Can subsets of species indicate overall patterns in biodiversity?

NIKOLAI KLIBANSKY [iD],† KYLE W. SHERTZER, G. TODD KELLISON, AND NATHAN M. BACHELER

Southeast Fisheries Science Center, National Marine Fisheries Service, NOAA, 101 Pivers Island Road, Beaufort, North Carolina 28516 USA

Abstract.   Resource limitations often allow only a subset of species to be counted. But using subsets may bias inferences on spatial or temporal trends in biodiversity. Using data from a video survey on reefs in the Gulf of Mexico for which all fish species observed were counted (243 species), we investigated how the use of reduced species lists (RSLs) can impact perceived patterns in biodiversity. We estimated four common biodiversity metrics (species richness, and Margalef's, Shannon's, and Simpson's indices) at each of 2115 sampling locations, using the total species list and RSLs. For all diversity metrics, correlations between estimates using the total species list and RSLs increased with the number of species in the list. Using a bootstrap approach, we randomly generated hypothetical lists equal in length to each empirical RSL to evaluate their performance; empirical RSLs tended to perform similar to random lists of equivalent length when estimating species richness or Margalef's index, and tended to outperform most hypothetical RSLs when estimating Shannon's and Simpson's indices. To understand how to create better performing RSLs, we extended the bootstrap analysis to select RSLs of all possible lengths, using four different selection methods related to species commonness; the functional relationships between correlation and number of species in an RSL were similar among metrics but were very different among selection methods. With each hypothetical RSL, we tested common biodiversity hypotheses such as relationships with depth and latitude and compared the outcomes with the best estimate of true relationships identified using the total list. Longer lists comprised of the most common species more often identified the true relationship, but results showed complex patterns. Many short lists of the most common species yielded results opposite the true relationships, and many lists of intermediate length failed to identify any relationship while the total list showed a significant trend. Overall, these analyses show that sampling methods used for biodiversity studies should be as unselective as possible, and datasets based on more selective methods should be interpreted carefully and should not be expected to reflect true patterns in biodiversity.

Key words: biodiversity; reef fish; species selectivity; video survey.

† **E-mail:** nikolai.klibansky@noaa.gov

INTRODUCTION

Simple measures of species diversity have long been used in community ecology, often to test hypotheses of temporal and spatial trends (Dobzhansky 1950, Stevens 1989, Willig et al. 2003). In this context, one must assume that species are equally detectable for all sampling units (Boulinier et al. 1998), and thus, it is advisable to use consistent sampling techniques and equipment when making direct comparisons of diversity estimates. In practice, this can be difficult to achieve, as the detectability of a species is often dependent on the sampling method. For

example, bird surveys conducted during the middle of the day will tend to under-represent crepuscular species (Bibby et al. 1998), pitfall traps used to collect spiders may over-represent very active species (Topping and Sunderland 1992), and highly clumped plant species will tend to appear uncommon in quadrat sampling (Buckland et al. 2007).

Video surveys are becoming increasingly common in biological studies, particularly in marine environments (Mallet and Pelletier 2014). Often these surveys are designed for estimating the abundance of individual species, but the data can also be valuable for biodiversity research. Video cameras offer a relatively unbiased view of fish assemblages, compared to other gear types (e.g., Parker et al. 2016), and are able to collect large amounts of visual data rather quickly. In systems where individuals and species are abundant, such as marine reefs, observations may be collected at such a high rate that subsampling is necessary. For various taxa, automated species identification software is developing rapidly, allowing many individuals to be quickly identified from video or still images. But current programs can often identify only subsets of species (e.g., White et al. 2006, Aguzzi et al. 2009, Kumar et al. 2012, Yu et al. 2013); thus, the number of datasets containing counts of only subsets of the species present are likely to grow. In video surveys in the Gulf of Mexico and Atlantic coast of the Southeast United States (SEUS) designed to estimate fish abundance, resource limitations dictate that data be subsampled with respect to both the number of video frames analyzed, and the number of species identified and enumerated. While the effect of sampling fewer frames on biodiversity estimates has been studied (Bacheler and Shertzer 2015), it is not clear how limiting data analysis to a subset of priority species affects the value of these datasets for biodiversity research.

