences in χ^2 and summation values (Fig. 4, Table 3), while non-Benford measures, including rarefaction, species diversity accumulation and sample coverage showed satisfactory completeness (Fig. 5). Benford's fit was lowest for Thailand and highest for Costa Rica.

We found a gradual increase in Benford conformity when randomly selecting increasing subsets of 20, 40, 60, and 80 % of the 3,876,668 observations from the Costa Rica dataset (Fig. 6, Table 4). At 20 %, the Benford fit resulted in non-conformity, at 40 % it reached marginally acceptable, at 60 % acceptable and at 80 % close conformity. The non-conformity of the small avian dataset from iNaturalist also fits this pattern.

The rank-abundance analysis of the plant, frog and spider datasets showed the patterns expected when the dataset contains few common and many rare species (Fig. 7). However, this distribution of abundance

Table 1
Number of countries with different levels of conformity to the Benford Law based the first digit of the number of observations for each bird species in eBird data. Mean Absolute Deviation (MAD) values based on Nigrini (2012). Countries in bold are shown in Fig. 1. For details of all countries see Table S1.

Benford fit	MAD	Number of countries	Examples	Number of observations per species (average ± SD)
Close conformity	0.000–0.006	1	**Costa Rica**	5048.9
Acceptable conformity	0.006–0.012	10	Argentina, Australia, Canada, Chile, Honduras, Mexico, Panama, Spain, Sweden, **United States**	38,455.6 ± 84,335.8
Marginally acceptable conformity	0.012–0.015	10	**Brazil**, Cuba, El Salvador, Greece, India, Malaysia, **Netherlands**, Portugal, Puerto Rico, Turkey	2439.4 ± 3089.1
Non-conformity	Above 0.015	136	Examples: **Djibouti**, **Thailand**	477.4 ± 955.6
Failed		94	Examples: Congo (3994 observations of 351 species), Martinique (7321 observations of 127 species)	24.0 ± 33.9

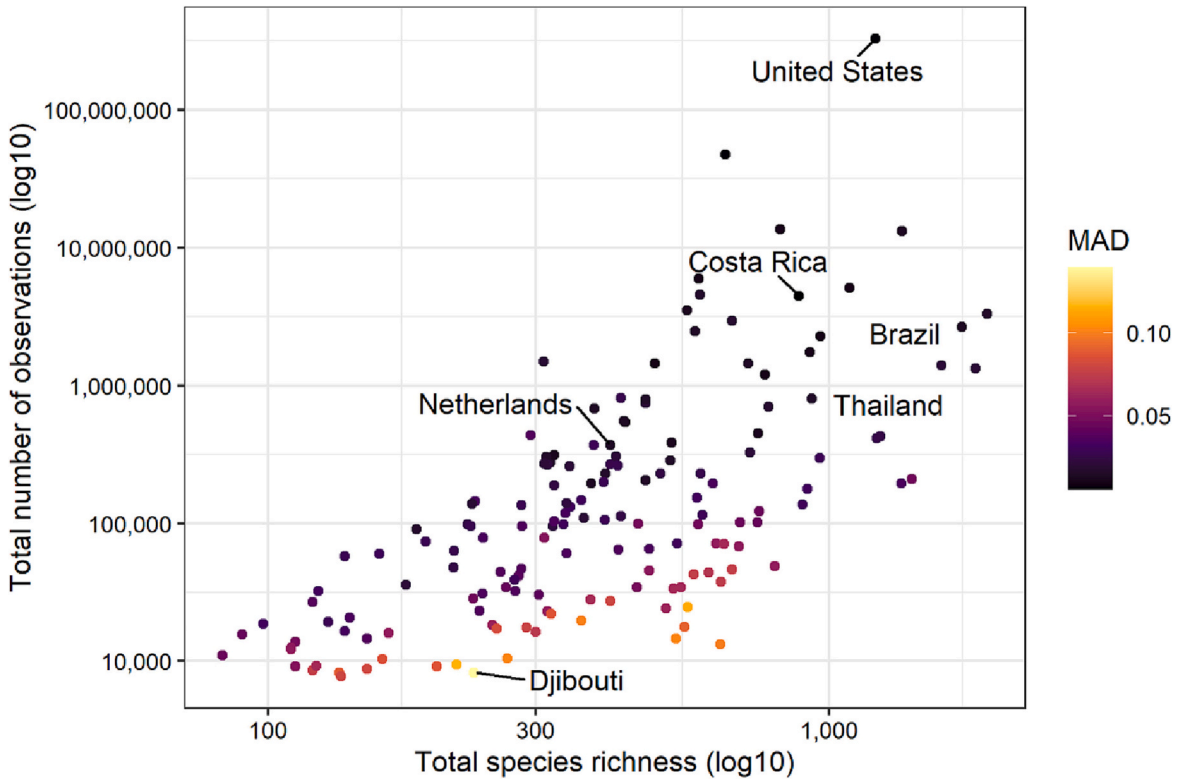


Fig. 1. Mean Absolute Deviation scores for all countries included in analysis (n = 157), as a function of their total bird species richness and the number of observations in eBird data. The six example countries highlighted in Table 1 are labelled.

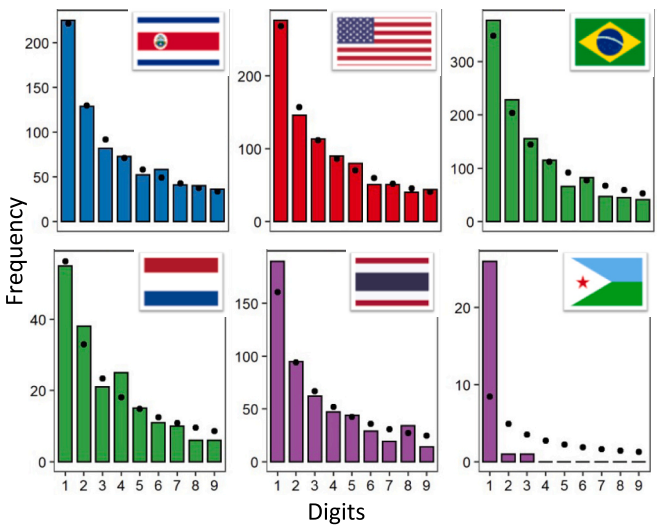


Fig. 2. Benford's law-based analysis for selected countries (Costa Rica, United States and Brazil top row from left to right and Netherlands, Thailand and Djibouti bottom row from left to right). The frequency of first digits of species observations in eBird data (coloured bars) compared to Benford's first digits' probability (black dots). Conformity is indicated by the colour of the bars; blue: close, red: acceptable, green: marginally acceptable and purple: nonconformity. Flag images are from https://flagpedia.net. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

was driven by one common species, particularly in the spider dataset. Even though in general, this is a diagnostic of well-sampled datasets, none of the rarefaction curves reached a satisfactory asymptote, similar to the one seen for birds in Costa Rica (Fig. 8). All rarefaction curves indicate that more species are likely to be included with new individuals sampled. While rank-abundance graphs seem to indicate robustness, particularly in the case of frogs and plants, these datasets did not obtain a good Benford fit.

4. Discussion

Our results suggest that analysing the conformity to Benford's law in large-scale community datasets can be a suitable test of data quality considering the fit for the best models of species abundance distributions. However, Benford conformity was not a reliable method to assess survey completeness in our sample datasets. As we hypothesised, datasets with good Benford conformity showed log-normal abundance distributions. Including rare species increased conformity in the tested datasets, suggesting that this method could be used to check an expected log-normal or gambin abundance distributions for real communities, which includes a long-tailed curve caused by the presence of rare species. Therefore, we suggest to use Benford conformity as complementary to traditional methods (e.g., rarefaction or Hill numbers) when assessing data quality.

