Received: 23 September 2019 | Revised: 3 February 2020 | Accepted: 4 February 2020

DOI: 10.1111/ddi.13049

BIODIVERSITY RESEARCH

Diversity and Distributions WILEY

Factors influencing the precision of species richness estimation in Japanese vascular plants

Werner Ulrich[1] | Buntarou Kusumoto[2,3] | Simone Fattorini[4] | Yasuhiro Kubota[2,5]

[1]Department of Ecology and Biogeography, Nicolaus Copernicus University, Toruń, Poland

[2]Faculty of Science, University of the Ryukyus, Nishihara, Japan

[3]Royal Botanic Gardens, Kew, Richmond, UK

[4]Department of Life, Health & Environmental Sciences, University of L'Aquila, L'Aquila, Italy

[5]Marine and Terrestrial Field Ecology, Tropical Biosphere Research Center, University of the Ryukyus, Nishihara, Japan

Correspondence
Werner Ulrich, Department of Ecology and Biogeography, Nicolaus Copernicus University, Toruń, Poland.
Email: ulrichw@umk.pl

Funding information
Japan Society for the Promotion of Science, Grant/Award Number: 4-1501 and 4-1802; Narodowe Centrum Nauki, Grant/Award Number: UMO-2017/27/B/NZ8/00316

Editor: Raimundo Real

Abstract

Aim: Estimating species richness from a series of samples is an important and widely debated issue in ecology and biodiversity conservation. Numerous tests of respective richness estimators gave insights into the precision, the limitations and the pitfalls of richness forecasting. However, few benchmark tests used almost complete empiric census data obtained at those spatial scales where richness estimation is most useful for conservation management.

Location: Japan.

Methods: We use an extraordinary dataset on the spatial distribution of Japanese plants containing complete information on the occurrence of each Japanese plant species at the 10 × 10 km^2 grid cell level. We link the estimates of four estimators representing different theoretical approaches, Chao2, rarefaction, species–area relationships (SAR) and species abundance distributions (SAD), to environmental data using a fully nested sampling design.

Results: Chao2 and rarefaction behaved very similar in all tests and significantly underestimated true richness below 40% sampling fraction. SAR and SAD were less precise than Chao2 and rarefaction at higher sampling fraction but also less affected by low sample size. In general, SAD provided robust estimates over the whole range of sampling fraction and 67.4% of estimates ranged within the 10% error level. Higher species spatial turnover increased and high evenness in occurrence decreased the precision of the SAD estimator. Precision of the four estimators was largely unaffected by environmental variability but increased with increasing latitude.

Main conclusions: Our results strongly indicate that the pattern of Japanese plant species spatial distribution is sufficiently scale invariant for richness estimators to provide precise forecasting results at the country level. The simplest process to generate such a spatial distribution is ecological drift.

KEYWORDS

biodiversity, Japanese vegetation, species abundance distribution, species richness estimation, species–area relationship

1 | INTRODUCTION

Estimating species richness from series of samples is an important and widely debated issue in ecology and biodiversity conservation (Chao & Chiu, 2016; Hortal, Borges, & Gaspar, 2006; Kunin et al., 2018; Magurran, 2004; Magurran & McGill, 2011). Numerous parametric or nonparametric richness estimators, based on either abundance or incidence data, have been developed and tested (reviewed in Chao & Chiu, 2015, 2016; Hortal et al., 2006; Walther & Moore, 2005).

Estimators of species richness can be classified into four groups (Chao & Chiu, 2016). First, nonparametric asymptotic estimators are often based on random sampling assuming a Poisson species spatial distribution (Chao, Colwell, Lin, & Gotelli, 2009). Second, parametric non-asymptotic rarefaction estimators extrapolate species richness of standardized samples within a common finite sample size towards larger sample sizes (Colwell, Chao, & Gotelli, 2012). Third, species accumulation curves and species distribution models extrapolate richness data for observed samples or areas towards the focal sample size or area (Scheiner, 2003; Shen & He, 2008; Ugland, Gray, & Ellingsen, 2003). These approaches also include spatially explicit ecological drift (Hubbell, 2001) and maximum entropy models (Harte & Newman, 2014). Fourth, Dewdney (1998) and Ulrich and Ollik (2005) have advocated the use of species–abundance or species–occupancy distributions (below abbreviated as *SAD*) obtained from finite samples to estimate richness of larger samples.

Tests of richness estimators were mainly based on simulation studies where samples were taken from idealized communities spread over many sample sites (e.g. Chao & Chiu, 2015). These tests provided insights into the behaviour of estimators, returned recommendations for estimator choice (Kunin et al., 2018; Walther & Moore, 2005) and demonstrated the comparable behaviour of different estimators that are based on the same theoretical assumptions (Chao & Chiu, 2015; Gwinn, Allen, Bonvechio, Hoyer, & Beesley, 2016). They also pointed to possible pitfalls. Particularly, Gwinn et al. (2016) demonstrated the sensitivity of all available nonparametric estimators to sample size and the shape of the underlying specie abundance distributions.

Tests of estimators with empirical data compared the species richness in a fully censused survey (often an area) with estimates obtained from series of samples (e.g. Palmer, 1990, 1991; Chiarucci, Enright, Perry, Miller, & Lamont, 2003; Herzog, Kessler, & Cahill, 2002; De Thoisy, Brosse, & Dubois, 2008; Walther & Moore, 2005). These tests, however, have been performed on small communities with low total species richness and their results cannot be extended in a straightforward way to systems with large spatial extent and with a large number of species. It is also not clear whether methods proposed for local scale sampling are also applicable at regional or even global scales (Fattorini, 2013). However, most environmental studies and also conservation planning use sample plots to predict ecological patterns at larger areas or even globally (Chao, Colwell, Gotelli, & Thorn, 2019; Chiarucci, Bacaro, & Scheiner, 2011). An exception is the recent comparative study by Kunin et al. (2018), who used fully censused plant records from Southern England to evaluate whether richness estimators based on extrapolation techniques (species–area relationships and species occupancy distributions) and species distribution modelling (maximum entropy and ecological drift) are able to forecast the true regional plant richness and the respective species accumulation with increasing space. However, all the methods tested by Kunin et al. (2018) required information on either small-scale species accumulation curves or species occupancy distributions that are not available in most ecological and conservation studies. Consequently, these latter studies still rely on traditional parametric and nonparametric estimators based on specific probability distributions (reviewed in Chao & Chiu, 2016). These estimators still have to be tested with large-scale empirical data.

Common richness estimators rely on the assumption that the distribution of abundances and the pattern of species spatial distribution are sufficiently scale invariant to meet the assumptions of the underlying model (Brose, Martinez, & Williams, 2003). For instance, fractal spatial species distributions would generate richness accumulation curves that follow a power function (Šizling & Storch, 2004), thus allowing its use for precise richness estimation. However, many studies revealed that patterns of species spatial aggregation (e.g. Kunin, 1998; Lennon, Koleff, Greenwood, & Gaston, 2001) and community evenness (e.g. Brose et al., 2003) change with increasing sample area. These changes appeared to be taxon and biome-specific, making it challenging to apply universal correctors for richness estimators.