Using a large video survey dataset from the Gulf of Mexico, for sites where all fish species observed were identified and counted, we address three main questions: (1) How does the number of species included in a reduced species list affect the ability of the dataset to predict the true diversity? (2) How do different patterns of species selection affect list performance? (3) How does the species selection method affect perceived patterns in biodiversity? By answering

these questions, we provide insight into how species selectivity of a sampling method affects the value of a dataset to biodiversity research.

MATERIALS AND METHODS

Data collection

The Gulf of Mexico Video Survey (GOMVS) dataset utilized in this study was produced by the combined efforts of the Southeast Area Monitoring and Assessment Program (SEAMAP) Reef Fish Survey and the Gulf of Mexico Marine Protected Area (MPA) Reef Fish Survey (www.sefsc.noaa.gov). Methods of data collection and video analysis were presented in detail by Campbell et al. (2015) and are summarized below.

The surveys covered a broad area of the Gulf of Mexico, from the US–Mexico border in Texas to the Dry Tortugas islands in Florida, spanning depths of 9–110 m (Fig. 1). The GOMVS dataset included SEAMAP data from the periods 1995 to 1997 and 2001 to 2007 and MPA data from 2001 to 2007. Sampling locations were selected with a two-stage sampling design, where first stratified random sampling was used to select among 10-by-10-min fixed sampling blocks, and then, sampling sites were randomly selected from previously identified reef habitat within each block. At each site, a four-video-camera array (Sony VX2000 DCR digital camcorders, Sony Corporation of America, New York, New York, USA) or a chevron-shaped fish trap with one mounted video camera was baited with squid and deployed during daylight hours. Video cameras were positioned approximately 30 cm above the bottom and recorded video for at least 20 min. In sites with four-video-camera arrays, one video was randomly chosen from among the videos facing reef habitat. Experienced technicians viewed each video for 20 min (beginning from the time that the silt plume raised by the trap landing on the sea bottom cleared), recorded times that each individual fish entered and exited the video frame, and identified each to the lowest taxonomic level.

Analysis

Data from the GOMVS were filtered to include only videos for which all taxa observed on frame were identified and counted, and which were read for exactly 20 min. Only one video was analyzed per site. Though other taxa were sometimes
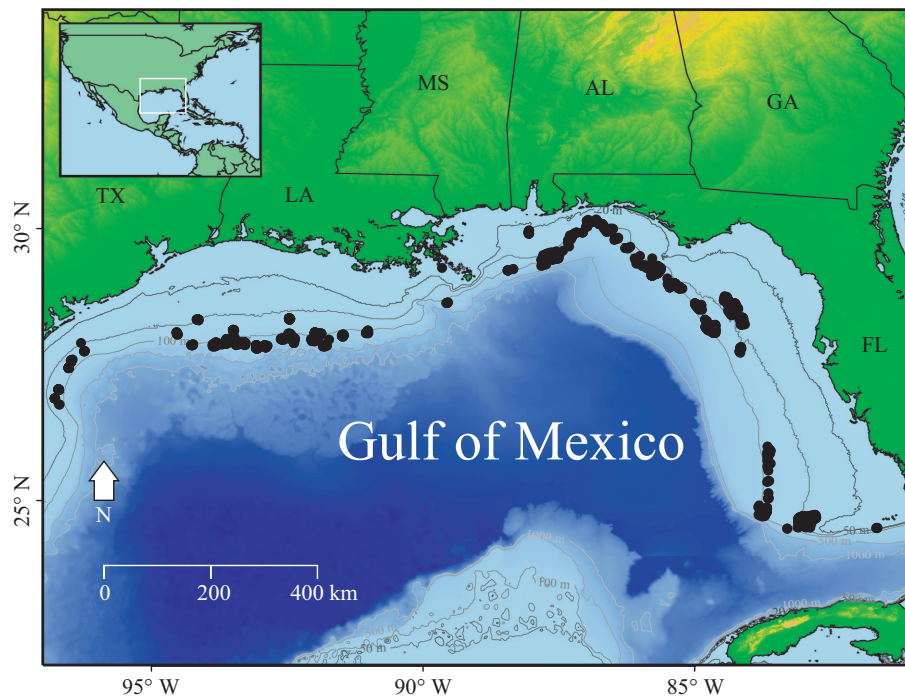
Fig. 1. Map of all 2115 sampling sites (filled circle) in the Gulf of Mexico Video Survey dataset. Note that many points overlap. Darker shading indicates deeper water. Contour lines are plotted at 20-, 50-, 100-, 500-, and 1000-m isobaths. The white rectangle in the inset map outlines the area shown in the main map.