Interestingly, the conformity to Benford's law increased in eBird datasets of some megadiverse countries that showed extremely high sampling effort and oversampling, such as the USA. Particularly for birds, biodiversity sampling effort is much higher in North America and Europe than in many developing nations (La Sorte and Somveille, 2019). Generally, the former countries also have government or NGO-coordinated systematic bird monitoring at the national scale, for instance the Breeding Bird Surveys coordinated by the United States Geological Survey in the USA and Canada, monitoring by the European Bird Census Council in European countries and by the British Trust for Ornithology in the United Kingdom. In countries with both formal surveys and ample citizen science data, bird population trend estimates based on these data sources are similar (Szabo et al., 2011; Horns et al., 2018). In contrast, many developing countries lack systematic surveys to

Table 2
Benford's law conformity decision parameters for simulated and biased community datasets. S: number of species, N: number of observations used, SAD: Species abundance distribution, $\mu \pm \sigma^2$: Mantissa mean and variance values, L2: Mantissa arc test statistic and associated *p*-value (in parentheses), MAD: Mean Absolute Deviation, NC: non-conformity, MAC: marginally acceptable conformity, AC: acceptable conformity, CC: close conformity, f: distortion factor. CV: coefficient of variation, N_s: the sum of abundances of individuals in a sample.

Dataset	S	N	SAD	$\mu \pm \sigma^2$	L2	MAD	F
Simulated 1[S1]	2024	7,839,439	Log-normal	0.495 ± 0.082	0.0011 (0.6079)	0.0037 CC	−1.9219
Simulated 2	2073	7,839,439	Log-series	0.543 ± 0.087	0.0131 (<0.0001)	0.0243 NC	1.0060
Simulated 3[S3]	2024	7,839,439	Poisson log-normal	0.518 ± 0.047	0.1499 (<0.0001)	0.0493 NC	−4.8750
Simulated 4[S4]	2024	7,839,439	MacArthur's broken stick	0.532 ± 0.083	0.0046 (0.0001)	0.0132 MAC	6.7930
Biased 1[B1]	2024	6,685,408	Log-normal	0.508 ± 0.082	0.0001 (0.8055)	0.0049 CC	1.4810
Biased 2[B2]	2024	8,997,605	Log-normal	0.4964 ± 0.0844	0.0002 (0.7359)	0.0041 CC	−0.5524
Biased 3[B3]	1974	9,013,735	Log-normal	0.5021 ± 0.0839	0.0001 (0.9592)	0.0046 CC	0.5860
Biased 4[B4]	1975	6,704,273	Log-normal	0.5062 ± 0.0819	0.0002 (0.665)	0.0067 AC	1.0505

Notes:
S1: abundance CV = 5.
S3: SAD coefficient properties: $\mu = 5$, $\sigma = 0.5$.
S4: SAD coefficient properties: $N_s = 100{,}000$.
B1: top20: divided by 2; middle: randomly ±1–20 % error; bottom50: multiplied by 50.
B2: top20: multiplied by 2; middle: randomly ±1–30 %; bottom50: multiplied by 10.
B3: top20: multiplied by 2; middle: randomly ±1–30 %; bottom50: eliminated.
B4: top20: divided by 2; middle: randomly ±1–30 %; bottom50: eliminated.

Table 3
Benford's law conformity decision and gambin model fit parameters for subsets of the three selected countries. CR: Costa Rica, B: Brazil, T: Thailand, no filter: all species observations used, >10: only species with >10 observations included, >100: only species with >100 observations included, S: number of species, $\mu \pm \sigma^2$: Mantissa mean and variance values, MAD: Mean Absolute Deviation, NC: non-conformity, MAC: marginally acceptable conformity, AC: acceptable conformity, CC: close conformity. f: distortion factor, L2: Mantissa arc test statistic and associated *p*-value (in parentheses), AIC_c: Akaike's information criterion value corrected for small sample sizes, α: estimated parameter of the gambit model, χ^2: Pearson's χ^2 statistic and associated *p*-value (in parentheses), and df is degrees of freedom with relation to the gambin model fit.

Dataset	S	$\mu \pm \sigma^2$	MAD	F	L2	AIC_c	α	χ^2	df
CR No filter	875	0.489 ± 0.088	0.0050 CC	−0.7101	0.0014 (0.2815)	4719.997	13.297	1956.82 (<0.0001)	14
CR > 10	822	0.497 ± 0.084	0.0050 CC	−0.7101	0.0005 (0.6438)	3968.961	19.422	92.378 (<0.0001)	14
CR > 100	730	0.490 ± 0.085	0.0057 CC	−2.0620	0.0003 (0.803)	3230.907	28.501	92.021 (<0.0001)	14
B No filter	1702	0.493 ± 0.082	0.0058 CC	−1.8731	0.0006 (0.3646)	8823.461	7.236	296.167 (<0.0001)	13
B > 10	1538	0.496 ± 0.079	0.0071 AC	−1.8731	0.0017 (0.0760)	7324.52	10.948	89.884 (<0.0001)	13
B > 100	1130	0.465 ± 0.077	0.0135 MAC	−8.5151	0.00824 (0.0912)	4950.417	19.759	332.001 (<0.0001)	13
T No filter	906	0.484 ± 0.089	0.0092 AC	−3.1895	0.0016 (0.2428)	4679.501	6.175	196.348 (<0.0001)	12
T > 10	782	0.486 ± 0.082	0.0098 AC	−3.1894	0.00119 (0.4301)	3646.414	10.834	57.268 (<0.0001)	12
T > 100	528	0.461 ± 0.083	0.0154 NC	−7.8992	0.0062 (0.0366)	2228.761	22.36	167.218 (<0.0001)	12

monitor birds and other biodiversity (Horns et al., 2018). Unfortunately, citizen science datasets are also often inadequate in these countries (Neate-Clegg et al., 2020), as also seen in our results. The fact that birds are not "typical" vertebrates with regard to sampling at the country level, is also reflected by the poor fit between our eBird country scores and the Species Sampling Effectiveness Index values from Oliver et al. (2021). We also demonstrated that sampling effort, as an isolated effect, was an important contributor to the conformity to Benford's law through resampling the same dataset increasing sample sizes, which led to better Benford's law conformity. Our results using a sub-country dataset (using Bird Conservation Regions as units) showed similar patterns to country-level datasets.

Testing four different biological groups in different spatial contexts (from subnational to global), we found similar results and identified variability among datasets. Among the non-avian datasets, we found the highest conformity to Benford's law in the plant dataset, which had the highest coverage, while the one with the lowest (spiders) had the highest MAD score. Testing Benford fit of simulated and biased datasets indicated better conformity to Benford's law for communities with log-normal abundance distribution. Data manipulation that did not change this distribution model were not detectable by the test (i.e., MAD scores did not increase).

While the rarefaction showed good sampling coverage for most datasets, the test for Benford fit resulted in variable conformity. When