Without knowledge of the scaling of species spatial distributions, extrapolations of species richness become increasingly uncertain, especially when based on very small sample sizes (Chiarucci et al., 2003). The scaling of the spatial distribution of species and their abundances is highly dependent on environmental factors (Mertes & Jetz, 2018). Consequently, we might expect that the performance of richness estimators is also influenced by these factors. A possible yet largely unexplored way to improve richness estimators would be to use environmental characteristics that determine community composition (defined by species identities) in a predictable way. Often, environmental variables are easier to assess than community patterns. However, respective studies on environmental constraints on richness estimation are largely missing. For instance, Beukema and Dekker (2012) found that species accumulation curves changed in shape along environmental gradients from near-shore to off-shore intertidal areas due to shifts in the relative abundances of rare and common species, allowing reliable richness estimates only for large aggregated samples.

Here, we try to fill this gap in our knowledge using the distributional data (presence–absence) of Japanese plant species as a model system. This dataset contains complete information on the occurrence of each Japanese plant species at the 10 × 10 km grid cell level (Kusumoto, Villalobos, Shiono, & Kubota, 2019). Thus, our data do not rely on samples, but on complete censuses. Using a fully nested sampling of these cells, we link the precision and the relative error of four common richness estimators to geographical position, climate and geographical variability, as well as to the pattern of species

spatial turnover and evenness. We ask whether and how (a) geographical position and altitude influence the accuracy of richness estimation, (b) the variability in climate variables, elevation and forest cover affects estimator performance and (c) evenness in occurrence and the spatial species turnover among cells influence estimation accuracy. For each of the four categories of estimators cited above (i.e. nonparametric asymptotic, parametric non-asymptotic rarefaction, species accumulation and SAD-based models), we selected that estimator known to perform best. We test them against our data to infer which estimator is best suited in a given large-scale, empirical situation.

2 | METHODS

2.1 | Study area

The East Asian islands mainly including the Japanese and Ryukyu archipelagos are located off the eastern coast of the Eurasian continent. The mean annual temperature ranges from −5.3 to 24.2°C. The annual precipitation ranges from 650 to 4,538 mm. Such a wide range of climate form diverse biomes across hemiboreal, temperate and subtropical vegetation.

2.2 | Species distribution data

Occurrence records for vascular plant species across Japan were compiled by searching the botanical literature on the flora of Japan (Kubota, Kusumoto, Shiono, & Tanaka, 2017; Kubota, Shiono, & Kusumoto, 2015). Most of the references were based on specimen records, local species checklists, expert species range maps and vegetation census records (phytosociological tables). We then georeferenced occurrences to latitudinal and longitudinal coordinates. Based on the resulting set of species occurrence data, we built a geographical distribution database for 5,614 species at the 10 × 10 km^2 grid cell level (4,852 cells). Therefore, our study is based on almost complete species census data at the grid level minimizing both commission (false presence) and omission (false absence) errors. Species names followed the Japanese Scientific Names Index (Yonekura & Kajita, 2003). In this analysis, exotic species including planted species were excluded from the dataset (Kusumoto et al., 2019). Kubota et al. (2015), Kubota et al. (2017) provided a more detailed description of how this dataset was created.

2.3 | Environmental data

For each of the 4,852 Japanese grid cells, we compiled an environmental dataset including temperature and precipitation seasonality (i.e. the difference between the annual maximum and minimum values), forest area (km^2), land and forest area, and the variability in forest cover and elevation within each grid cell. Temperature and precipitation data refer to the period between 1971 and 2000, and were extracted from the 1-km gridded data in the Mesh Climate Data 2000 (Okada, Iizumi, Nishimori, & Yokozawa, 2008).

We split the 4,852 grid cells into 46 windows of 100 geographically adjacent cells each (excluding cells that covered more than 50% sea area). We excluded another 252 cells that were either isolated (violating the nested design) or lacked plant records (inflating zero counts). We calculated for each of these windows four estimators described below based on 10, 20, 30, … 90, 99 cells (in a fully nested design) resulting in a total of 460 estimates.

2.4 | Statistical analysis

Here, we use one well-performing estimator of each of the estimator groups mentioned in the introduction: the Chao2 (Chao, 1987), rarefaction (Chao et al., 2009; Chiarucci, Bacaro, Rocchini, & Fattorini, 2008), power function SAR (Kunin et al., 2018; Rosenzweig, 1995) and the SAD estimator of Ulrich and Ollik (2005).

The Chao2 estimator is given by.

$$S_{\text{Chao}} = S_{\text{pobs}} + \frac{N}{N-1}\frac{Q_1^2}{2Q_2} \mid Q_2 > 0 \quad (1)$$

$$S_{\text{Chao}} = S_{\text{pobs}} + \frac{N}{N-1}\frac{Q_1\,(Q_1-1)}{2} \mid Q_2 = 0 \quad (2)$$

where S_{obs} is the observed richness, N the number of sample units, and Q_1 and Q_2 are the numbers of species that occurred exactly in one and in two of the sample units (Chao & Chiu, 2016).

The rarefaction estimate (Chao & Chiu, 2016; Chiarucci et al., 2008) comes from

$$S_{\text{RF}}(K) = S_{\text{obs}} + (\text{Chao2} - S_{\text{obs}})\left[1 - \left(1 - \frac{Q_1}{N(\text{Chao2} - S_{\text{obs}}) + Q_1}\right)^{(K-N)}\right] \quad (3)$$

where K is the total number of sample units (Chao & Chiu, 2016).

Among various species accumulation models, species–area relationships (SARs) are most often applied in ecology and biodiversity conservation. There is a long-standing debate whether and in which situation SARs are better described by the power function or other models, such as the logarithmic function (e.g. Connor & McCoy, 1979; Triantis, Guilhaumon, & Whittaker, 2012). Most work favoured power functions (Rosenzweig, 1995; Dengler et al., 2020). Both the power function and the logarithmic model might be able to provide upper and lower limits of richness, but the logarithmic function frequently underestimates richness when extrapolated to larger areas (Dengler, 2009). The estimate based on the power function species–area relationship (Rosenzweig, 1995) is given by

$$S_{\text{SAR}} = S_0 A_{\text{total}}^{z} \quad (4)$$

where S_0, the richness at unit area, and z, the slope value, are the parameters inferred from the observed range of area. Here, we extrapolated eq. 4 using its linearized version

$$\ln S_{\text{powSAR}} = \ln S_0 + z \ln A_{\text{total}} \quad (5)$$

Accordingly, extrapolations from the logarithmic SAR were based on

$$S_{\log \text{SAR}} = S_0 + z \ln A_{\text{total}} \quad (6)$$

where both estimates for the intercept S_0 and the slope z came from ordinary least squares regressions for the richness data within the observed range of areas.