observed (e.g., dolphins, sea turtles), we restricted our analysis to fish taxa, which all fell within classes Actinopterygii and Chondrichthyes. Because fish passed in and out of the video field and individuals were not easily identified if they appeared multiple times in the video, fish were not strictly counted. Instead, we used the Mean-Count method (Schobernd et al. 2014) as a proxy for counts of each species at each site, calculated as the mean number of fish observed per second, over the entire 20-min video. Multiple counting methods have been developed for enumerating species in videos, but we used the MeanCount method because it has been shown to be unbiased relative to true abundance (Schobernd et al. 2013). In total, 2115 sites met our filtering criteria, including videos recorded between 1995 and 2007, containing a total of 243 fish species (Appendix S1: Table S1).

At each site, we calculated four common biodiversity metrics (hereafter, metrics). The first was simply the number of species observed at a site, or species richness (S; Magurran 2004). This value is sometimes referred to as species density (Gotelli 2008). The second was Margalef's index [$M = (S - 1)/\ln(N)$], where N is the number of individuals observed at a site, and S is the number of species observed at a site, which attempts to compensate S for variation in sampling effort and population density (Magurran 2004). The third was one of the most commonly used metrics, Shannon's index ($H' = -\sum p_i \ln(p_i)$), where p_i is the number of individuals of species i observed. The fourth metric was Simpson's index ($D = \sum p_i^2$), where p_i is the number of individuals of species i observed, which quantifies the probability that a pair of organisms drawn randomly from an infinite community are from the same species (Magurran 2004). Though many other metrics exist, these were chosen due to their common usage (Kenchington and Kenchington 2013), computational simplicity, and dependence on a single data source.

To simulate the use of reduced species lists (RSLs) in sampling fish communities, we filtered the GOMVS dataset to include only species in a particular RSL and then re-estimated metrics at each site. The estimates were compared to

metrics that were calculated based on all fish species observed, which should be the closest available estimates of the actual fish biodiversity, and are hereafter referred to as the true metrics. (While these estimates may not represent true fish biodiversity in the purest sense, they are the best available estimates of the theoretical true values in this dataset and are considered to be true within the context of the analysis. Since they are used as reference estimates, it should not affect the results of the study if they differ from theoretical true values. Furthermore, while there is sampling selectivity associated with any sampling method, it is possible that fish occupy sites but are not visible to the camera, such as fish hidden by substrate or structure; for fish species moving about in the water column within the camera's field of view, these estimates should be reasonably close to true estimates.)

We identified three RSLs relevant to the GOMVS dataset that were currently utilized (hereafter, empirical RSLs). The first was a list of 122 species used by the GOMVS to reduce video processing cost in recent years. The second was a list of 107 species used by the largest video survey of reef fish in the SEUS Atlantic (SouthEast Reef Fish Survey [SERFS]; www.sefsc.noaa.gov), an adjacent region sharing most of the same species. The third contained the 85 species in the US Fish Stock Sustainability Index (FSSI; www.nmfs. noaa.gov) found in the Gulf of Mexico. The filtered GOMVS dataset included 103, 53, and 37 species from the GOMVS, SERFS, and FSSI lists, respectively (Appendix S1: Table S1).