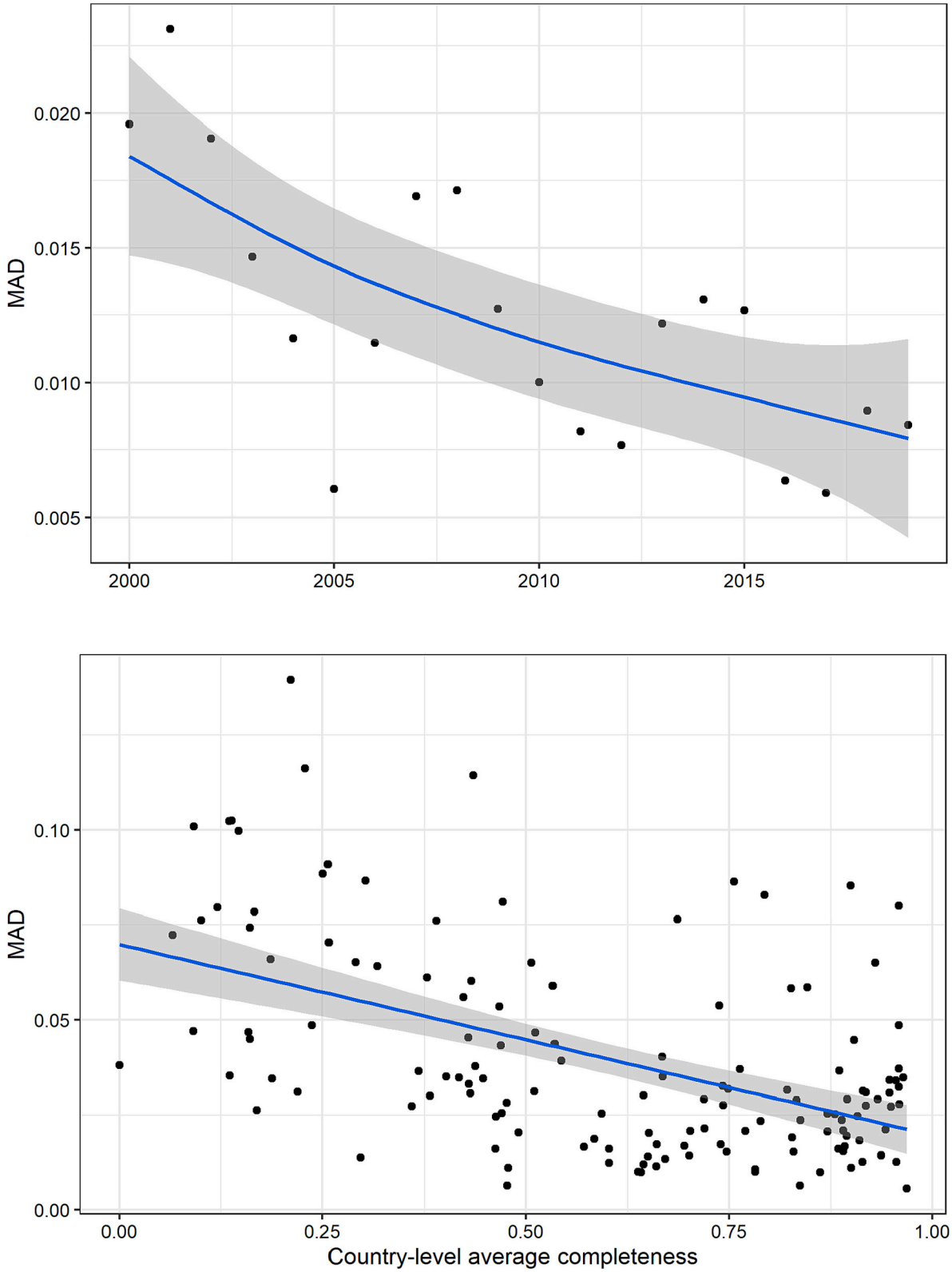


Fig. 3. a) Yearly Mean Absolute Deviation values of cumulative eBird data from the USA in 2000–2019. b) Mean Absolute Deviation score as a function of country-level averaged completeness of surveys based on eBird data. Each point represents a country ($n = 140$, countries, where no MAD value was obtained were excluded). The blue line represents a linear model regression fit. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

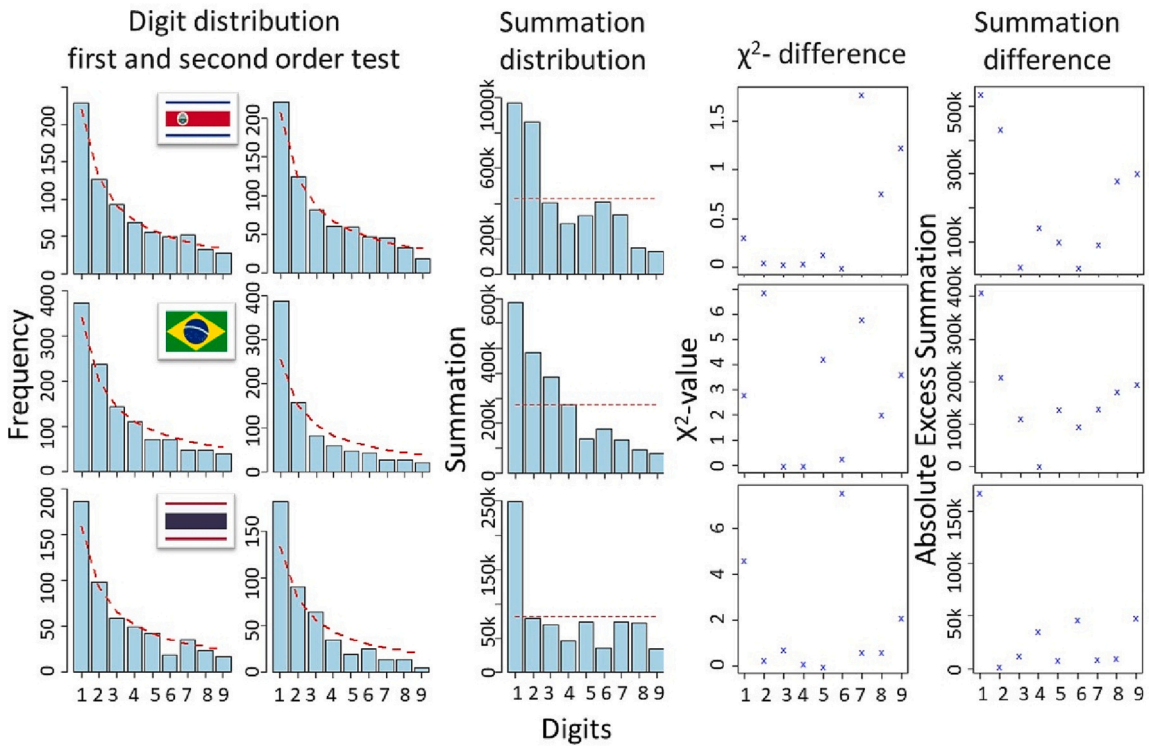


Fig. 4. Five Benford characteristics for three selected countries. In the columns from left to right: digit distribution for first order test, digit distribution for second order test, summation distribution by digits, Chi-square difference and Summation difference. The three selected countries are in the rows: Costa Rica (top), Brazil (middle) and Thailand (bottom). Flag images are from https://flagpedia.net.

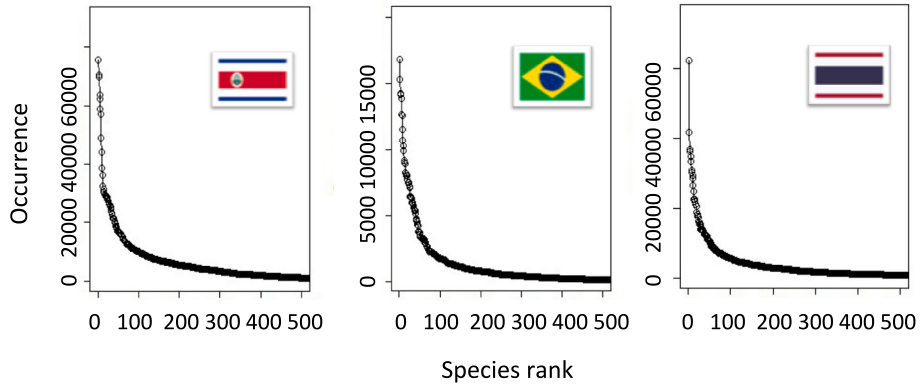


Fig. 5. Rank-abundance graphs for the three selected countries, Costa Rica (left), Brazil (central) and Thailand (right). Flag images are from https://flagpedia.net.

comparing eBird datasets for the three selected countries, we found that Benford's fit was lowest for Thailand and highest for Costa Rica. However, the rarefaction curves suggested that it would be unlikely to obtain more species with more individuals sampled for Thailand and Brazil, indicating that Benford's law is not ideal to check survey completeness and is somewhat sensitive to sample size, i.e., does not perform well for small datasets. Therefore, we confirm that Benford's conformity values are dependent on sample size, in our case, species richness matters for occurrence datasets, as suggested by Nigrini (2012) and others for non-biological datasets. While the Benford test is generally recommended for datasets with over 1000 records (to obtain numbers with at least four digits), in case of fewer digits, there is only a slightly larger bias in favour of the lower digits (Nigrini, 2012). Benford's law theorists warn that if smaller datasets are analysed, the deviation from the Benford proportions could be higher and the first digit test is generally recommended for small datasets, which are often the case for biological data. For instance, in our eBird dataset, only 11 countries had over 1000 species and 39 countries had only 10–99 observations.