Recent meta-analyses (Ulrich, Nakadai, Matthews, & Kubota, 2018; Ulrich, Ollik, & Ugland, 2010) and estimator comparisons (Kunin et al., 2018) have corroborated the view that the vast majority of observed species abundances distributions fall within boundaries set by the log-series (Fisher, Corbet, & Williams, 1943) and the lognormal distribution (Preston, 1948). When applied to a finite area both distributions allow for an assessment of total richness. Therefore, we used the method of Ulrich and Ollik (2005) to obtain an estimate of the upper limit of species richness from the observed distribution of species occurrences (cf. Kunin et al., 2018). Under the assumption that the observed species rank occurrence distribution follows a log-series, the upper limit of expected richness S_{LS} is given by

$$S_{\text{LS}} = \frac{\text{Int} + \ln(N_{\text{total}}) - \ln(N_{\text{obs}})}{-\text{slope}} \quad (7)$$

The lower limit of richness comes from the assumption that the SAD follows a lognormal distribution

$$S_{\text{LN}} = \frac{2\text{Int} + \ln(N_{\text{total}}) - 2\ln(N_{\text{obs}})}{-\text{slope}} \quad (8)$$

where *Int* is the intercept and *slope* the slope of the linear regression for the observed log-occurrence—species rank order distribution (SAD). N_{obs} and N_{total} are the observed number of sample plots used for constructing the SAD and the total number of plots, respectively.

We assessed the relative error (relative bias) and the precision from

$$\text{Relative bias} = \frac{S_{\text{est}} - S_{\text{total}}}{S_{\text{total}}} \quad (9)$$

and

$$\text{Precision} = \max\left(0; 1 - \left|\frac{S_{\text{est}} - S_{\text{total}}}{S_{\text{total}}}\right|\right) \quad (10)$$

where S_{est} and S_{total} are the estimated and total richness, respectively.

We assessed the dependence of relative bias on sampling fraction (the proportion of cells sampled) by the coefficient of correlation r_{bias} between relative bias and ln-transformed sampling fraction. Below, species coverage refers to the proportion of species sampled with respect to the total (the empirical) number in the focal window.

For each sample, we assessed the total area sampled, the proportions of area sampled (the sampling fraction) and of forest cover, and the averaged values of the aforementioned variability in forest cover, elevation, precipitation and temperature. Additionally, we calculated for each window the degree of species spatial segregation from the C-score (Stone & Roberts, 1990), which is a normalized count of the number of pairwise checkerboard occurrences summed over all species pairs. As raw metrics of co-occurrences are biased by matrix size and fill (Ulrich et al., 2018), we use a null model approach and compared observed scores with those obtained from an equiprobable reshuffling of the occurrences of each species among the cells of each window. This null model retained for each window the observed species composition but randomized the occurrences of each species among the samples. The motivation behind this random assumption was that there is no a priori reason to constrain the occurrences of a focal species towards certain cells within a window of generally similar climatic conditions. We used standardized effect sizes ($\text{SES} = \frac{\text{CS}_{\text{obs}} - \text{CS}_{\text{exp}}}{\sigma_{\text{exp}}}$; where CS_{obs} is the observed score and CS_{exp} and σ_{exp} are the mean and the standard deviation of the null distribution, respectively) to relate the degree of spatial segregation to the error in richness estimation. High values of CS_{obs} and SES point to significant species spatial turnover among cells. Finally, we calculated for each window and each sample the degree of evenness in the number of occurrences from $E = \frac{H}{\ln S_{\text{total}}}$; where H denotes the Shannon diversity and ln S_{total} the log-transformed total number of species in the focal window (sample).

We used a general linear model (GLM) to link the relative bias for each sample size and window to environmental variables, the degree of species spatial segregation and evenness. We note that this nested sampling design caused spatial non-independence of single samples (10 to 99 grid cells) that might affect significance testing. To account for this non-independence, we added the dominant eigenvector (PCA1) of the geographical Euclidean distance matrix of the single cells as a covariate in the GLM (reviewed by Hawkins, 2012). Additionally, we the log-transformed sampling fraction entered the models. Coverage increased in a nested manner and accounted for the richness increase from the area effect only. Prior to analyses, we calculated pairwise Pearson's correlations between predictor variables. These were at most moderately correlated (r < .40) except for PCA1 and temperature seasonality (r = .88). Excluding PCA1 from the analysis did not change the remaining results qualitatively.

3 | RESULTS

Japan shows a strong latitudinal gradient in species richness above 35° north (Figure 1a). The proportion of species detected increased logarithmically with the proportion of area sampled (the sampling fraction, Figure 1b). However, this proportion–area relationship was independent from total richness (Figure 1c).

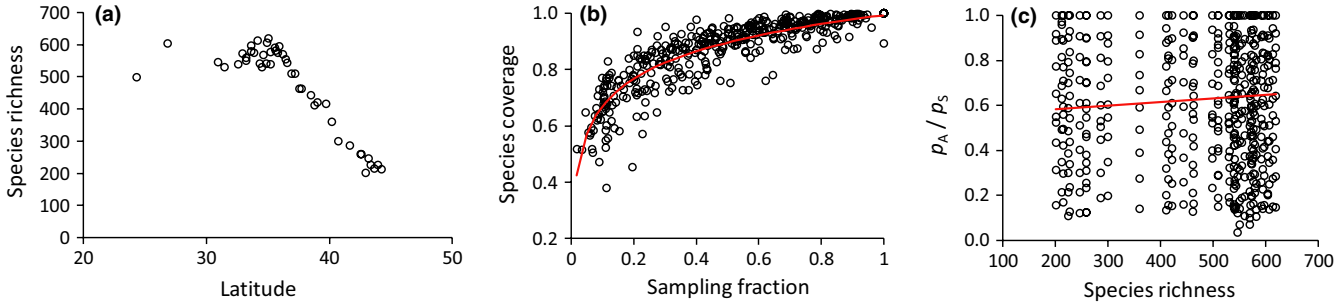
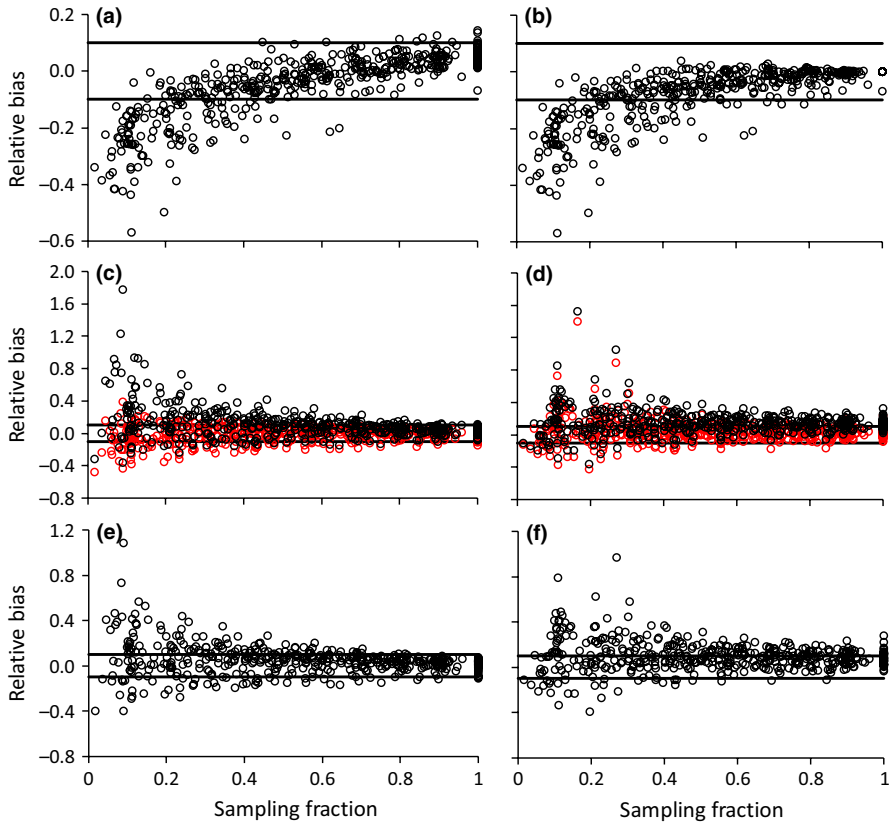


