RESEARCH PAPER

Performance of species richness estimators across assemblage types and survey parameters

Gordon C. Reese[1]*, Kenneth R. Wilson[1] and Curtis H. Flather[2]

[1]*Department of Fish, Wildlife, and Conservation Biology, Colorado State University, Fort Collins, CO 80523, USA,*
[2]*US Forest Service, Rocky Mountain Research Station, Fort Collins, CO 80526, USA*

ABSTRACT

Aim A raw count of the species encountered across surveys usually underestimates species richness. Statistical estimators are often less biased. Nonparametric estimators of species richness are widely considered the least biased, but no particular estimator has consistently performed best. This is partly a function of estimators responding differently to assemblage-level factors and survey design parameters. Our objective was to evaluate the performance of raw counts and nonparametric estimators of species richness across various assemblages and with different survey designs.

Location We used both simulated and published field data.

Methods We evaluated the bias, precision and accuracy of raw counts and 13 nonparametric estimators using simulations that systematically varied assemblage characteristics (number of species, species abundance distribution, total number of individuals, spatial configuration of individuals and species detection probability), sampling effort and survey design. Results informed the development of an estimator selection framework that we evaluated with field data.

Results When averaged across assemblages, most nonparametric estimators were less negatively biased than a raw count. Estimators based on the similarity of repeated subsets of surveys were most accurate and their accumulation curves appeared to reach asymptotes fastest. Number of species, species abundance distribution and effort had the largest effects on performance, ultimately by affecting the proportion of the species pool contained in a sample. Our estimator selection framework showed promising results when applied to field data.

Main conclusions A raw count of the number of species in an area is far from the best estimate of true species richness. Nonparametric estimators are less biased. Newer largely unused, estimators perform better than more well known and longer established counterparts under certain conditions. Given that there is generally a trade-off between bias and precision, we believe that estimator variance, which is often not reported when presenting species richness estimates, should always be included.

*Correspondence: Gordon C. Reese, Department of Fish, Wildlife, and Conservation Biology, Colorado State University, Fort Collins, CO 80523, USA.
E-mail: greese126@gmail.com*

INTRODUCTION

All aspects of biological diversity, from genes to ecosystems, can inform decisions in ecological monitoring, conservation management and reserve design. How best to quantify diversity at a species level remains a topic of much debate (Brose *et al.*, 2003).

Attempts to quantify species diversity often include: (1) a count or estimate of the unique species in a delineated area (species richness), (2) a measure of the uniformity of abundances among species (species evenness), or (3) some measure of species composition (species similarity; see Magurran, 2004). Of these, species richness (SR) is the most conceptually simple and

frequently used (Moreno *et al.*, 2006; Drake, 2007; Lopez *et al.*, 2012).

The number of species observed (S_{obs}) is the most straightforward estimate of the true number of species in an assemblage (S_{true}), but it has a known negative bias and is based on the naïve assumption that all species are detected with a probability equal to one (Palmer, 1990; Nichols *et al.*, 1998). Fortunately, species distribution and abundance patterns can be used to inform statistically derived estimators of S_{true}. All estimators have underlying assumptions, but the statistical estimators of SR are typically less negatively biased than S_{obs} (Baltanás, 1992; Chiarucci *et al.*, 2003; Walther & Moore, 2005).

Besides S_{obs}, there are three categories of SR estimators. One category includes approaches for extrapolating a species accumulation curve to an asymptote, often using a negative exponential model (Holdridge *et al.*, 1971), the Michaelis–Menten equation (Michaelis & Menten, 1913) or a power model (Arrhenius, 1921; Tjørve, 2009). A second category includes parametric methods that involve: (1) interpolating under a distribution fit to abundance data or (2) applying an estimator that assumes that all species are equally detectable. A third category includes nonparametric estimators, which make no assumption about the underlying distribution of the data.

The performance of SR estimators has largely depended on whether underlying assumptions are met, making factors such as the species abundance distribution, species detection probability (p), survey effort and S_{true} (Baltanás, 1992; Keating & Quinn, 1998; Brose *et al.*, 2003) important to understanding their performance. Ultimately, no estimator has consistently performed best, but the nonparametric estimators have generally performed better than the other categories (see Table 1 in Cao *et al.* 2004; Table 3 in Walther & Moore 2005; Table 1 in Reese 2012). We thus focus on evaluating nonparametric estimators, including several that are relatively untested (see Burnham & Overton, 1978; Pledger, 2000; Cao *et al.*, 2001; Cao *et al.*, 2004), by comprehensively and consistently varying factors that can compromise estimator assumptions.

Most of the nonparametric SR estimators can be categorized as those that model either heterogeneity in p (commonly denoted as model M_h; see Otis *et al.*, 1978) or the similarity between replicate subsets of the survey data. Many in the first group were developed for population estimation using mark–recapture data under an assumption of geographic and demographic closure, an assumption that also applies to all SR estimators. Brose *et al.* (2003) detailed additional challenges that arise when an M_h population estimator is used to estimate S_{true}. First, differences in detectability between species can be larger and more difficult to model than those between the individuals composing a population of one species. Second, when an M_h estimator is used to estimate population size from encounter histories of individuals in surveys repeated at the same location, an assumption is that p will vary across individuals but remain constant over time. When those estimators are instead used to estimate S_{true} from encounter histories of species in surveys replicated at different locations, the comparable assumption is that detection probabilities vary among

Table 1 Levels of factors simulated to evaluate species richness estimators.

Description	Symbol	Levels
Total (true) number of species	S_{true}	25
		100
		500
Total abundance across all species	N	6250
		12500
Species abundance distribution	*Abund*	Log-series
		Log-normal
Spatial relationship between individuals of a species	*Config*	Aggregated
		Hyper-dispersed
		Random*
Mean detection probability of three species abundance groups	p	(0.5, 0.5, 0.5)
		(0.9, 0.9, 0.9)
		(0.5, 0.7, 0.9)
		(0.9, 0.7, 0.5)
Spatial arrangement of surveyed grid cells	*Design*	Random
		Transect (linear)†
Amount of landscape surveyed	*Effort*	1% (100 cells)
		5% (500 cells)

*Individuals were spaced less regularly (*aggregated*) or more evenly (*hyperdispersed*) than expected by chance (*random*).
†Surveys were configured randomly or as random linear transects of 50 grid cells.

species but are constant across space. This assumption can be violated when distributions are spatially heterogeneous, which occurs regularly in natural systems (Legendre, 1993; Deblauwe *et al.*, 2008).