To quantify the retention of fish diversity information using specific RSLs, we calculated nonparametric same-site correlations (Spearman's ρ) between true (all species) and RSL-estimated metrics. High same-site correlation would suggest that the RSL should perform well in biodiversity studies, closely estimating the true values of biodiversity metrics. To evaluate the performance of each empirical RSL, we conducted bootstrap analysis comparing the observed same-site correlations with a null distribution of same-site correlations from 1000 randomly generated RSLs containing the same number of species as each empirical RSL. For example, if an empirical RSL contained 50 species, then for each of 1000 bootstrap replicates, 50 species would be randomly drawn, without replacement, from the total list of

species observed in the dataset, to generate a random RSL; the same-site correlation would then be calculated for this RSL. Comparing the same-site correlation of an empirical RSL with the corresponding bootstrap distribution allows one to evaluate how they perform relative to randomly devised lists. Lists with observed same-site correlations above or below the 95% bootstrap confidence intervals were considered high or low performing, respectively. To measure the influence of each species i on RSL performance, we calculated the change in correlation due to excluding only that species as $1 - \rho(i)$, where $\rho(i)$ was the correlation for the list with only species i removed.

To understand how to create an RSL that would capture true patterns in biodiversity, we sought to identify how the selection of species affects performance (i.e., same-site correlation). Although a number of factors influenced the selection of species for the empirical lists, one major factor was species commonness, or prevalence, since many of the species were chosen for their importance to fisheries. Thus, we created a species commonness gradient, by ranking all 243 species from most to least common, based on their frequency of occurrence (Appendix S1: Table S1), and we identified four methods of selecting n species (where $1 \leq n \leq 243$). These methods involved choosing n species randomly, at even intervals along the commonness gradient, choosing the n most common species, or the n least common species. For each method, and for each value of n, we created an RSL, calculated values of each metric at each site, and then calculated same-site correlations with true values. For the random method, for each value of n we repeated 1000 times the process of selecting species and calculating same-site correlations and then used the 1000 values to compute means and 95% confidence intervals. This procedure produced a universe of same-site correlations for each metric, for all possible RSLs for the three deterministic selections methods, and $\approx 243{,}000$ possible lists for the random selection method. It thus provided broad context for the performance of our empirical lists and elucidated how the performance of hypothetical lists varied among selection methods.

Because the overall goal was to understand how the RSLs affect the value of a dataset for biodiversity research, we evaluated common

biodiversity hypotheses and examined the sensitivity of results to the species list. For the three empirical RSLs described above, we also tested for dependence of each metric on latitude (decimal degrees) or depth (meters) using linear regression.

We further considered the relative robustness of diversity metrics to the use of RSLs in these tests with depth and latitude, with two simple approaches. Aggregating results for lists of all lengths, we calculated absolute relative errors of the slopes of regression of each diversity metric on each predictor [|(observed slope − true slope)/ true slope|]. Note that error rate can range from 0 to ∞ with lower values indicating better estimates and a value of 0 indicating perfect estimation. We also calculated the error rate in these hypothesis tests as the proportion of the tests using RSLs that did not agree with the true test result.

To investigate how apparent temporal trends in biodiversity varied dependent on the RSL used, we calculated total number of species, rarefied species richness, and asymptotic species richness (i.e. Chao2 index; Chao 1989) for each year, for all three empirical species lists, as well as for the total dataset. Calculations of rarefied species richness, Chao2, and corresponding 95% confidence intervals were based on several sources (Longino and Colwell 1997, Colwell et al. 2004, Colwell 2013). Note that these time series were developed for the purposes of our methodological analysis only, were not standardized, and were not intended to represent actual trends in biodiversity in the Gulf of Mexico.