Hill numbers and other methods can inform us about biodiversity, while accounting for coverage and relative abundance distribution (Hill, 1995; Chao et al., 2014). On the other hand, Benford's law can provide additional information on data heterogeneity or the evenness of the sample. Like Hill numbers, Benford fit can also be somewhat sensitive to the number of individuals sampled, but unlike Hill numbers, it does not explicitly measure species richness, but rather indicates the reliability of a dataset for studying relative abundance or occupancy of the species within a community. Similarly, we found that gambin fit (Ugland et al., 2007) improved with increasing coverage and the smallest adjustment (lowest AIC_c value) was seen in the only dataset categorised as non-conforming to Benford's law, i.e., the Thailand eBird dataset with no rare species. The fact that including rare and very rare species in the subsets increased conformity to Benford's law can be seen as a virtue of

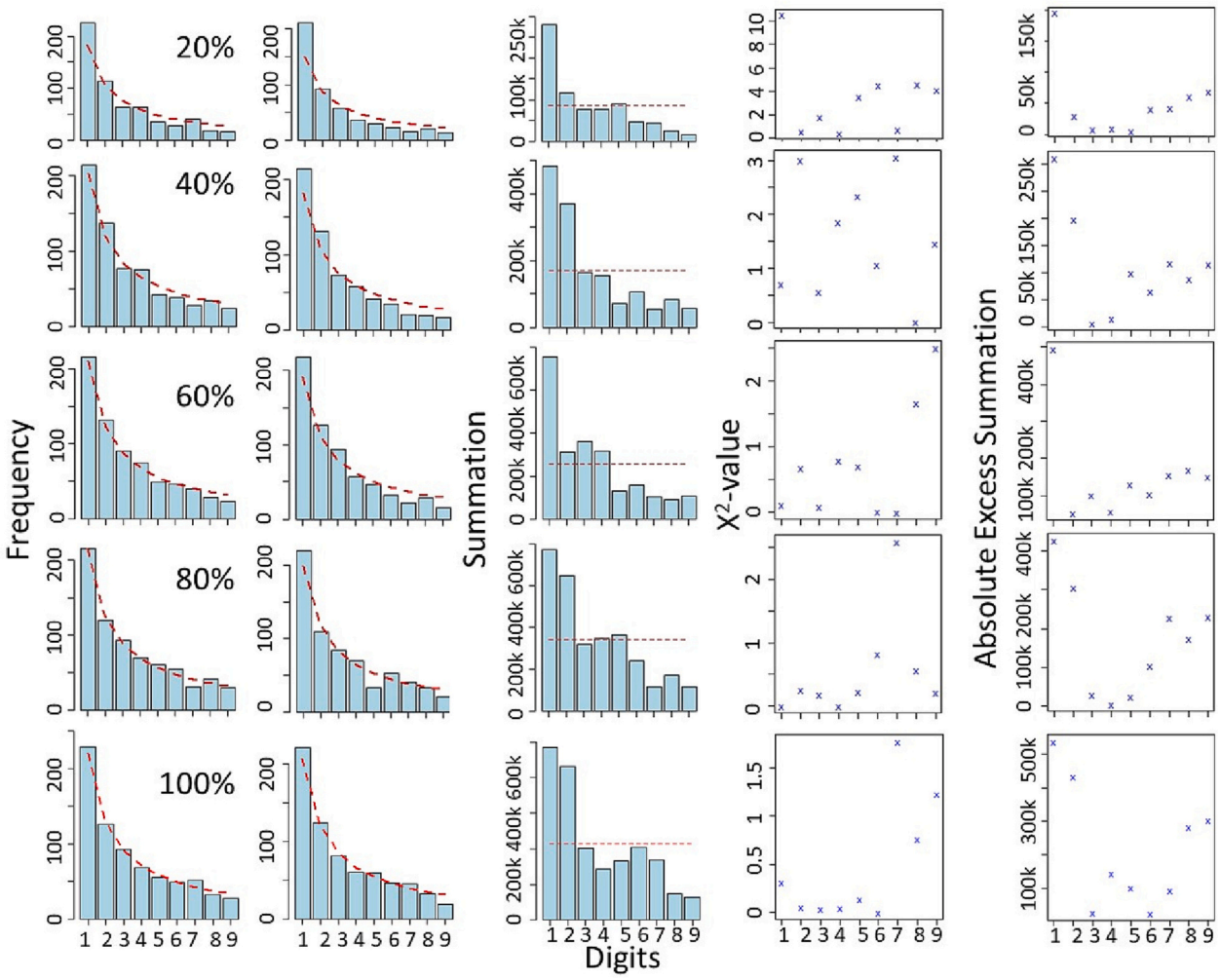


Fig. 6. Benford characteristics for subsets of the Costa Rica eBird data. In the columns from left to right: digit distribution for first order test, digit distribution for second order test, summation distribution by digits, Chi-square difference and Summation difference. In the rows from top to bottom: 20–40–60–80 and 100 % of the data. Plots in the last row are the same as plots in the same row of Fig. 5.

Table 4
Parameters for the decision to evaluate the conformity to Benford's law for 20 %, 40 %, 60 %, 80 % and 100 % subsets of eBird data from Costa Rica, Brazil and Thailand and GBIF datasets for spiders, frogs, and plants. N: number of observations used, N2: number of observations used for second order, $\mu \pm \sigma^2$: Mantissa mean and variance values, χ^2: Pearson's χ^2 statistic and associated p-value (in parentheses), L2: Mantissa arc test statistic and associated p-value (in parentheses), MAD: Mean Absolute Deviation, NC: non-conformity, MAC: marginally acceptable conformity, AC: acceptable conformity, CC: close conformity.

Dataset	N	N2	$\mu \pm \sigma^2$	χ^2	L2	MAD
20 % Costa Rica eBird	605	499	0.443 ± 0.082	30.597 (0.0002)	0.0103 (0.0223)	0.0334 NC
40 % Costa Rica eBird	671	606	0.475 ± 0.079	14.063 (0.0801)	0.0056 (0.0231)	0.0141 MAC
60 % Costa Rica eBird	698	637	0.481 ± 0.080	6.5671 (0.584)	0.0011 (0.4738)	0.0077 AC
80 % Costa Rica eBird	713	663	0.501 ± 0.082	4.8273 (0.7759)	0.0002 (0.8877)	0.0058 CC
100 % Costa Rica eBird	875	755	0.489 ± 0.088	3.1601 (0.9239)	0.0014 (0.2815)	0.0050 CC
Brazil eBird	1130	846	0.493 ± 0.082	6.6414 (0.5758)	0.0006 (0.3646)	0.0058 MAC
Thailand eBird	906	531	0.484 ± 0.089	0.001 (0.2428)	0.0016 (0.2428)	0.00922 NC

this method, as rare species often represent a problem for fitting SAD (Magurran and Henderson, 2003). The gambin fit proved to be the best model of the SAD of real communities and Benford conformity showed a similar pattern to its results, suggesting that testing Benford fit is similar to testing for a log-normal distribution with an adjustment for a long-tail caused by the presence of rare species.

The results of Benford fit also diverged from traditional measures of completeness for the non-avian datasets. The highest conformity to Benford's law was achieved for plants, possibly as this dataset had the larger sample size and higher species representativeness than frogs or spiders. While the rarefaction curve for frogs reached an asymptote, this dataset did not fit the distribution predicted by Benford's law. These comparisons also suggest that testing for Benford conformity is not a suitable indicator of the completeness of a biodiversity survey, although we know that many frog species are absent from this dataset.