Our primary objective was to evaluate nonparametric SR estimators across assemblages that were systematically varied both in their attributes (number of species, total number of individuals, species abundance distribution, spatial aggregation and species detection probability) and how they were sampled (effort and survey design). Given the large number of factors and the benefits of knowing the true factor values, we relied heavily on simulated data, but also evaluated estimators using field data (see the Supporting Information). An important secondary objective was to expand the estimator selection approach proposed by Brose *et al.* (2003). Their framework targeted estimator accuracy and based selection on the ratio of S_{obs} to the mean SR estimate, which is an estimate of sample coverage (*sc*) or the fraction of a species pool represented in a sample (i.e. S_{obs}/S_{true}). Our contribution included additional selection criteria based on bias, precision and an estimate of *sc* based on Jaccard's similarity coefficient.

METHODS

Simulation procedure

We evaluated the performances of S_{obs} and 13 nonparametric SR estimators across simulated assemblages using the program

SimAssem (Reese *et al.*, 2013). We systematically varied the total number of species, total assemblage abundance, type of species abundance distribution, spatial structure of individual occurrences within a species and detection probability of species abundance groups in a factorial design (Table 1). Similarly, we simulated a sampling procedure in which we varied the number of surveyed grid cells and how they were selected. Based on our review of the literature and past experience, we selected these factors and factor levels to cover a wide range of realistic scenarios. We generated 42 replicates of each factor combination (24,192 total simulations) as a compromise between computer processing time and having a sufficient number of replicates to ensure an adequate sample size after we removed replicates involving estimation issues (e.g. lack of convergence for the Mixture estimator).

We simulated assemblages with 25, 100 and 500 species (factor S_{true}), populating them with a total of 6250 and 12,500 individuals (factor *N*). These values restricted the average number of individuals encountered per species to the range of several datasets (see Williams, 1939; Lewis & Taylor, 1967; Dallmeier *et al.*, 1991). To simulate species abundance patterns (factor *Abund*), we used log-normal (Preston, 1948) and log-series distributions (Fisher *et al.*, 1943). SimAssem generated: (1) log-normal distributions by drawing a random log-normal variate ($\mu = 0$, $\sigma = 1$) for each species, dividing each variate by the sum of all the variates and rescaling by multiplying each variate by *N* and (2) log-series distributions by calculating the number of species to populate with *z* individuals, where $z = 1, 2, \ldots, N$ (see Magurran, 2004). For comparison, we also simulated evenly distributed assemblages with particulate-niche distributions (MacArthur, 1957) where each individual was randomly assigned to a species. Given that: (1) MacArthur considered this distribution unsatisfactory (Magurran, 2004) and (2) it is infrequently referenced in the literature, we provide the results in Table S1 in Supporting Information (see also Reese, 2012).

We assigned *x*- and *y*-coordinates [0, 1] to each individual via three spatial configuration options available in SimAssem (factor *Config*). The *random* configuration located each individual with a pair of random uniform variates [0, 1]. For the *hyperdispersed* configuration, SimAssem assigned each individual a square territory with a linear dimension of $1/\sqrt{n_i}$, where n_i is the abundance of species *i*. Territories were adjacent in the horizontal and vertical directions and collectively formed a grid across the entire landscape. Individuals were located a random distance and random direction from the bottom left corner of their territory. When a randomized location in an overlapping territory fell outside the top or right landscape boundary, as happened occasionally, the individual was randomly placed on the landscape. To determine the locations around which species could aggregate with the *aggregated (centres equal abun)* option, SimAssem generated n_i random uniform variates (RUV) for each species *i*, and every RUV ≥ 0.98 increased by one the number of randomly placed aggregate centres, resulting in approximately one centre for every 50 individuals. Each individual of a species was then randomly allocated to a centre and located on the landscape when two conditions were met. First,

0.95^d had to equal or exceed a RUV, where *d* is a randomly selected distance [0, 1]. Second, the individual's location had to fall on or within landscape boundaries when placed distance *d* in a random direction (0–359°) from the selected centre.

Species-specific detection probabilities (factor *p*) were randomly drawn from beta distributions with expected means of 0.5, 0.7 and 0.9. To isolate the effect of *p*, we constrained variances to the range 0.010–0.015. For two factor levels, we drew each *p* from the same distribution, with an expected mean of 0.5 or 0.9 (α and β parameters equalled 10 and 10, or 4.5 and 0.5, respectively). We evaluated two additional factor levels where *p* varied as a function of abundance (see Selmi & Boulinier, 2004; Pagano & Arnold, 2009). We grouped species into thirds based on abundance and randomly selected abundance groups to accommodate the additional species since the tested levels of S_{true} are not divisible by three. For one factor level we assigned each species in the least, moderate and most abundant groups a *p* that was randomly drawn from a distribution with an expected mean of 0.5, 0.7 ($\alpha = 14$, $\beta = 6$) or 0.9, respectively. We created another factor level by reversing the expected means of the abundance groups.

In SimAssem, every cell of an overlaid 100×100 grid is a potential survey site. We evaluated two levels of effort (factor *Effort*) by sampling either 100 (1%) or 500 (5%) grid cells. In addition, we used two survey designs (factor *Design*), random and a linear transect design. Each transect contained 50 adjacent cells in a randomly selected horizontal or vertical orientation and transects were added until the specified number of grid cells was surveyed. Previously surveyed grid cells intersected by a new transect were applied to the transect length, but were not double-counted. An individual was encountered when two conditions were met. First, a surveyed cell had to contain at least one individual. Second, a RUV, where one was generated for each individual in the cell, had to be $\leq p$.