RESULTS

Metrics from RSLs with more species were more highly correlated with true metrics (Fig. 2). Bootstrap analysis showed that empirical RSLs performed similar to random lists including the same number of species, when estimating species richness. The FSSI and SERFS lists (n = 85 and 107 species, respectively) outperformed many randomly generated lists when estimating Margalef's index, while the GOMVS list still performed similar to the average random list. For Shannon's and Simpson's indices, all empirical lists outperformed most random lists of their same lengths (Fig. 3). Correlations generally increased with the number of species included in
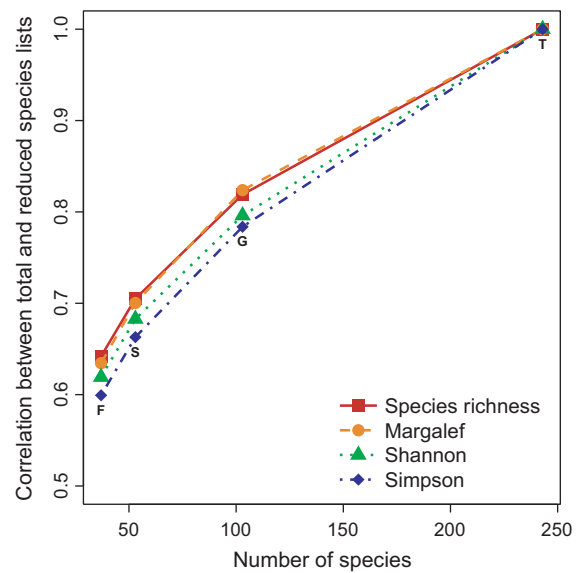

Fig. 2. Correlations in diversity metrics between estimates based on the total dataset and data filtered with reduced species lists, plotted by the number of species in each list found in the dataset. A single letter plotted below each set of points indicates each species list (F = FSSI, S = SERFS, G = GOMVS, and T = total). Symbols and line types for each diversity metric are provided in the legend.

the reduced list. For all metrics, confidence intervals around the mean correlation of the random lists overlapped for all empirical RSLs.