Biodiversity datasets are often incomplete, especially those from tropical areas or consisting of hyperdiverse taxa, and in these cases, true species richness or species occurrences are underestimated (Colwell and Coddington, 1994). However, the completeness of a dataset does not necessarily indicate its usefulness, as there are many ecological and conservation questions that can be answered from datasets that are not necessarily complete (Wilson et al., 2005; Grantham et al., 2009). The question is to know when the dataset is reliable. While in general we suggest that high conformity is correlated with the representativeness of the sample based on species abundance heterogeneity, we did not

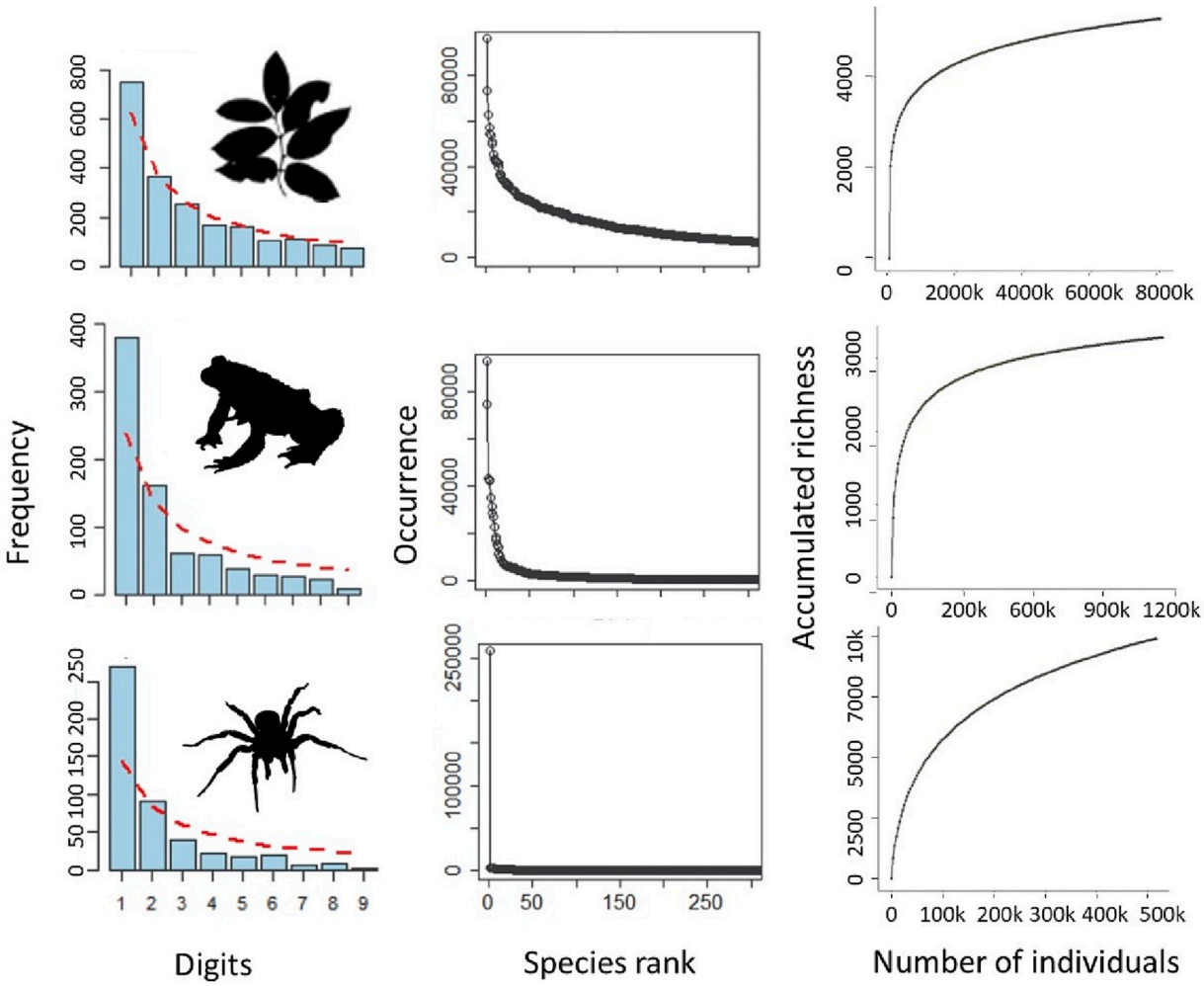


Fig. 7. Benford fit (left column), rarefaction (central column) and species accumulation curves as a function of the number of individuals for three GBIF datasets and the Plants of the Parisian Basin (marginal Benford conformity – top row) and frogs of the Southern Hemisphere (non-conformity – central row) and spiders of the Southern Hemisphere (non-conformity – bottom row). PhyloPic under Creative Commons licenses CC0 1.0 and CC BY 3.0.

explicitly test if a better fit to Benford's law actually means 'better' data, and we acknowledge that more data alone do not necessarily translate to effective biodiversity knowledge (Oliver et al., 2021). Even within a given country, the data will continue to have gaps and redundancies, potentially limiting our ability to understand biodiversity fully in that region. Nevertheless, our intent was to demonstrate the potential for using Benford's law in the age of big data in ecology and evolution. Hence, we suggest that Benford's law provides a complementary approach to identify reliable datasets for inferences on the relative abundances of species. We do not intend Benford conformity to replace traditional data quality checking approaches, but give more evidence to their usefulness.

As analyses look to answer large ecological and conservation research questions, an important first question is identifying 'good enough' data. Often, regions with little to no data need to be discarded for practical reasons (Oliver et al., 2021), and the conformity to Benford's law may provide a quick method to assess which regions have relatively good quality data as well as temporal progress towards increased data representativeness. Higher local participation in citizen science initiatives can help improve our ability to effectively monitor biodiversity in the future (Pocock et al., 2018). As these big datasets continue to increase, each country is increasingly responsible for monitoring its biodiversity, and metrics such as Benford's law may prove useful for countries to track their progress in closing biodiversity data gaps and species abundances representativeness. Therefore, we suggest that along with the more traditional methods checking the fit to Benford's law can also be useful, particularly for large-scale occurrence data. In conclusion, Benford conformity test is a way to check data quality based on the heterogeneity of species abundance, because it fits well to log-normal distribution, especially when longer tails of rare species are included. We suggest its use as an initial screening process to access the quality of citizen-science data and other large-scale biodiversity datasets to indicate their potential reliability.

Thresholds of acceptability or conformity to Benford's law vary depending on sample size (i.e., the number of species), the traits of these species (e.g., their rarity or detectability that will determine the number of observations) and also on the species abundance distribution. Although we present the results based on categorical responses (e.g., close and acceptable conformity), the use of the MAD score as a continuous measure may prove useful to track biodiversity data into the future, especially to account for the expected SAD. Future work should focus on (1) understanding the applicability of using Benford's law across different spatial scales for data collected using the same method; (2) repeating for different large-scale biodiversity datasets across different taxonomic groups, both structured and unstructured, collected using different methods; and (3) testing whether high conformity to Benford's law matches 'better' knowledge about biodiversity in a given region.