Species richness estimators

In addition to S_{obs}, we evaluated the performance of 13 estimators, where some are variants of others (see Table 2 for details and abbreviations). Those estimators belonging to the M_h class included two based on abundance patterns (i.e. the number of species with an exact number of individuals encountered) and nine based on incidence patterns (i.e. the number of species encountered in an exact number of surveys). We used the Rmark package (Version 1.9.5; Laake & Rexstad, 2008) in the program R (R Development Core Team, 2009) to generate Mixture estimates based on two groups using program MARK (Version 6.0; White & Burnham, 1999).

Two additional estimators, CY-1 and CY-2, are based on the similarity of two replicate subsets of surveys. A CY-1 estimate equals average SR across the replicate sets of surveys ($\overline{\text{SR}}$) divided by the Jaccard coefficient (JC), $\text{JC} = c/(a + b + c)$, where *a* and *b* are the numbers of species unique to each subset and *c* is the number of species common to both subsets. Thus, CY-1 is only calculable when individuals are encountered in two or more surveys. A CY-2 estimate equals the slope plus the inter-

Estimator name	Abbreviation	Reference
Abundance-based coverage*†	ACE	Chao & Lee (1992)
Chao1 (bias-corrected)*†	Chao1	Chao (1984)
Bootstrap*‡	Boot	Smith & van Belle (1984)
Chao2 (bias-corrected)*‡	Chao2	Chao (1987)
Incidence-based coverage*‡	ICE	Lee & Chao (1994)
1st-order jackknife*‡	Jack1	Burnham & Overton (1978)
2nd-order jackknife*‡	Jack2	Burnham & Overton (1978)
3rd-order jackknife*‡	Jack3	Burnham & Overton (1978)
4th-order jackknife*‡	Jack4	Burnham & Overton (1978)
5th-order jackknife*‡	Jack5	Burnham & Overton (1978)
Mixture-model*‡§	Mixture	Pledger (2000)
CY-1¶	CY-1	Cao *et al.* (2001)
CY-2¶	CY-2	Cao *et al.* (2004)
Number of species observed	S_{obs}	

Table 2 Evaluated species richness estimators.

*Estimation by modelling heterogeneity in detection probability of species.
†Estimation using sample abundance patterns (i.e. number of individuals).
‡Estimation using sample incidence patterns (i.e. number of surveys).
§Estimation using maximum likelihood by modelling heterogeneity using mixtures.
¶Estimation using similarity of replicate surveys of species with the Jaccard coefficient.

cept of a regression line fitted to $\overline{SR}$ versus JC, where each value was the average of 100 iterations (see Cao *et al.*, 2004). SimAssem used five regression points ($\overline{SR}$–JC pairs) when there were between 10 and 19 surveys with encounters and 10 regression points when there were 20 or more surveys with encounters.

Across estimators, the range of possible estimates is large. The bootstrap estimator can extrapolate to $S_{obs} \times 2$ (Colwell & Coddington, 1994), the jackknife estimators nearly to S_{obs} (the order of the estimator + 1), Chao2 to $S_{obs}^2/2$ and CY-2 to $\overline{SR}^2$ (Cao *et al.*, 2004). Higher-order jackknife estimates can be $< S_{obs}$ (Lam & Kleinn, 2008) or even negative. We therefore report the proportion of estimates that were $< S_{obs}$.

Performance evaluation

To evaluate estimator performance, we computed bias, precision, and accuracy (Walther & Moore, 2005). We assessed overall performance by combining all simulations (i.e. all factors and all factor levels) and performance for a specific factor level by combining the replicates of all remaining factors and factor levels. Estimator bias was evaluated with scaled mean error (SME; Walther & Moore, 2005)

$$SME = \frac{1}{X} \sum_{j=1}^{X} \left(\frac{\hat{S}_{est(j)} - S_{true}}{S_{true}} \right)$$

where X is the number of replicates across combined factors, j is the replicate, $j = 1, 2, \ldots, X$, and $\hat{S}_{est}$ is an estimate of S_{true}. Negative and positive values of SME indicate average underestimation and overestimation, respectively, and SME = 0 indicates an unbiased estimator. We evaluated estimator precision with the sample standard deviation of a group of scaled estimates (SD):

$$SD = \sqrt{ \frac{ \sum_{j=1}^{X} \left(\frac{\hat{S}_{est(j)}}{S_{true}} - \frac{\overline{\hat{S}_{est}}}{S_{true}} \right)^2 }{X - 1} }$$

where $\hat{S}_{est}/S_{true}$ is the scaled estimate of the *j*th replicate, $j = 1, 2, \ldots, X$. We scaled the estimates in order to compare performances across the three levels of S_{true}. For accuracy, we used scaled mean square error (SMSE; Walther & Moore, 2005):

$$SMSE = \frac{1}{X} \sum_{j=1}^{X} \left(\frac{\hat{S}_{est(j)} - S_{true}}{S_{true}} \right)^2 .$$

Values of SD and SMSE are always positive and, as with SME, those closer to zero indicate better performance.

We treated all factors as random and evaluated relative factor effects with random-effects models, using the 'proc mixed' procedure in SAS (version 9.3; SAS Institute, 1999), by computing variance components and proportionally allocating total variance into S_{true}, *N*, *Abund*, *Config*, *p*, *Effort* and *Design*. By using the averages (SME and SMSE) and standard deviations (SD) of the replicates for each unique factor combination, the variance components only reflect between-factor effects. We considered only main effects, so the residual represents the variance explained by all interactions plus residual error.

We also evaluated performance against *sc*, which is possibly the most important factor driving estimator performance (Baltanás, 1992; Brose *et al.*, 2003). We used *sc* to group simulations into 10 equally sized bins (i.e. $0.0 < sc \leq 0.1$, $0.1 < sc \leq 0.2, \ldots, 0.9 < sc \leq 1.0$). Estimator performance was averaged over each coverage range, which thereby assesses performance at various levels of sampling completeness. Except when data are simulated, *sc* is usually not known. Brose *et al.* (2003) estimated *sc* by dividing S_{obs} by the average SR estimate.