When we generated a broader universe of hypothetical lists, it became evident that RSL performance tended to increase across the entire range of n species, for all metrics and selection patterns (Fig. 4). The functional relationships between correlation and number of species differed widely among selection methods, but were quite similar among metrics. When the most common species were preferentially included in RSLs, correlations increased very rapidly with number of species initially, but gains in performance then diminished. For all diversity metrics, a ρ of 0.9 was reached once a list contained the 49 most common species. By contrast, lists composed of the least common species generally performed very poorly until they contained a large number of species. For this selection pattern, a ρ of 0.9 was not reached by all metrics until the list contained the 242 least common species. The performance of lists chosen by even selection was
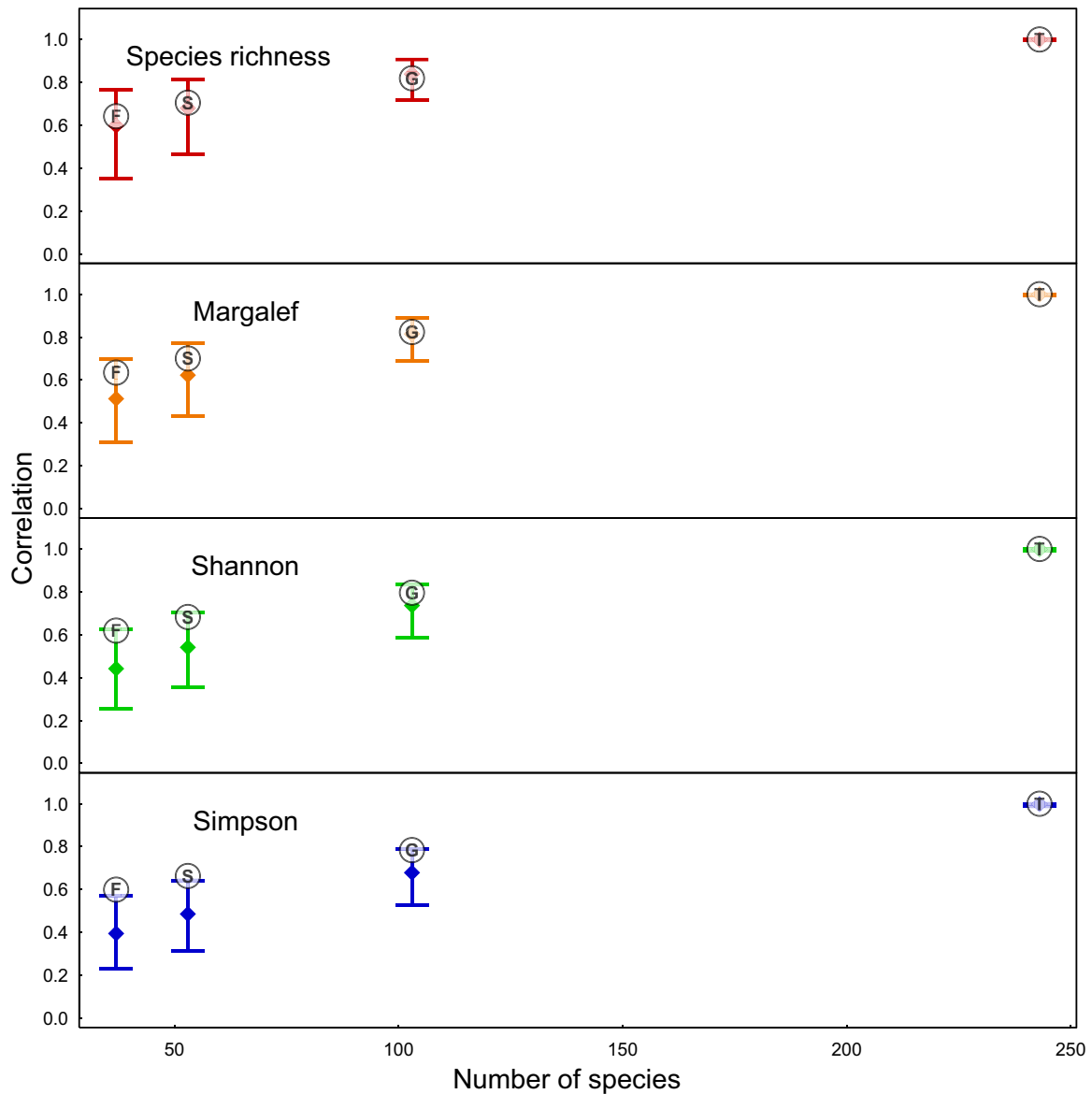
Fig. 3. Mean and 95% confidence intervals (error bars) for the random list correlations. The observed correlations for empirical lists are plotted as the encircled first initial of the list name (Ⓕ = FSSI, Ⓢ = SERFS, Ⓖ = GOMVS, and Ⓣ = total).