In conclusion, our analyses highlight the potential of Benford's law to be used as a fast and efficient first-pass complementary method to

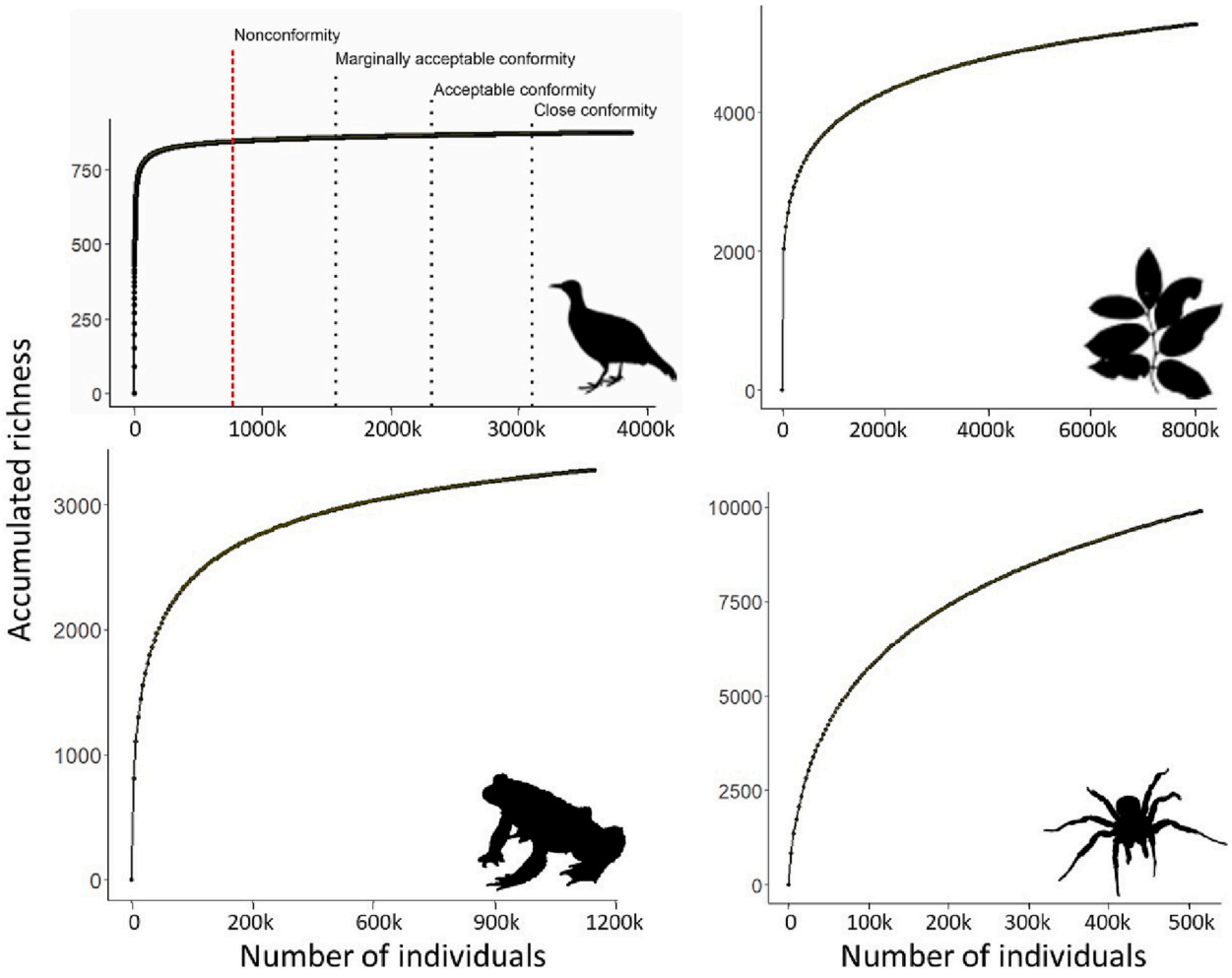


Fig. 8. Accumulated species richness values for A) eBird data for Costa Rica, showing number of individuals with different levels of Benford conformity, B) Plants of the Parisian Basin from GBIF (marginal Benford conformity) and C) frogs and D) spiders of the Southern Hemisphere from GBIF (both non-conformity). PhyloPic under Creative Commons licenses CC0 1.0 and CC BY 3.0.

traditional ways to assess the 'reliability' of large-scale biodiversity datasets, including those collected by citizen scientists with regard to the distribution of abundance among species.

CRediT authorship contribution statement

JKS conceived the ideas, designed methodology, and led the writing of the manuscript. LRF conceived the ideas, analysed the data, and reviewed the drafts. CTC designed methodology, collected, organized and analysed the data, as well as reviewed the drafts. All authors gave final approval for publication.

Declaration of competing interest

The authors declare no conflict of interest.

Data availability

eBird data are freely available from www.ebird.org and the completeness scores are available from La Sorte and Someveille (2019). GBIF datasets are available at https://doi.org/10.15468/dl.2da96q - plants of the Parisian Basin (Flore du Bassin Parisien), https://doi.org/10.15468/dl.g48yd6 - anurans of the Southern Hemisphere, and https://doi.org/10.15468/dl.6wfshp - the spiders of the Southern Hemisphere. The code to reproduce our analyses is provided in an archived Zenodo repository at https://doi.org/10.5281/zenodo.7470188.

Acknowledgements

LRF received a fellowship from the Coordination for the Improvement of Higher Education Personnel (CAPES - Finance Code 001). CTC was supported by a Marie Skłodowska-Curie Individual Fellowship (No 891052). The authors are indebted to the two anonymous reviewers and the associate editor for their insightful comments that have greatly improved the article.

References

Bakker, V.J., Baum, J.K., Brodie, J.F., Salomon, A.K., Dickson, B.G., Gibbs, H.K., Jensen, O.P., McIntyre, P.B., 2010. The changing landscape of conservation science funding in the United States. Conserv. Lett. 3, 435–444.
Beck, J., Schwanghart, W., 2010. Comparing measures of species diversity from incomplete inventories: an update. Methods Ecol. Evol. 1, 38–44.
Benford, F., 1938. The law of anomalous numbers. Proc. Am. Philos. Soc. 78, 551–572.
Burgess, H.K., DeBey, L.B., Froehlich, H.E., Schmidt, N., Theobald, E.J., Ettinger, A.K., HilleRisLambers, J., Tewksbury, J., Parrish, J.K., 2017. The science of citizen science: exploring barriers to use as a primary research tool. Biol. Conserv. 208, 113–120.
Callaghan, C.T., Poore, A.G.B., Mesaglio, T., Moles, A.T., Nakagawa, S., Roberts, C., Rowley, J.J.L., Vergés, A., Wilshire, J.H., Cornwell, W.K., 2021. Three Frontiers for the future of biodiversity research using citizen science data. Bioscience 71, 55–63.