Table 3 Average performance of species richness estimators across all factors (see Table 1) based on bias, measured as scaled mean error (SME); precision, measured as standard deviation of scaled estimates (SD); and accuracy, measured as scaled mean square error (SMSE). Subscripts are estimator rank for a given performance measure.

Estimator	Bias (SME)	Precision (SD)	Accuracy (SMSE)
ACE	-0.32_7	0.26_4	0.17_4
Boot	-0.49_{12}	0.27_6	0.31_{11}
Chao1	-0.37_{10}	0.26_2	0.20_8
Chao2	-0.37_9	0.25_1	0.20_7
CY-1	-0.19_3	0.30_9	0.13_1
CY-2	-0.15_1	0.34_{11}	0.14_2
ICE	-0.32_6	0.26_5	0.17_3
Jack1	-0.41_{11}	0.28_7	0.25_{10}
Jack2	-0.33_8	0.29_8	0.19_6
Jack3	-0.27_5	0.33_{10}	0.18_5
Jack4	-0.22_4	0.41_{12}	0.22_9
Jack5	-0.18_2	0.59_{13}	0.38_{13}
Mixture	0.64_{14}	13.41_{14}	180.22_{14}
S_{obs}	-0.56_{13}	0.26_3	0.38_{12}

We estimated sc ($\widehat{sc}$) by averaging the Jaccard similarity coefficients across the 100 iterations used to calculate CY-1.

RESULTS

Simulations

The Mixture estimator failed to converge (i.e. it failed to produce an estimate) in 0.19% of the simulations and in as many as five replicates (i.e. simulations with identical factor levels). A small number of simulations (0.39%) resulted in less than the number of encounters needed to compute CY-1 and CY-2, so those estimates equalled zero. There were also simulations where one or more of the estimators returned an estimate $< S_{obs}$ including CY-2 (0.45%), Jack2 (3.60%), Jack3 (7.73%), Jack4 (13.45%) and Jack5 (18.10%). We removed simulations in which Mixture did not converge and rebalanced the factorial by removing five or more replicates from every factor combination (i.e. resulting in 37 replicates of every factor combination); when possible, we removed replicates that also had one or more estimates $< S_{obs}$.

Averaged over all simulations, Mixture was the only positively biased estimator and the only nonparametric estimator more biased than S_{obs} (Table 3). The least biased estimator, CY-2, underestimated S_{true} by 15% on average. CY-1 was the most accurate estimator. Estimator rank depended on the evaluation metric and we noted a general trade-off between bias and precision (i.e. the least biased estimators were the least precise, and vice versa). However, Mixture was both the most biased and the least precise estimator (Table 3). Relative ranks were rarely consistent across factor levels, but estimators that performed best overall were generally among the best at individual factor levels. Below, we mainly focus on factor-level comparisons; additional estimator-specific results are presented in Reese (2012).

Except for the bias of Mixture and the accuracy of higher-order jackknife estimators (Jack3–Jack5), bias and accuracy worsened with increasing S_{true} (Table S1.1 in Supporting Information). The relationship between S_{true} and estimator precision was much less consistent, with numerous instances of both positive and negative relationships. Only three estimators, S_{obs}, Boot and Jack1, ranked consistently across all factor levels and only for precision. The greatest change in the relative performance of an estimator was the bias of Mixture which was the least biased estimator with $S_{true} = 100$ and one of the two most biased estimators with the other factor levels.

Variation in the other assemblage factors also affected estimator performance. Except for Jack5 and Mixture, estimator bias and accuracy improved with increasing evenness (i.e. in moving from log-series to log-normal distributions) and all estimators were more precise with log-series distributions (Table S1.2). The relative bias of Mixture was inconsistent, ranking best with log-series distributions and worst with log-normal distributions. By all measures, the nonparametric estimators performed better at the larger N level (Table S1.3). The precision of S_{obs}, however, decreased slightly with the increase in N. Mixture was the most biased estimator at $N = 6250$ and the least biased at $N = 12,500$. Performances varied little across configuration patterns, but estimators were generally least biased and least precise when individuals were hyperdispersed (Table S1.4). The nonparametric estimators performed better when average species detection probability ($\overline{p}$) equalled 0.9 than when $\overline{p} = 0.5$ (Table S1.5). When $\overline{p}$ increased with ranked abundance, estimators were more precise and generally more biased and more accurate than when $\overline{p}$ decreased with abundance. Mixture was again either the most or least biased estimator, depending on the factor level.

The effects of the tested survey factors, *Effort* and *Design*, were relatively consistent. The nonparametric estimators performed better at the larger *Effort* level (Table S1.6). Only the precision of S_{obs} decreased with the increase in *Effort*. Estimators were generally, often to a small degree, less biased, more precise and more accurate when individual survey locations were selected randomly than when they were placed along randomly located linear transects (Table S1.7).

Averaged across all estimators, S_{true}, *Effort* and *Abund* had the largest effects on estimator performance (Table S2). The results were generally similar with individual estimators; however, N or p occasionally ranked higher. Main effects accounted for approximately 87%, 71% and 81% of the variation in bias, precision and accuracy, respectively, but explained $< 50\%$ of the variation in the performance of Mixture.

We also evaluated estimator performance against sample coverage (i.e. sc). Estimators were generally less biased and more accurate with a larger sc. There was no apparent trend in precision. Our results suggest that the best estimator for bias reduction depends largely on the sc range; only Jack5 is recommended first in more than one sc range (Table 4). By contrast, S_{obs} and Boot were among the three most precise estimators for all coverage ranges. CY-1 and Boot were the most accurate estimators in the two smallest and two largest coverage ranges, respectively, and one of the jackknife estimators was most accurate in the

Table 4 The three best performing species richness estimators, averaged over all simulations (without the particulate-niche distribution), in the specified sample coverage (*sc*) range (the number of simulations in each range is given in parentheses). See Table 2 for estimator abbreviations and Table 3 for specific performance metrics.