FIGURE 1 (a) Above 35°N, Japanese tree species richness strongly declines with increasing latitude. Data from the 46 grid windows only. (b) The species coverage (p_S = observed richness/total richness) in dependence on the sampling fraction (p_A). Data from 4,600 10 × 10 km^2 grid cells. The increase in sampling fraction is well described by a logarithmic function: fra = 0.15 ln(p_A) + 1; OLS fit: r^2 = .84). (c) The increase in sampling fraction with increasing area was independent from total richness in each window (r^2 < .01)

FIGURE 2 The relative bias in richness estimation of the Chao2 (a), rarefaction (b), species−-area relationship (c: black dots: power function, red dots: logarithmic function; e: arithmetic mean of both estimators), and SAD estimators (d: black dots: log-series, red dots: lognormal, f: arithmetic mean of both estimators) in relation to sampling fraction. Data from 46 Japanese areas of 100 grid cells and 10 samples for each window. The horizontal lines denote the 10% error level



When more than 40% of area was sampled, 93.1% of the Chao2 and 94.1% of the rarefaction estimates had errors of <10% (Figure 2a,b), with a tendency of richness underestimation. Chao2 and rarefaction behaved very similar in all comparisons (Figure 2, Tables 1 and 2) and significantly underestimated true richness below 40% sampling fraction (r_{bias} = .84). The power function SAR approach (powSAR) tended to overestimate richness at lower sampling fraction (r_{bias} = .81, Figure 2c) and was generally less precise than Chao2 and rarefaction but also less affected by sampling fraction (Figure 2c). Only 53.3% of the powSAR estimates were within the 10% error level. In turn, the logarithmic SAR (logSAR) estimates were always below those of the powSAR and tended to underestimate true richness below 40% sampling fraction (r_{bias} = −0.39, Figure 2c). 76.3% of the logSAR estimates were within the 10% error level without a stronger relative at lower sampling fraction (r_{bias} = .22). The SAD approach also provides upper and lower limits in richness. Indeed, the S_{LS} estimator consistently overestimated observed richness and behaved similarly to the powSAR approach (Figure 2d). Only 30.9% of the estimates were within the 10% error level (Figure 2d). The S_{LN} provided robust estimates over the whole range of sampling fraction (r_{bias} = .04), with 67.4% of estimates ranging within the 10% error boundaries (Figure 2d). S_{LN} behaved similar to the logSAR approach (Figure 2c,d) and was slightly positively biased at low sampling fraction (r_{bias} = .13).

Because the SAR and the SAD estimators were designed to provide lower and upper limits to richness, average values might reduce the relative bias and consequently the precision of the estimates (Figure 2e, f). This was indeed the case although the SAD approach still tended to overestimate richness irrespective of sampling fraction

TABLE 1 General linear models of richness coverage (= observed richness/ total richness) and precision in richness estimation in relation to longitude, latitude, the standardized effect sizes of species co-occurrences (SES_{CS}), and the degree of evenness

		Precision					
Variable	**Richness coverage**	**Chao2**	**Rarefaction**	**powSAR**	**logSAR**	**S_{LS}**	**S_{LN}**
Longitude	(–) <0.01	(–) 0.02	(–) 0.02	(–) <0.01	(–) 0.06	(–) 0.01	(–) 0.01
Latitude	0.03	0.05	0.07	(–) <0.01	0.13	0.03	0.06
SES_{CS}	0.01	0.01	0.02	<0.01	0.02	0.09	0.09
Evenness	0.07	0.01	0.01	(–) <0.01	0.09	(–) 0.19	(–) 0.05
ln sampling fraction	0.85	0.60	0.72	0.31	0.31	0.04	0.14
r^2	.86***	.62***	.74***	.31***	.42***	.27***	.26***

Note: Log-transformed percentages of forest cover served as covariate. $N = 460$. Given are partial η^2-values and the coefficients of determination (r^2) of the whole model. Negative signs of the regression parameters are given in brackets.

***Parametric significances of r^2 and $p < .001$.

TABLE 2 General linear modelling of richness coverage (= observed richness/ total richness) and precision in richness estimation in relation to climate, elevation, and forest cover variability

		Precision					
Variable	**Richness coverage**	**Chao2**	**Rarefaction**	**powSAR**	**logSAR**	**S_{LS}**	**S_{LN}**
Elevation variability	0.02	0.01	0.02	(–) 0.02	<0.01	(–) 0.08	(–) 0.02
Precipitation seasonality	0.03	<0.01	0.03	(–) 0.02	(–) <0.01	(–) <0.01	<0.01
Temperature seasonality	(–) <0.01	<0.01	<0.01	0.02	0.10	0.05	0.03
Variability in forest area	0.01	0.02	0.03	(–) <0.01	<0.01	<0.01	<0.01
ln sampling fraction	0.81	0.51	0.65	0.29	0.27	0.02	0.10
PCA1	(–) 0.02	(–) <0.01	(–) 0.01	0.03	0.03	0.02	<0.01
r^2	.62***	.62***	.75***	.32***	.45***	.14***	.20***

Note: Log-transformed sampling fraction and the dominant eigenvector (PCA1) of the geographical distance matrix served as covariates. $N = 460$. Given are partial η^2-values and the coefficients of determination (r^2) of the whole model. Negative signs of the regression parameters are given in brackets.

***Parametric significances: $p < .001$.

(Figure 2f). 66.1% of the averaged SAR and 48.5% of the averaged SAD estimates were within the 10% error level. Both, SAR and SAD, performed as well as the Chao2 and rarefaction approaches when relaxing the error level to 20% (86.5% and 83.0%, respectively, of estimates within the 20% error level) but were not significantly biased by sampling fraction (Figure 2, Table 1).

We found a weak increase in precision of the studied estimators (except of powSAR with increasing latitude (Table 1), which explained between 3% and 7% of the variance in precision, respectively. Further, the pattern of tree species co-occurrences, as quantified by the degree of species spatial segregation, did not markedly affect the Chao2, the rarefaction and both SAR estimators (Table 1). Higher species spatial turnover (positive SES_{CS}) increased the precision of S_{LS} and S_{LN}. Evenness did not affect the precision of the Chao2, the rarefaction and the SAR estimators, while high evenness negatively affected the precision of S_{LS} and S_{LN} (Table 1). In turn, S_{LS} and S_{LN} were least influenced by the proportion of area sampled (Table 1) and provided unbiased estimated even at low cell coverage (Figure 2d).