Campos, L., Salvo, A.E., Flores-Moya, A., 2016. Natural taxonomic categories of angiosperms obey Benford's law, but artificial ones do not. Syst. Biodivers. 14, 431–440.
CBD, 2022. In: Monitoring Framework for the Kunming-Montreal Global Biodiversity Framework, pp. 1–26. https://www.cbd.int/doc/decisions/cop-15/cop-15-dec-05-en.pdf.
Ceballos, G., Ehrlich, P.R., Dirzo, R., 2017. Biological annihilation via the ongoing sixth mass extinction signaled by vertebrate population losses and declines. PNAS 114, E6089–E6096.
Cerri, J., 2018. In: A Fish Rots From the Head Down: How to Use the Leading Digits of Ecological Data to Detect Their Falsification. bioRxiv, p. 368951.
Chandler, M., See, L., Copas, K., Bonde, A.M.Z., López, B.C., Danielsen, F., Legind, J.K., Masinde, S., Miller-Rushing, A.J., Newman, G., Rosemartin, A., Turak, E., 2017. Contribution of citizen science towards international biodiversity monitoring. Biol. Conserv. 213, 280–294.
Chao, A., Gotelli, N.J., Hsieh, T.C., Sander, E.L., Ma, K.H., Colwell, R.K., Ellison, A.M., 2014. Rarefaction and extrapolation with hill numbers: a framework for sampling and estimation in species diversity studies. Ecol. Monogr. 84, 45–67.
Chao, A., Jost, L., 2012. Coverage-based rarefaction and extrapolation: standardizing samples by completeness rather than size. Ecology 93, 2533–2547.
Chao, A., Kubota, Y., Zelený, D.Chiu, Li, C.-F., Kusumoto, B., Yasuhara, M., Thorn, S., Wei, C.-L., Costello, M.J., Colwell, R.K., CH, 2020. Quantifying sample completeness and comparingdiversities among assemblages. Ecological Res. 35, 292–314.
Chiarucci, A., Bacaro, G., Rocchini, D., Ricotta, C., Palmer, M.W., Scheiner, S.M., 2009. Spatially constrained rarefaction: incorporating the autocorrelated structure of biological communities into sample-based rarefaction. Community Ecology 10, 209–214.
Cinelli, C., 2014. Benford.analysis: Benford Analysis for data validation and forensic analytics. R package version 0.1.5. https://www.rdocumentation.org/packages/benford.analysis/versions/0.1.5.
Colwell, R.K., Coddington, J.A., 1994. Estimating terrestrial biodiversity through extrapolation. Philos. Trans. R. Soc. B 345, 101–118.
Cooper, C.B., Shirk, J., Zuckerberg, B., 2014. The invisible prevalence of citizen science in global research: migratory birds and climate change. PLoS ONE 9, e106508.
Costas, E., Lopez-Rodas, V., Toro, F.J., Flores-Moya, A., 2008. The number of cells in colonies of the cyanobacterium Microcystis aeruginosa satisfies Benford's law. Aquat. Bot. 89, 341–343.
Cox, K.D., Black, M.J., Filip, N., Miller, M.R., Mohns, K., Mortimor, J., Freitas, Thaise R., Loerzer, R.G., Gerwing, T.G., Juanes, F., Dudas, S.E., 2017. Community assessment techniques and the implications for rarefaction and extrapolation with hill numbers. Ecol. Evol. 7, 11213–11226.
Docampo, S., del Mar Trigo, M., Aira, M.J., Cabezudo, B., Lores-Moya, A., 2009. Benford's law applied to aerobiological data and its potential as a quality control tool. Aerobiologia 25, 275.
Friar, J.L., Goldman, T., Perez-Mercader, J., 2012. Genome sizes and the benford distribution. PLoS ONE 7, e36624.
Gotelli, N.J., Colwell, R.K., 2011. Estimating species richness. In: Magurran, A.E., McGill, B.J. (Eds.), Biological diversity: frontiers in measurement and assessment. Oxford University Press, New York, New York, USA, pp. 39–54.
Grantham, H.S., Wilson, K.A., Moilanen, A., Rebelo, T.G., Possingham, H.P., 2009. Delaying conservation actions for improved knowledge: how long should we wait? Ecol. Lett. 12, 293–301.
Harrison, P.J., Buckland, S.T., Yuan, Y., Elston, D.A., Brewer, M.J., Johnston, A., Pearce-Higgins, J.W., 2014. Assessing trends in biodiversity over space and time using the example of British breeding birds. J. Appl. Ecol. 51, 1650–1660.
Hill, T.P., 1995. A statistical derivation of the significant-digit law. Stat. Sci. 10, 354–363.
Horns, J.J., Adler, F.R., Şekercioğlu, Ç.H., 2018. Using opportunistic citizen science data to estimate avian population trends. Biol. Conserv. 221, 151–159.
Hortal, J., Jiménez-Valverde, A., Gómez, J.F., Lobo, J.M., Baselga, A., 2008. Historical bias in biodiversity inventories affects the observed environmental niche of the species. Oikos 117 (6), 847–858.
Horton, J., Kumar, D.K., Wood, A., 2020. Detecting academic fraud using Benford law: the case of professor James Hunton. Res. Policy 49, 104084.
Jetz, W., McGeoch, M.A., Guralnick, R., Ferrier, S., Beck, J., Costello, M.J., Fernandez, M., Geller, G.N., Keil, P., Merow, C., Meyer, C., Muller-Karger, F.E., Pereira, H.M., Regan, E.C., Schmeller, D.S., Turak, E., 2019. Essential biodiversity variables for mapping and monitoring species populations. Nat. Ecol. Evol. 3, 539–551.
Klimek, P., Jiménez, R., Hidalgo, M., Hinteregger, A., Thurner, S., 2018. Forensic analysis of turkish elections in 2017–2018. PLoS ONE 13, e0204975.
La Sorte, F.A., Somveille, M., 2019. Survey completeness of a global citizen-science database of bird occurrence. Ecography 42, 1–10.
Leong, M., Dunn, R.R., Trautwein, M.D., 2018. Biodiversity and socioeconomics in the city: a review of the luxury effect. Biol. Lett. 14, 20180082.
Lobo, J.M., 2008. Database records as a surrogate for sampling effort provide higher species richness estimations. Biodivers. Conserv. 17 (4), 873–881.
Lobo, J.M., Hortal, J., Yela, J.L., Millán, A., Sánchez-Fernández, D., García-Roselló, E., González-Dacosta, J., Heine, J., González-Vilas, L., Guisande, C., 2018. KnowBR: an application to map the geographical variation of survey effort and identify well-surveyed areas from biodiversity databases. Ecol. Indic. 91, 241–248.
McGill, B.J., Etienne, R.S., Gray, J.S., Alonso, D., Anderson, M.J., Benecha, H.K., Dornelas, M., Enquist, B.J., Green, J.L., He, F., Hurlbert, A.H., 2007. Species abundance distributions: moving beyond single prediction theories to integration within an ecological framework. Ecol. Lett. 10 (10), 995–1015.
Magurran, A.E., Henderson, P.A., 2003. Explaining the excess of rare species in natural species abundance distributions. Nature 422, 714–716.
Magurran, A.E., 2004. Measuring Biological Diversity. Blackwell, Oxford, UK.
Magurran, A.E., Baillie, S.R., Buckland, S.T., Dick, J.M., Elston, D.A., Scott, E.M., Watt, A. D., 2010. Long-term datasets in biodiversity research and monitoring: assessing change in ecological communities through time. Retrieved from Trends Ecol. Evol. 25, 574–582. http://www.iucnredlist.org/technical-documents/categories-and-criteria.
Marquet, P.A., Abades, S.R., Labra, F.A., 2007. Biodiversity power laws. In: Storch, D., Marquet, P.A., Brown, J.H. (Eds.), Scaling Biodiversity Cambridge. Cambridge University Press, U. K, pp. 441–461.
Matthews, T.J., Borregaard, M.K., Ugland, K.I., Borges, P.A.V., Rigal, F., Cardoso, P., Whittaker, R.J., 2014. The gambin model provides a superior fit to species abundance distributions with a single free parameter: evidence, implementation and interpretation. Ecography 37, 1002–1011.
May, F., Gerstner, K., McGlinn, D.J., Xiao, X., Chase, J.M., 2018. Mobsim: an r package for the simulation and measurement of biodiversity across spatial scales. Methods Ecol. Evol. 9, 1401–1408.
Melo, A.S., Pereira, R.A.S., Santos, A.J., Shepherd, G.J., Machado, G., Medeiros, H.F., Sawaya, R.J., 2003. Comparing species richness among assemblages using sample units: why not use extrapolation methods to standardize different sample sizes? Oikos 101 (2), 398–410.
McGill, B.J., Etienne, R.S., Gray, J.S., Alonso, D., Anderson, M.J., Benecha, H.K., Dornelas, M., Enquist, B.J., Green, J.L., He, F.L., Hurlbert, A.H., Magurran, A.E., Marquet, P.A., Maurer, B.A., Ostling, A., Soykan, C.U., Ugland, K.I., White, E.P., 2007. Species abundance distributions: moving beyond single prediction theories to integration within an ecological framework. Ecol. Lett. 10, 995–1015.
Meyer, C., Kreft, H., Guralnick, R., Jetz, W., 2015. Global priorities for an effective information basis of biodiversity distributions. Nat. Commun. 6, 8221.
Michener, W.K., 2006. Meta-information concepts for ecological data management. Ecol. Informatics 1, 3–7.
Navarro, L.M., Fernández, N., Guerra, C., Guralnick, R., Kissling, W.D., Londoño, M.C., Muller-Karger, F.E., Turak, E., Balvanera, P., Costello, M.J., Delavaud, A., El Serafy, G., Ferrier, S., Geijzendorffer, I., Geller, G.N., Jetz, W., Kim, E.-S., Kim, H., Martin, C.S., McGeoch, M.A., Mwampamba, T.H., Nel, J.L., Nicholson, E., Pettorelli, N., Schaepman, M.E., Skidmore, A.K., Sousa Pinto, I., Vergara, S., Vihervaara, P., Xu, H., Yahara, T., Gill, M., Pereira, H.M., 2017. Monitoring biodiversity change through effective global coordination. Curr. Opin. Environ. Sustain. 29, 158–169.
Neate-Clegg, M.H.C., Horns, J.J., Adler, F.R., Kemahlı Aytekin, M.Ç., Şekercioğlu, Ç.H., 2020. Monitoring the world's bird populations with community science data. Biol. Conserv. 248, 108653.
Newcomb, S., 1881. Note on the frequency of use of the different digits in natural numbers. Am. J. Math. 4, 39–40.
Nigrini, M.J., 2012. Benford's Law: Applications for Forensic Accounting, Auditing, and Fraud Detection. John Wiley & Sons.
Oksanen, J., Blanchet, F.G., Friendly, M., Kindt, R., Legendre, P., McGlinn, D., Wagner, H., 2020. vegan: Community Ecology Package. R package version 2.5-7. https://CRAN.R-project.org/package=vegan.
Oliver, R.Y., Meyer, C., Ranipeta, A., Winner, K., Jetz, W., 2021. Global and national trends, gaps, and opportunities in documenting and monitoring species distributions. PLoS Biol. 19 (8), e3001336.
Özkundakci, D., Pingram, M.A., 2019. Nature favours "one" as the leading digit in phytoplankton abundance data. Limnologica 78, 125707.
Pavlacky, D.C., Lukacs, P.M., Blakesley, J.A., Skorkowsky, R.C., Klute, D.S., Hahn, B.A., Dreitz, V.J., George, T.L., Hanni, D.J., 2017. A statistically rigorous sampling design to integrate avian monitoring and management within bird conservation regions. PLoS ONE 12, e0185924.
Perazzoni, F., Bacelar-Nicolau, P., Painho, M., 2020. Geointelligence against illegal deforestation and timber laundering in the Brazilian Amazon. ISPRS Int. J. Geo Inf. 9, 398.
Pereira, H.M., Ferrier, S., Walters, M., Geller, G.N., Jongman, R.H.G., Scholes, R.J., Bruford, M.W., Brummitt, N., Butchart, S.H.M., Cardoso, A.C., Coops, N.C., Dulloo, E., Faith, D.P., Freyhof, J., Gregory, R.D., Heip, C., Höft, R., Hurtt, G.C., Jetz, W., Karp, D.S., McGeoch, M.A., Obura, D.O., Onoda, Y., Pettorelli, N., Reyers, B., Sayre, R., Scharlemann, J.P.W., Stuart, S.N., Turak, E., Walpole, M., Wegmann, M., 2013. Essential biodiversity variables. Science 339, 277–278.
Pocock, M.J.O., Chandler, M., Bonney, R., Thornhill, I., Albin, A., August, T.A., Bachman, S., Brown, P.M.J., Gasparini Fernandes Cunha, D., Grez, A., Jackson, C., Peters, M., Romer Rabarijaonkk, N., Roy, H.E., Zaviezo, T., Danielsen, F., 2018. A vision for global biodiversity monitoring with citizen science. Advances in Ecological Research 59, 169–223.
R Core Development Team, 2020. R: A Language and Environment for Statistical Computing. (Foundation for Statistical Computing, Vienna, Austria.
Roswell, M., Dushoff, J., Winfree, R., 2021. A conceptual guide to measuring species diversity. Oikos 130, 321–338.
Sambridge, M., Jackson, A., 2020. National COVID numbers — Benford's law looks for errors. Nature 581, 384.
Sullivan, B.L., Aycrigg, J.L., Barry, J.H., Bonney, R.E., Bruns, N., Cooper, C.B., Damoulas, T., Dhondt, A.A., Dietterich, T., Farnsworth, A., Fink, D., Fitzpatrick, J. W., Fredericks, T., Gerbracht, J., Gomes, C., Hochachka, W.M., Iliff, M.J., Lagoze, C., La Sorte, F.A., Merrifield, M., Morris, W.K., Phillips, T.B., Reynolds, M., Rodewald, A.D., Rosenberg, K.V., Trautmann, N.M., Wiggins, A., Winkler, D.W., Wong, W.-K., Wood, C.L., Yu, J., Kelling, S., 2014. The eBird enterprise: An integrated approach to development and application of citizen science. Biol. Conserv. 169, 31–40.