Coverage range	Bias (SME)			Precision (SD)			Accuracy (SMSE)		
	1st	2nd	3rd	1st	2nd	3rd	1st	2nd	3rd
$0.0 < sc \leq 0.1$ (1891)	−0.30	−0.32	−0.56	0.02	0.02	0.03	0.21	0.30	0.39
	CY-2	CY-1	ICE	S_{obs}	Boot	Jack1	CY-1	CY-2	ICE
$0.1 < sc \leq 0.2$ (2264)	−0.25	−0.29	−0.51	0.03	0.04	0.05	0.24	0.28	0.29
	CY-1	CY-2	Jack5	S_{obs}	Boot	Jack1	CY-1	CY-2	Jack5
$0.2 < sc \leq 0.3$ (3209)	−0.11	−0.30	−0.30	0.03	0.04	0.05	0.17	0.18	0.19
	Mixture	CY-1	CY-2	S_{obs}	Boot	Jack1	Jack4	CY-2	Jack5
$0.3 < sc \leq 0.4$ (3451)	−0.22	−0.27	−0.27	0.03	0.04	0.07	0.14	0.15	0.15
	Jack5	Jack4	CY-2	S_{obs}	Boot	Jack1	Jack3	Jack4	CY-2
$0.4 < sc \leq 0.5$ (2719)	−0.07	−0.13	−0.17	0.03	0.04	0.07	0.07	0.08	0.09
	Jack5	Jack4	CY-2	S_{obs}	Boot	Jack1	Jack3	Jack2	CY-2
$0.5 < sc \leq 0.6$ (2563)	−0.04	−0.09	−0.09	0.03	0.04	0.08	0.06	0.06	0.08
	Jack5	Jack4	CY-2	S_{obs}	Boot	Jack1	Jack2	CY-2	CY-1
$0.6 < sc \leq 0.7$ (1313)	0.03	−0.03	−0.04	0.02	0.04	0.08	0.02	0.03	0.04
	Jack3	CY-1	Jack2	S_{obs}	Boot	Jack1	Jack1	Jack2	CY-1
$0.7 < sc \leq 0.8$ (1304)	−0.03	0.05	−0.06	0.03	0.04	0.08	0.01	0.02	0.02
	Jack1	Jack2	ICE	S_{obs}	Boot	Jack1	Jack1	Boot	ICE
$0.8 < sc \leq 0.9$ (1065)	−0.02	−0.03	0.03	0.02	0.04	0.08	0.00	0.01	0.01
	ICE	ACE	Jack1	S_{obs}	Boot	Jack1	Boot	Jack1	ACE
$0.9 < sc \leq 1.0$ (1534)	0.00	0.00	0.00	0.03	0.03	0.04	0.00	0.00	0.00
	Chao2	Chao1	ACE	S_{obs}	Mixture	Boot	Boot	Mixture	S_{obs}

SME, scaled mean error; SD, standard deviation of scaled estimates; SMSE, scaled mean square error.

other ranges. The CY-2 estimator was one of the three most accurate estimators in the largest number of *sc* ranges (six) and CY-1 was next best (four).

DISCUSSION

The case for statistical estimators

Using S_{obs} as an estimate of SR frequently leads to large underestimates, due to a strong dependence on sampling effort and the assumption that all species are detected (Nichols *et al.*, 1998; Kéry & Plattner, 2007). Other than Mixture, the nonparametric estimators were generally less biased and more accurate than S_{obs} (Table 3), supporting other studies (see Wagner & Wildi, 2002; Brose *et al.*, 2003; Walther & Moore, 2005). Thus, S_{obs} is far from the best approach for estimating SR. As effort increases, S_{obs} will approach S_{true}, but estimators generally provide a more reliable approach (see Table S1.6). A measure of precision is another advantage provided by some estimators (Nichols *et al.*, 1998; Reese, 2012).

Statistical estimator comparisons

Averaged across all factor levels, our results suggest that ACE, CY-1, CY-2, ICE and Jack3 can provide less biased and more accurate estimates than the more popular Chao1, Chao2, Jack1 and Jack2 estimators (Table 3). The higher-order jackknife and bias-corrected Chao estimators were among the least and intermediately biased estimators, respectively, supporting Brose *et al.* (2003). In particular, the overall biases of Jack5, CY-1 and CY-2

were less than the ≥ 20% bias reported for other estimators (see Brose *et al.*, 2003; Canning-Clode *et al.*, 2008; Jobe, 2008). Bias reduction involves increased extrapolation, which could partly explain the loss of precision that accompanied the least biased estimators. As was found in our study, estimation involves a tradeoff between bias and precision (see Burnham & Overton, 1979; Brose *et al.*, 2003; Willie *et al.*, 2012).

Among the most accurate and least biased estimators at many factor levels, CY-1 and CY-2 could be promising newer estimators (Table 3, Appendix S1). It is noteworthy that both performed at their best with relatively uneven log-normal and log-series distributions (Table S1) which might best approximate true species abundance patterns (Sugihara, 1980; Ulrich *et al.*, 2010). However, CY-1 and CY-2 were outperformed at some factor levels (see Appendix S1). Cao *et al.* (2004) stated that sample size would affect CY-2 less than other estimators and it was the least biased estimator at the smaller effort level (Table S1.6). However, CY-1 was more accurate and less affected (bias) by variation in *Effort* than any other estimator (Table S2; see also Drake, 2007). It is possible that the number of regression points used in the program SIMASSEM, five or ten, does not fully expose the performance potential of CY-2. Both estimators were relatively imprecise and prone to overestimation with particulate-niche distributions (Table S1) and larger levels of N, *Effort* and *p* (Appendix S1), factor levels generally associated with larger *sc* values.

Another largely untested estimator, Mixture, often performed relatively poorly. However, based on bias, Mixture ranked more favourably in assemblages with an intermediate number of species (Table S1.1), log-series distributions (Table S1.2), larger

N (Table S1.3) and either a larger $\bar{p}$ or a $\bar{p}$ that increased with abundance groups (Table S1.5). Precision improved considerably when effort increased from 100 to 500 cells (Table S1.6). These results suggest that some factor combinations did not provide enough data to estimate probabilities for the two groups of the mixture model. An inspection of the raw results (G.C.R., unpublished data) showed that the poorest performances resulted from an increased number of large estimates. In other words, Mixture occasionally produced large estimates (e.g. 211,111) in assemblages with, for example, $S_{true} = 500$, $Abund$ = log-normal, $N = 6250$, and $\bar{p} = 0.5$, possibly a result of poor convergence. The conditions under which the performance of Mixture improved would indicate that it should be further evaluated in comparatively data-rich environments.