Relative bias and precision of the four estimators were largely unaffected by environmental variability (Table 2, Figure 3). After accounting for cell coverage and spatial distribution, elevation variability, precipitation and temperature seasonality, the variability in forest cover explained at most 3% of the variance in the Chao2, rarefaction, and the SAR estimators (Table 2). Exception were the S_{LS} and S_{LN} estimators that turned out to be the least biased at intermediate levels of variability in elevation (Table 2, Figure 3d) and forest cover (Table 2, Figure 3l).

4 | DISCUSSION

Our study focused on two important aspects of species richness forecasting. First, we tested the performance of four types of estimators against empirical large spatial scale data. Second, we assessed how precision and relative bias of these estimators were influenced by environmental variables and community composition. Importantly, our study is the first that uses almost complete census data on species occurrences at a large spatial scale. These data enable a precise analysis of the behaviour of each estimator at different fractions of area covered.

Our first task largely confirmed prior tests based on simulation studies (e.g. Chao & Chiu, 2015) and empirical data at small and regional scale (Herzog et al., 2002; Walther & Moore, 2005). The

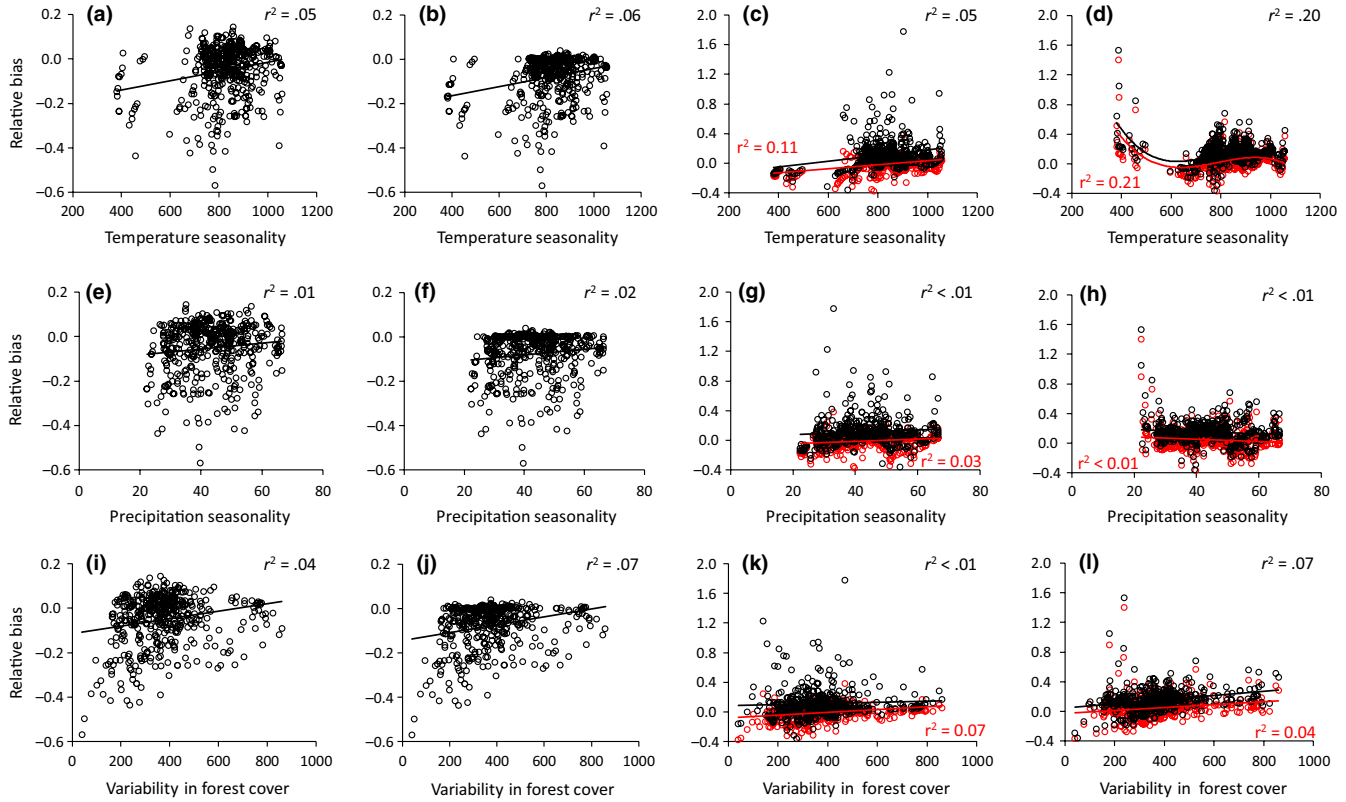


FIGURE 3 Relative bias in richness estimation of the Chao2 (a, e, i), rarefaction (b, f, j), species –area relationship (c, g, k: black dots: power function, red dots: logarithmic function), and SAD estimators (d, h, l: black dots: log-series, red dots: lognormal) in relation to the precipitation and temperature variability, and the variability in forest cover among grid cells of each study area. Data from 460 Japanese areas of 100 contiguous grid cells each. Coefficients of determination refer to ordinary least square linear regressions (in *d* to a third order polynomial regression)

nonparametric Chao2 estimator and rarefaction performed well only at high sampling fraction (Figure 2), being negatively biased at low sample sizes. At higher sampling fraction (say above 40%), rarefaction was superior to Chao2 because all estimates ranged within the 10% error boundary (Figure 2b) without overestimating richness. At lower coverage, Chao2 and rarefaction performed nearly identical and underestimated true richness (even by 60%). Nevertheless, even at low sampling fraction both estimators performed not worse or even better than the parametric SAR and SAD estimators (Figure 2). This is a promising result obtained from large-scale empirical data meaning that we have very different approaches to diversity forecasting that provide reliable results independent of the available data structure. For instance, in many cases available data contain only species lists or information on total abundances and do not allow for the construction of rarefaction or species accumulation curves, while Chao2 or SAD provides alternatives.