Sullivan, B.L., Phillips, T., Dayer, A.A., Wood, C.L., Farnsworth, A., Iliff, M.J., Davies, I.J., Wiggins, A., Fink, D., Hochachka, W.M., Rodewald, A.D., Rosenberg, K.V., Bonney, R., Kelling, S., 2017. Using open access observational data for conservation action: a case study for birds. Biol. Conserv. 208, 5–14.

Szabo, J.K., Baxter, P.W.J., Vesk, P.A., Possingham, H.P., 2011. Paying the extinction debt: woodland birds in the mount lofty ranges, South Australia. Emu 111, 59–70.

Szabo, J.K., Fuller, R.A., Possingham, H.P., 2012. A comparison of estimates of relative abundance from a weakly structured mass-participation bird atlas survey and a robustly designed monitoring scheme. Ibis 154, 468–479.

Theobald, E.J., Ettinger, A.K., Burgess, H.K., DeBey, L.B., Schmidt, N.R., Froehlich, H.E., Wagner, C., HilleRisLambers, J., Tewksbury, J., Harsch, M.A., Parrish, J.K., 2015. Global change and local solutions: tapping the unrealized potential of citizen science for biodiversity research. Biol. Conserv. 181, 236–244.

Troia, M.J., McManamay, R.A., 2016. Filling in the GAPS: evaluating completeness and coverage of open-access biodiversity databases in the United States. Ecol. Evol. 6 (14), 4654–4669.

Troudet, J., Grandcolas, P., Blin, A., Vignes-Lebbe, R., Legendre, F., 2017. Taxonomic bias in biodiversity data and societal preferences. Sci. Rep. 7, 9132.

Tulloch, A., Szabo, J.K., 2012. A behavioural ecology approach to understand volunteer surveying for citizen science datasets'. Emu 112, 313–325.

Ugland, K.I., Lambshead, J.D., McGill, B.J., Gray, J.S., O'Dea, N., Ladle, R.J., Whittaker, R.J., 2007. Modelling dimensionality in species abundance distributions: description and evaluation of the Gambin model. Evol. Ecol. Res. 9, 313–324.

Verberk, W.C.E.P., van der Velde, G., Esselink, H., 2010. Explaining abundance–occupancy relationships in specialists and generalists: a case study on aquatic macroinvertebrates in standing waters. J. Anim. Ecol. 79, 589–601.

Ward, D.F., 2014. Understanding sampling and taxonomic biases recorded by citizen scientists. J. Insect Conserv. 18, 753–756.

Warren II, R.J., Skelly, D.K., Schmitz, O.J., Bradford, M.A., 2011. Universal ecological patterns in college basketball communities. PLoS ONE 6 (3), e17342.

Wilson, K.A., Westphal, M.I., Possingham, H.P., Elith, J., 2005. Sensitivity of conservation planning to different approaches to using predicted species distribution data. Biol. Conserv. 22 (1), 99–112.

Zizka, A., Rydén, O., Edler, D., Klein, J., Perrigo, A., Silvestro, D., Jagers, S.C., Lindberg, S.I., Antonelli, A., 2021. Bio-dem, a tool to explore the relationship between biodiversity data availability and socio-political conditions in time and space. J. Biogeogr. 48, 2715–2726.