Factor effects

The factors S_{true}, $Effort$, and $Abund$ had the largest effects on estimator performance, supporting Brose et $al.$ (2003), and the strongest correlations with sc (i.e. the correlation between sc and $S_{true} = -0.66$, $Effort = 0.48$, $Abund = 0.37$, $N = 0.18$, $P = 0.04$, $Config = 0.01$ and $Design = 0.01$); note that the sign of correlation is irrelevant for the nominal variables $Abund$, $Config$, $Design$ and p. This suggests that sc is the means by which factors affect performance and supports Brose et $al.$ (2003). The negative correlation that we found between S_{true} and sc, however, contradicts the positive and relatively weak correlation (i.e. $r = 0.14$) found by Brose et $al.$ (2003).

The negative bias of most estimators was largest with the relatively uneven log-series distribution (Table S1), which supports Wagner & Wildi (2002) and Brose et $al.$ (2003). Wagner & Wildi (2002) found less bias with log-normal distributions than with the distribution selected for its relative evenness, the broken-stick distribution. However, we did not find a significant difference between simulated broken-stick and log-normal distributions using Kolmogorov–Smirnov goodness of fit tests (G.C.R., unpublished analysis; for the test see Magurran, 2004), which could explain our contradictory result when using particulate-niche distributions instead (Table S1). Previous comparisons, therefore, might not have included enough variation to reveal actual relationships between estimator bias and assemblage evenness.

At larger values of S_{true}, estimators were generally more biased, supporting Baltanás (1992). Brose et $al.$ (2003) reported a contradictory result where estimators were more biased at $S_{true} = 25$ than at $S_{true} = 500$; however, it is unclear whether that relationship held across their other levels of S_{true} (see their Fig. 1). We also found that estimators were less accurate at larger values of S_{true} which contradicts Walther & Morand (1998; see Table S1.1). Better performances at the larger $Effort$ level, which supports Brose et $al.$ (2003) and Wagner & Wildi (2002), could be caused by more effort resulting in more data, particularly encounters (see Table S1.6). The number of encounters can also be increased with larger values of N and p. Estimator performance improved when N was increased from 6250 to 12,500

individuals (Table S1.3). Positive relationships have been similarly reported with N in the form of density (Baltanás, 1992; Walther & Morand, 1998).

When the two levels where $\bar{p}$ was a function of ranked abundance were removed, the correlation between sc and p increased from $r^2 = 0.04$ to $r^2 = 0.18$. When $\bar{p}$ generally decreased with abundance, estimators were less biased (there were exceptions with Boot, Mixture and S_{obs}), though often with less precision and accuracy, than when $\bar{p}$ increased with abundance (see Table S1.5). Less abundant species, particularly when p is small, are often not detected, but a larger p could result in such species being encountered, thereby reducing bias. Given the formulae of many of the nonparametric estimators, bias will be further reduced when such species are encountered in a small number of surveys. Using either a fixed p for each abundance group or larger differences in $\bar{p}$ might show more consistent trends in bias and accuracy across estimators.

Estimator performance was little affected by $Config$, supporting both Wagner & Wildi (2002) and Brose et $al.$ (2003). The three configuration patterns differed considerably for a single species, but at the assemblage level the differences were minimal. This emergent property could partly explain the relatively weak effects that $Config$ had on estimator performance and the effects of an assemblage-wide configuration pattern could be greater. Most estimators were more negatively biased with increased aggregation, which supports Baltanás (1992, their Fig. 3), but not Wagner & Wildi (2002). Furthermore, we did not find the positive relationship between accuracy and aggregation reported by Walther & Morand (1998).

Survey design also had small effects on performance. Its importance could be greater if, for example, SR varied along one or more gradients and the orientation of linear transects failed to fully represent an area. We are unaware of other investigations of the relationships between survey design and SR estimates, but differences between survey designs are important to other estimation issues (see Reese et $al.$, 2005), including practicality due to logistics when implementing sampling regimes.

Estimator selection framework

Near the boundaries of sc, we found that accuracy can be maximized with different estimators than those proposed in Brose et $al.$ (2003), specifically CY-1 when $sc \leq 20\%$ and Boot when $sc > 80\%$ (Table 4). We also found the best performing estimator to vary between bias, precision and accuracy; therefore selection should be application specific. Our selection framework and the one by Brose et $al.$ (2003) are based on sc (i.e. S_{obs}/S_{true}) a quantity that, if known, would make estimation unnecessary. Across our simulations, the correlation between sc and $\hat{sc}$ exceeded 0.80. Thus, we recommend using the program SimAssem to process data and the reported $\hat{sc}$ to locate the coverage range of suggested estimators. Given the inconsistent performance of Mixture, we recommend caution when considering it for bias reduction (see Table 4 and Appendix S1). We also tested our selection framework against field data and present the results in Appendix S3.

Statistical estimation issues

Our results included estimates that were $< S_{obs}$, even negative, and others that were unrealistically large. We therefore evaluated the effects of sensible substitutions. When the count associated with one or more of the negative terms of a jackknife estimator is large, the final estimate can be $< S_{obs}$ (Lam & Kleinn, 2008). Otis *et al.* (1978) considered the observed number, S_{obs} in species richness applications, a reasonable lower confidence interval boundary because at least that many species are known to exist. We tested three approaches when a jackknife estimate was $< S_{obs}$. First, an estimate $< S_{obs}$ was set equal to S_{obs}. Second, we substituted the next lower-order jackknife estimate that was $\geq S_{obs}$. Third, we substituted the larger of S_{obs} or the largest jackknife estimate, regardless of its order. Burnham & Overton (1978) introduced a statistical test for selecting a jackknife estimator that was not tested here (but see Brose *et al.*, 2003). Of the three approaches, we found that substituting the larger of S_{obs} or the largest jackknife estimate resulted in the best performance and therefore recommend that this approach be considered when using the jackknife estimators (see Appendix S2).