The SAR and SAD estimators were designed to assess lower and upper limits of richness. Indeed, both estimators performed well with this task (Figure 2) although tending to overestimate richness irrespective of sampling fraction. With respect to the SAD estimators, this result confirms the findings of Kunin et al. (2018) based on British vascular plants. The behaviour of the SAD estimator allows some conclusions about the underlying occupancy—rank order distribution. The log-series assumption caused overestimation of richness (Figure 2). The log-series is a sample distribution (Fisher et al., 1943) and implies fairly constant proportions of species along the occupancy axis. This distribution is predicted from ecological drift (Hubbell, 2001) and a variety of niche-based SAD models (Tokeshi, 1998). In turn, recent meta-analyses on global plant species abundance distributions pointed to a higher proportion of lognormal type SADs with a lot of rare species (Ulrich et al., 2016a; Ulrich et al., 2016, Ulrich et al., 2018). Importantly, Ulrich et al. (2016a), Ulrich, Soliveres, et al. (2016) revealed a latitudinal gradient in SAD shape with a preponderance of lognormal distributions at higher latitudes. This is in line with the present findings from the Japanese flora. The fact that the lognormal based SAD constantly performed better than the log-series based estimator (Figure 2) is a strong corroboration for the prevalence of lognormal types SADs in vascular plants. Of course, our results regard regional (large-scale) abundance distributions, whereas the surveys of Ulrich et al. (2016a), Ulrich et al. (2016), Ulrich et al. (2018) focused on local communities. However, sampling theory (Green & Plotkin, 2007) predicts local log-series sample distributions to scale up to the regional scale, while local lognormal distributions behave scale invariant only if the constituent species are not dispersal limited (Green & Plotkin, 2007; Hubbell, 2001). Therefore, our finding of a better fit of regional lognormal abundance distributions corroborates the view that local plant communities at least at higher latitudes are dispersal limited and lognormally

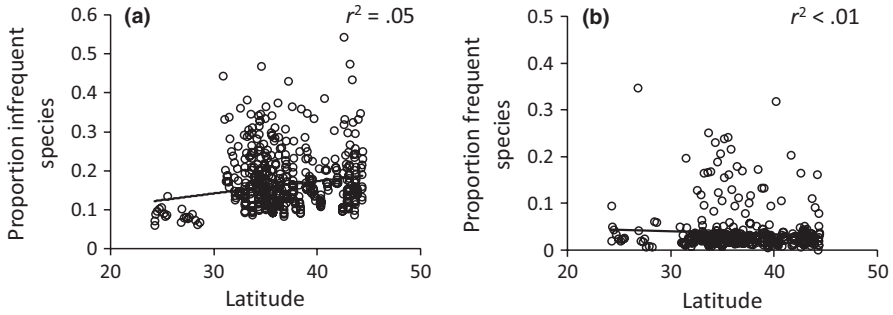


FIGURE 4 Proportions of (a) infrequent (species with only single or double incidence among the 100 cell of each window) and (b) frequent (species occurring in all grid cells of each window) in dependence of latitude. Coefficients of determination refer to ordinary least squares linear regressions. Significance of the regression in (a) $p(F_{1,458}) < .001$

distributed. Respective studies for lower latitudinal, tropical communities are missing.

With respect to the SAR estimators, we found a consistently better estimate of the logarithmic SAR (Figure 2). The power function SAR constantly tended to overestimate richness, which indicates deviations from the power function species accumulation irrespective of spatial scale (Figure 2). This finding stands in certain contrast to recent analyses of richness patterns that clearly pointed to a prevalence of power function SARs at local (Dengler et al., 2020) and regional scales (Dengler, 2009). However, direct comparisons of SAR model fits do not exclude spatial scale dependent deviations from the model. With respect to extrapolations, even small (and possibly statistically insignificant) deviations might cause imprecise richness estimates. Therefore, observed positive relative biases in richness estimation demonstrate that the accumulation of species richness in natural systems at regional scales occurs slower than predicted by a power function. In this respect, Dengler (2009) and Dengler et al. (2020) reported that the power function SAR reliably predicts the increase in terrestrial vegetation over ten orders of magnitude while logarithmic SAR models proved inappropriate. However, the data of Dengler et al. (2020) covered local samples of continuous vegetation. Rosindell and Cornell (2007) and Storch (2016) demonstrated that SARs can be triphasic with shallower slopes at regional scales. Consequently, extrapolating across the local–regional border would cause either an over- or an underestimation of richness depending on the proportion of data points from local and regional samples used to construct the SAR. This is the pattern reported here. In this respect, Storch, Keil, and Jetz (2012) found that local to continental SARs collapse into one common power function after area was rescaled by mean species range sizes. Unfortunately, for most datasets (including the present) such data are unavailable.

Regarding our second main task, we first asked whether and how geographical and climate variability influences the accuracy of richness estimation. This was indeed the case. Although weak, we found positive latitudinal gradients in the precision of all estimators with the exception of powSAR (Table 1). This trend was apparent even after correcting for sampling fraction. Latitude is only a surrogate variable. We hypothesize that this latitudinal gradient is caused by respective latitudinal changes in community composition. Surprisingly, we found an increase in relatively infrequent species (those that occurred only in one or in two grid cells) at higher latitudes (Figure 4a), while the proportion of frequent species (occurring in all cells) did not show any significant latitudinal gradient (Figure 4b). Particularly, the southernmost Japanese islands of the Ryukyus archipelago with subtropical climate had the lowest proportions of infrequent species (Figure 4a). The pattern on the Ryukyus islands (<30 degree south) might be explained by the comparatively homogeneous vegetation across the archipelago (Kubota, unpublished) and lack of higher mountains. Species composition of small islands is basically a subset of the larger island (i.e. nested), and the former do not have own endemic species. Such a pattern naturally results in a low proportion of infrequent species. In turn, the middle part of Japan (30–40 degrees south) is characterized by higher mountains and many regionally endemic species or disjunctive distributed species in their summits. Consequently, the proportion of rare species is relatively high.

We also asked whether the variability in environmental variables affects estimator performance. This was only marginally the case (Table 2, Figure 3). After correcting for covariates (Table 2), only temperature variability appeared to markedly influence estimator performance (Table 2) explaining between 3% and 10% of variability in logSAR and the SAD estimators. Importantly, most robust were Chao2 and rarefaction indicating that the underlying theoretical models are indeed sufficiently independent of environmental covariates. Again, environmental variability might only indirectly influence estimator performance via respective changes in the spatial distribution of species and/or the distribution of abundances. Both distributions are known to change along gradients of environmental variability (Ackerly, Knight, Weiss, Barton, & Starmer, 2002; Pottier et al., 2013). The fact that this did not significantly influence estimator performance is an important and a strong argument in favour of the underlying theoretical assumptions of scale invariant species spatial distribution patterns.

Finally, we asked whether and how evenness in species incidence and the spatial species turnover among grid cells influence estimation accuracy. To our surprise, higher spatial species turnover was positively related only to the performance of the SAD estimators. We expected to see a signal for SAR as the SAR slope is directly linked to the decay in compositional similarity among communities (Soininen, McDonald, & Hillebrand, 2007). Low spatial species turnover might decrease the proportion of infrequent species in dispersal limited communities as predicted by neutral theory (Hubbell, 2001). Low proportions of infrequent species are equivalent to a higher evenness in species incidence. If evenness

were higher than predicted by the log-series distribution, both SAD estimators would perform weak as species richness is also higher than the predicted upper limit estimated by the S_{LS} estimator (Ulrich & Ollik, 2005). This is the pattern observed here (Table 2).

ACKNOWLEDGMENTS

W.U. was supported by the Polish National Science Centre (UMO-2017/27/B/NZ8/00316). B.K and Y.K is supported by the Program for Advancing Strategic International Networks to Accelerate the Circulation of Talented Researchers, the Japan Society for the Promotion of Science and the Environment Research and Technology Development Fund (4-1501 and 4-1802) of the Ministry of the Environment, Japan.

DATA AVAILABILITY STATEMENT

The raw data used in the present publication have been published by Kubota et al. (2015) and Kusumoto et al. (2015).