There were also simulations where encounters occurred in too few surveys for the computation of CY-1 and CY-2 and the CY-2 estimate was negative in other runs (0.14%), demonstrating the limitations of these estimators. A CY-2 estimate can be $< S_{obs}$ when $\overline{SR}$ and JC are negatively related, a chance relationship when data are repeatedly randomized. In both cases, we tested the effect of replacing the non-negative or negative estimate with S_{obs}. In simulations where CY-2 failed to produce an estimate, Mixture always either failed to converge or gave an unreasonably large estimate [e.g. $\geq (8059 \times S_{obs})$], providing further evidence that sparse data affected estimates. Systematically varying both the number of surveys and the number of encounters will probably be required to define the thresholds at which valid Mixture estimates become possible. Relative to S_{obs}, CY-2 returned the next largest estimate (approximately $S_{obs} \times$ 76) which we used as a proportionality threshold above which Mixture estimates were replaced with S_{obs}. These substitutions had minimal effects on the performances of CY-1 and CY-2, but occasionally dramatic effects on the performance of Mixture (Appendix S2).

Implications

Our study suggests that differences between assemblage characteristics and survey designs partly explain reported differences in estimator performance. Inconsistencies in the use of performance metrics almost certainly further complicate comparisons between studies (see Walther & Moore, 2005). Despite simulating a wide range of conditions, additional real-world factors and factor levels will undoubtedly complicate estimation efforts; therefore, the application of our results will be most beneficial when combined with simulations using programs such as SimAssem (Reese *et al.*, 2013).

The most biased estimators were generally the most precise. This is a particularly dangerous combination because a precise and biased estimator can lead to a false sense of confidence when compared with an imprecise, but unbiased, estimator. In our study, precision was based on empirical estimates of variance across the replicated simulations, but a single SR estimate is of limited value when not accompanied by some measure of its reliability. Although rarely reported in the literature, SR variance estimators have been derived for many of the SR estimators and their performance should also be considered when selecting a species richness estimator (Reese, 2012).

REFERENCES

Arrhenius, O. (1921) Species and area. *Journal of Ecology*, **9**, 95–99.

Baltanás, A. (1992) On the use of some methods for the estimation of species richness. *Oikos*, **65**, 484–492.

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.

Burnham, K.P. & Overton, W.S. (1978) Estimation of the size of a closed population when capture probabilities vary among animals. *Biometrika*, **65**, 623–633.

Burnham, K.P. & Overton, W.S. (1979) Robust estimation of population size when capture probabilities vary among animals. *Ecology*, **60**, 927–936.

Canning-Clode, J., Valdivia, N., Molis, M., Thomason, J.C. & Wahl, M. (2008) Estimation of regional richness in marine benthic communities: quantifying the error. *Limnology and Oceanography: Methods*, **6**, 580–590.

Cao, Y., Larsen, D.P. & Hughes, R.M. (2001) Estimating total species richness in fish assemblage surveys: a similarity based approach. *Canadian Journal of Fisheries and Aquatic Sciences*, **58**, 1782–1793.

Cao, Y., Larsen, D.P. & White, D. (2004) Estimating regional species richness using a limited number of survey units. *Ecoscience*, **11**, 23–35.

Chao, A. (1984) Non-parametric estimation of the number of classes in a population. *Scandinavian Journal of Statistics*, **11**, 265–270.

Chao, A. (1987) Estimating the population size for capture–recapture data with unequal catchability. *Biometrics*, **43**, 783–791.

Chao, A. & Lee, S.M. (1992) Estimating the number of classes via sample coverage. *Journal of the American Statistical Association*, **87**, 210–217.

Chiarucci, A., Enright, N.J., Perry, G.L.W. & Miller, B.P. (2003) Performance of nonparametric species richness estimators in a high diversity plant community. *Diversity and Distributions*, **9**, 283–295.

Colwell, R.K. & Coddington, J.A. (1994) Estimating terrestrial biodiversity through extrapolation. *Philosophical Transactions of the Royal Society B: Biological Sciences*, **345**, 101–118.

Dallmeier, F., Foster, R.B., Romano, C., Rice, R. & Kabel, M. (1991) *A user's guide to the Beni Biosphere Reserve biodiversity plots*. Smithsonian Institution, Washington, DC.

Deblauwe, V., Barbier, N., Couteron, P., Lejeune, O. & Bogaert, J. (2008) The global biogeography of semi-arid periodic vegetation patterns. *Global Ecology and Biogeography*, **17**, 715–723.

Drake, M.T. (2007) Estimating sampling effort for biomonitoring of nearshore fish communities in small central Minnesota lakes. *North American Journal of Fisheries Management*, **27**, 1094–1111.

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.

Holdridge, L.R., Grenke, W.C., Hatheway, W.H., Liang, T. & Tosi, J.A. (1971) *Forest environments in tropical life zones*. Pergamon Press, Oxford.

Jobe, R.T. (2008) Estimating landscape-scale species richness: reconciling frequency- and turnover-based approaches. *Ecology*, **89**, 174–182.

Keating, K.A. & Quinn, J.F. (1998) Estimating species richness: the Michaelis–Menten model revisited. *Oikos*, **81**, 411–416.

Kéry, M. & Plattner, M. (2007) Species richness estimation and determinants of species detectability in butterfly monitoring programmes. *Ecological Entomology*, **32**, 53–61.

Laake, J. & Rexstad, E. (2008) RMark – an alternative approach to building linear models in MARK. *Program mark: 'a gentle introduction'*, 7th edn (ed. by E. Cooch and G.C. White), pp. C1–C115. Available at: http://www.phidot.org/software/mark/docs/book/ (accessed 24 October 2013).

Lam, T.Y. & Kleinn, C. (2008) Estimation of tree species richness from large area forest inventory data: evaluation and comparison of jackknife estimators. *Forest Ecology and Management*, **255**, 1002–1010.

Lee, S.M. & Chao, A. (1994) Estimating population size via sample coverage for closed capture–recapture models. *Biometrics*, **50**, 88–97.

Legendre, P. (1993) Spatial autocorrelation: trouble or new paradigm? *Ecology*, **74**, 1659–1673.

Lewis, T. & Taylor, L.R. (1967) *Introduction to experimental ecology*. Academic Press, London.

Lopez, L.C.S., Fracasso, M.P.D.A., Mesquita, D.O., Palma, A.R.T. & Riul, P. (2012) The relationship between percentage of singletons and sampling effort: a new approach to reduce the bias of richness estimates. *Ecological Indicators*, **14**, 164–169.

MacArthur, R.H. (1957) On the relative abundance of bird species. *Proceedings of the National Academy of Sciences USA*, **43**, 293–295.

Magurran, A.E. (2004) *Measuring biological diversity*. Blackwell Publishing, Malden, MA.

Michaelis, M. & Menten, M.L. (1913) Der Kinetic der Invertinwirkung. *Biochemische Zeitschrift*, **49**, 333–369.

Moreno, C., Zuria, I., García-Zenteno, M., Sánchez-Rojas, G., Castellanos, I., Martínez-Morales, M. & Rojas-Martínez, A. (2006) Trends in the measurement of alpha diversity in the last two decades. *Interciencia*, **31**, 67–71.

Nichols, J.D., Boulinear, T., Hines, J.E., Pollock, K.H. & Sauer, J.R. (1998) Inference methods for spatial variation in species richness and community composition when not all species are detected. *Conservation Biology*, **12**, 1390–1398.

Otis, D.L., Burnham, K.P., White, G.C. & Anderson, D.R. (1978) Statistical inference from capture data on closed animal populations. *Wildlife Monographs*, **62**, 1–135.

Pagano, A.M. & Arnold, T.W. (2009) Estimating detection probabilities of waterfowl broods from ground-based surveys. *Journal of Wildlife Management*, **73**, 686–694.

Palmer, M.W. (1990) The estimation of species richness by extrapolation. *Ecology*, **71**, 1195–1198.

Pledger, S. (2000) Unified maximum likelihood estimates for closed capture–recapture models using mixtures. *Biometrics*, **56**, 434–442.

Preston, F.W. (1948) The commonness, and rarity, of species. *Ecology*, **29**, 254–283.

R Development Core Team (2009) *R: a language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. Available at: http://www.r-project.org (accessed 24 October 2013).

Reese, G.C. (2012) *Simulating species assemblages and evaluating species richness estimators*. PhD thesis, Colorado State University, Fort Collins. Available at: http://hdl.handle.net/10217/68193 (accessed 24 October 2013).

Reese, G.C., Wilson, K.R., Hoeting, J.A. & Flather, C.H. (2005) Factors affecting species distribution predictions: a simulation modeling experiment. *Ecological Applications*, **15**, 554–564.

Reese, G.C., Wilson, K.R. & Flather, C.H. (2013) Program SimAssem: software for simulating species assemblages and estimating species richness. *Methods in Ecology and Evolution*, **4**, 891–896.

SAS Institute (1999) *SAS/STAT user's guide, version 8*. SAS Institute, Cary, NC.

Selmi, S. & Boulinier, T. (2004) Distribution–abundance relationship for passerines breeding in Tunisian oases: test of the sampling hypothesis. *Oecologia*, **139**, 440–445.

Smith, E.P. & van Belle, G. (1984) Nonparametric estimation of species richness. *Biometrics*, **40**, 119–129.

Sugihara, G. (1980) Minimal community structure: an explanation of species abundance patterns. *The American Naturalist*, **116**, 770–787.

Tjørve, E. (2009) Shapes and functions of species–area curves (II): a review of new models and parameterizations. *Journal of Biogeography*, **36**, 1435–1445.

Ulrich, W., Ollik, M. & Ugland, K.I. (2010) A meta-analysis of species-abundance distributions. *Oikos*, **119**, 1149–1155.

Wagner, H.H. & Wildi, O. (2002) Realistic simulation of the effects of abundance distribution and spatial heterogeneity on non-parametric estimators of species richness. *Ecoscience*, **9**, 241–250.

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.

Walther, B.A. & Morand, S. (1998) Comparative performance of species richness estimation methods. *Parasitology*, **116**, 395–405.

White, G.C. & Burnham, K.P. (1999) Program MARK: survival estimation from populations of marked animals. *Bird Study*, **46**, Supplement, 120–138.

Williams, C.B. (1939) An analysis of four years' captures of insects in a light trap. Part I. *Transactions of the Royal Entomological Society of London*, **89**, 79–132.

Willie, J., Petre, C.A., Tagg, N. & Lens, L. (2012) Evaluation of species richness estimators based on quantitative performance measures and sensitivity to patchiness and sample grain size. *Acta Oecologica*, **45**, 31–41.

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of this article at the publisher's web-site.

Table S1 Performance of species richness estimators as a function of species abundance distribution, log-normal, log-series and particulate-niche, averaged across all factor levels.

Table S2 A variance component analysis of relative factor effects.

Appendix S1 Estimator bias, precision and accuracy across factor levels (Tables S1.1–S1.7).

Appendix S2 Estimator bias, precision and accuracy across factor levels, after accounting for estimation issues (Tables S2.1–S2.7).

Appendix S3 Case study of the performance of our proposed selection framework and the asymptotic properties of estimators.

BIOSKETCHES

Gordon C. Reese recently completed his PhD in the Graduate Degree Program in Ecology at Colorado State University. He is researching the factors affecting the performance of species richness estimators and also the relative performance of several approaches for estimating the variance of species richness estimators.

Kenneth R. Wilson is a professor in the Department of Fish, Wildlife, and Conservation Biology and Graduate Degree Program in Ecology at Colorado State University. He has an interest in wildlife management, conservation biology and ecology specifically related to: (1) impacts of human activities on wildlife, (2) population ecology, and (3) understanding patterns of species richness.

Curtis H. Flather is a research ecologist with the US Forest Service, Rocky Mountain Research Station in Fort Collins, CO. His research programme focuses on estimating the response of species assemblages to natural and human-caused disturbance to support resource planning within the agency.

Editor: Gary Mittelbach