ORCID

Werner Ulrich https://orcid.org/0000-0002-8715-6619
Buntarou Kusumoto https://orcid.org/0000-0002-5091-3575

REFERENCES

Ackerly, D. D., Knight, C. A., Weiss, S. B., Barton, K., & Starmer, K. P. (2002). Leaf size, specific leaf area and microhabitat distribution of chaparral woody plants: Contrasting patterns in species level and community level analyses. *Oecologia*, *130*, 449–457. https://doi.org/10.1007/s004420100805

Beukema, J. J., & Dekker, R. (2012). Estimating macrozoobenthic species richness along an environmental gradient: Sample size matters. *Estuarine, Coastal and Shelf Science*, *111*, 67–74. https://doi.org/10.1016/j.ecss.2012.06.013

Brose, U., Martinez, N. D., & Williams, R. J. (2003). Estimating species richness: Sensitivity to sample coverage and insensitivity to spatial patterns. *Ecology*, *84*, 2364–2377. https://doi.org/10.1890/02-0558

Chao, A. (1987). Estimating the population size for capture-recapture data with unequal catchability. *Biometrics*, *43*, 783–791. https://doi.org/10.2307/2531532

Chao, A., & Chiu, C. H. (2015). *Nonparametric estimation and comparison of species richness*. In: eLS. Chichester: John Wiley & Sons Ltd.

Chao, A., & Chiu, C. H. (2016). Species richness: estimation and comparison. Wiley Stats Ref: Statistics Reference Online, 1-26.

Chao, A., Colwell, R. K., Lin, C. W., & Gotelli, N. J. (2009). Sufficient sampling for asymptotic minimum species richness estimators. *Ecology*, *90*, 1125–1133. https://doi.org/10.1890/07-2147.1

Chao, A., Colwell, R. K., Gotelli, N. J., & Thorn, W. (2019). Proportional mixture of two rarefaction/extrapolation curves to forecast biodiversity changes under landscape transformation. *Ecology Letters*, *22*(11), 1913–1922. https://doi.org/10.1111/ele.13322

Chiarucci, A., Bacaro, G., & Scheiner, S. M. (2011). Old and new challenges in using species diversity for assessing biodiversity. *Philosophical Transactions of the Royal Society B*, *366*, 2426–2437. https://doi.org/10.1098/rstb.2011.0065

Chiarucci, A., Bacaro, G., Rocchini, D., & Fattorini, L. (2008). Discovering and rediscovering the sample-based rarefaction formula in the ecological literature. *Community Ecology*, *9*, 121–123. https://doi.org/10.1556/ComEc.9.2008.1.14

Chiarucci, A., Enright, N. J., Perry, G. L. W., Miller, B. P., & Lamont, B. B. (2003). Performance of non-parametric species richness estimators in a high diversity plant community. *Diversity and Distributions*, *9*, 283–295. https://doi.org/10.1046/j.1472-4642.2003.00027.x

Colwell, R. K., Chao, A., Gotelli, N., Lin, S. Y., Mao, C. X., Chazdon, R. L., & Longino, J. T. (2012). Models and estimators linking individual-based and sample-based rarefaction, extrapolation, and comparison of assemblage. *Journal of Plant Ecology*, *5*, 3–21.

Connor, E. F., & McCoy, E. D. (1979). The statistics and biology of the species–area relationship. *American Naturalist*, *113*, 791–833. https://doi.org/10.1086/283438

Dengler, J., Matthews, T. J., Steinbauer, M. J., Boch, S., Chiarucci, A., Conradi, T., Biurrun, I. (2020). Species-area relationships in continuous vegetation: evidence from Palaearctic Grasslands. *Journal of Biogeography 47*, 72–86.

Dengler, J. (2009). Which function describes the species-area relationship best? A review and empirical evaluation. *Journal of Biogeography*, *36*, 728–744. https://doi.org/10.1111/j.1365-2699.2008.02038.x

De Thoisy, B., Brosse, S., & Dubois, M. A. (2008). Assessment of large-vertebrate species richness and relative abundance in Neotropical forest using line-transect censuses: What is the minimal effort required? *Biodiversity and Conservation*, *17*, 2627–2644. https://doi.org/10.1007/s10531-008-9337-0

Dewdney, A. K. (1998). A general theory of the sampling process with application to the "veil line". *Theoretical Population Biology*, *54*, 294–302.

Fattorini, S. (2013). Regional insect inventories require long time, extensive spatial sampling and good will. *PLoS One*, *8*(4), e62118. https://doi.org/10.1371/journal.pone.0062118

Fisher, R. A., Corbet, A. S., & Williams, C. B. (1943). The relation between the number of species and the number of individuals in a random sample of an animal population. *Journal of Animal Ecology*, *12*, 42–58. https://doi.org/10.2307/1411

Green, J. L., & Plotkin, J. B. (2007). A statistical theory for sampling species abundances. *Ecology Letters*, *10*, 1037–1045. https://doi.org/10.1111/j.1461-0248.2007.01101.x

Gwinn, D. C., Allen, M. S., Bonvechio, K. I., Hoyer, M. V., & Beesley, L. S. (2016). Evaluating estimators of species richness: The importance of considering statistical error rates. *Methods in Ecology and Evolution*, *7*, 294–302. https://doi.org/10.1111/2041-210X.12462

Hawkins, B. A. (2012). Eight (and a half) deadly sins of spatial analysis. *Journal of Biogeography*, *39*, 1–9. https://doi.org/10.1111/j.1365-2699.2011.02637.x

Harte, J., & Newman, E. A. (2014). Maximum information entropy: A foundation for ecological theory. *Trends in Ecology and Evolution*, *29*, 384–389. https://doi.org/10.1016/j.tree.2014.04.009

Herzog, S. K., Kessler, M., & Cahill, T. M. (2002). Estimating species richness of tropical bird communities from rapid assessment data. *The Auk*, *119*, 749–769.

Hortal, J., Borges, P. A. V., & Gaspar, C. (2006). Evaluating the performance of species richness estimators: Sensitivity to sample grain size. *Journal of Animal Ecology*, *75*, 274–287. https://doi.org/10.1111/j.1365-2656.2006.01048.x

Hubbell, S. P. (2001). *The unified theory of biogeography and biodiversity*. Princeton, NJ: Princeton University Press.

Kubota, Y., Kusumoto, B., Shiono, T., & Tanaka, T. (2017). Phylogenetic properties of Tertiary relict flora in the east Asian continental islands: Imprint of climatic niche conservatism and in situ diversification. *Ecography*, *40*, 436–447. https://doi.org/10.1111/ecog.02033

Kubota, Y., Shiono, T., & Kusumoto, B. (2015). Role of climate and geohistorical factors in driving plant richness patterns and endemicity on the east Asian continental islands. *Ecography*, *38*, 639–648. https://doi.org/10.1111/ecog.00981

Kunin, W. E. (1998). Extrapolating species abundance across spatial scales. *Science*, *281*, 1513–1515. https://doi.org/10.1126/science.281.5382.1513

Kunin, W. E., Harte, J., He, F., Hui, C., Jobe, R. T., Ostling, A., … Varma, V. (2018). Upscaling biodiversity: Estimating the species-area relationship from small samples. *Ecological Monographs*, *88*, 170–187.

Kusumoto, B., Villalobos, F., Shiono, T., & Kubota, Y. (2019). Reconciling Darwin's naturalization and pre-adaptation hypotheses: An inference from phylogenetic fields of exotic plants in Japan. *Journal of Biogeography*, *46*(11), 2597–2608. https://doi.org/10.1111/jbi.13683

Lennon, J. J., Koleff, P., Greenwood, J. J. D., & Gaston, K. J. (2001). The geographical structure of British bird distributions: Diversity, spatial turnover and scale. *Journal of Animal Ecology*, *70*, 966–979. https://doi.org/10.1046/j.0021-8790.2001.00563.x

Magurran, A. E. (2004). *Measuring biological diversity*. Oxford, UK: Blackwell.

Magurran, A. E., & McGill, B. J. (2011). *Biological diversity: frontiers in measurement and assessment*. Oxford, UK: Oxford University Press.

Mertes, K., & Jetz, W. (2018). Disentangling scale dependencies in species environmental niches and distributions. *Ecography*, *41*, 1604–1615.

Okada, M., Iizumi, T., Nishimori, M., & Yokozawa, M. (2008). Mesh climate change data of Japan Vers. 2 for climate change impact assessments under IPCC SRSS A1B and A2. *Journal of Agricultural Meteorology*, *65*, 97–109.

Palmer, M. W. (1990). The estimation of species richness by extrapolation. *Ecology*, *71*, 1195–1198. https://doi.org/10.2307/1937387

Palmer, M. W. (1991). Estimating species richness: The second order jackknife reconsidered. *Ecology*, *72*, 1512–1513. https://doi.org/10.2307/1941127

Pottier, J., Dubuis, A., Pellissier, L., Maiorano, L., Rossier, L., Randin, C. F., … Guisan, A. (2013). The accuracy of plant assemblage prediction from species distribution models varies along environmental gradients. *Global Ecology and Biogeography*, *22*, 52–63. https://doi.org/10.1111/j.1466-8238.2012.00790.x

Preston, F. W. (1948). The commonness, and rarity of species. *Ecology*, *29*, 254–283. https://doi.org/10.2307/1930989

Rosenzweig, M. L. (1995). *Species diversity in space and time*. Cambridge, UK: Cambridge University Press.

Rosindell, J., & Cornell, S. J. (2007). Species–area relationships from a spatially explicit neutral model in an infinite landscape. *Ecology Letters*, *10*, 586–595. https://doi.org/10.1111/j.1461-0248.2007.01050.x

Scheiner, S. M. (2003). Six types of species–area curves. *Global Ecology and Biogeography*, *12*, 441–447. https://doi.org/10.1046/j.1466-822X.2003.00061.x

Shen, T. J., & He, F. L. (2008). An incidence-based richness estimator for quadrats sampled without replacement. *Ecology*, *89*, 2052–2060. https://doi.org/10.1890/07-1526.1

Šizling, A., & Storch, D. (2004). Power-law species–area relationships and self-similar species distributions within finite areas. *Ecology Letters*, *7*, 60–68. https://doi.org/10.1046/j.1461-0248.2003.00549.x

Soininen, J., McDonald, R., & Hillebrand, H. (2007). The distance decay of similarity in ecological communities. *Ecography*, *30*, 3–12. https://doi.org/10.1111/j.0906-7590.2007.04817.x

Stone, L., & Roberts, A. (1990). The checkerboard score and species distributions. *Oecologia*, *85*, 74–79. https://doi.org/10.1007/BF00317345

Storch, D., Keil, P., & Jetz, W. (2012). Universal species–area and endemics–area relationships at continental scales. *Nature*, *488*, 78–81. https://doi.org/10.1038/nature11226

Storch, D. (2016). The theory of the nested species–area relationship: Geometric foundations of biodiversity scaling. *Journal of Vegetation Science*, *27*, 880–891. https://doi.org/10.1111/jvs.12428

Tokeshi, M. (1998). *Species coexistence: Ecological and evolutionary perspectives*. Oxford, UK: Blackwell.

Triantis, K. A., Guilhaumon, F., & Whittaker, R. J. (2012). The island species–area relationship: Biology and statistics. *Journal of Biogeography*, *39*, 215–231. https://doi.org/10.1111/j.1365-2699.2011.02652.x

Ugland, K. I., Gray, J., & Ellingsen, K. E. (2003). The species accumulation curve and estimation of species richness. *Journal of Animal Ecology*, *72*, 888–897. https://doi.org/10.1046/j.1365-2656.2003.00748.x

Ulrich, W., Kusumoto, B., Shiono, T., & Kubota, Y. (2016a). Climatic and geographical correlates of global forest tree species abundance distributions and community evenness. *Journal of Vegetation Science*, *27*, 295–305.

Ulrich, W., Soliveres, S., Thomas, A. D., Dougill, A. J., & Maestre, F. T. (2016). Environmental correlates of species rank – abundance distributions in global drylands. *Perspectives in Plant Ecology, Evolution and Systematics*, *20*, 56–64. https://doi.org/10.1016/j.ppees.2016.04.004

Ulrich, W., Nakadai, R., Matthews, T., & Kubota, Y. (2018). The two-parameter Weibull distribution as a universal tool to model the variation in species relative abundances. *Ecological Complexity*, *36*, 110–116. https://doi.org/10.1016/j.ecocom.2018.07.002

Ulrich, W., & Ollik, M. (2005). Limits to the estimation of species richness: The use of relative abundance distributions. *Diversity and Distributions*, *11*, 265–273. https://doi.org/10.1111/j.1366-9516.2005.00127.x

Ulrich, W., Ollik, M., & Ugland, K. I. (2010). A meta-analysis of species - abundance distributions. *Oikos*, *119*, 1149–1155. https://doi.org/10.1111/j.1600-0706.2009.18236.x

Walther, B. A., & Moore, J. L. (2005). The concepts of bias, precision, and accuracy, and their use in testing the performance of species richness estimators, with a literature review of estimator performance. *Ecography*, *28*, 815–829. https://doi.org/10.1111/j.2005.0906-7590.04112.x

Yonekura, K., & Kajita, T. (2003). *BG plants Japanese name – scientific names index (YList)*. Retrieved from http://bean.bio.chiba-u.jp/bgplants/ylist_main.html

BIOSKETCH

Authors are macroecologists with a focus on bioconservation, the dynamics of local community assembly and large-scale distributions of species within an evolutionary and environmental context. They use local and large-scale biogeographical datasets, simulation studies and matrix-based analytical tools to infer the patterns and processes of species co-existence, abundances and meta-community dynamics.

Author contributions: WU designed the study, analysed the data and wrote the first draft. B.K. and Y.K. provided the data and drafted parts of the text. S.F. gave input to the concept, introduction and discussion. All authors contributed to the final text version.

How to cite this article: Ulrich W, Kusumoto B, Fattorini S, Kubota Y. Factors influencing the precision of species richness estimation in Japanese vascular plants. *Divers Distrib*. 2020;26:769–778. https://doi.org/10.1111/ddi.13049