similar to the mean performance of lists chosen randomly, across the entire range of n species, always falling within the 95% confidence intervals around mean ρ. Both selection methods performed much better than preferential selection of uncommon species; however, correlations did not reach 0.9 for all metrics until lists included 175 species for even selection and 194 species for random selection. When species were individually removed from the total dataset, the greatest decreases in correlation tended to be associated with the removal of the most common species. The five most common species were red porgy (*Pagrus pagrus*), scamp (*Mycteroperca phenax*), almaco jack (*Seriola rivoliana*), red snapper (*Lutjanus campechanus*), and greater amberjack (*Seriola dumerili*), respectively (Appendix S1: Table S1; Fig. 5).
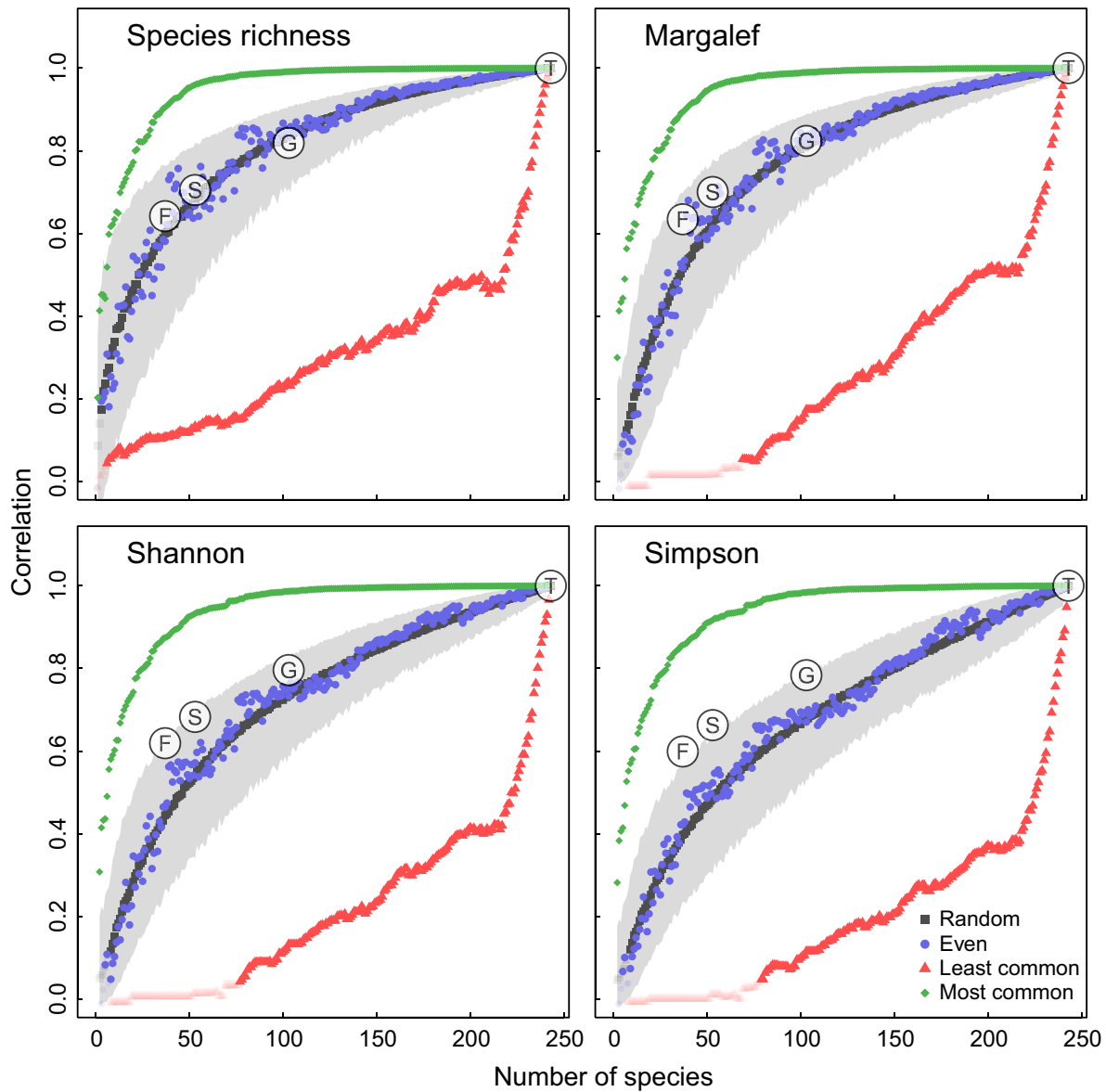
Fig. 4. Correlations between diversity metrics estimated with reduced species lists and true values. Lists were generated by selecting *n* species from the set of all species ordered along a gradient of commonness in the dataset. Species were selected randomly (mean and 95% CI of 1000 random draws; gray shading), evenly spaced along the commonness gradient, or from the least, or most common species. Correlations for the empirical lists are plotted as the encircled first initial of the list name (Ⓕ = FSSI, Ⓢ = SERFS, Ⓖ = GOMVS, and Ⓣ = total). Faded symbols indicate non-significant correlations.

In the total dataset, all metrics decreased significantly with depth; thus, diversity was higher at shallower sites. When conducting this same hypothesis test with datasets employing RSLs, results varied widely (Fig. 6). For all selection methods, longer lists were more likely to yield the true result. Even selection correctly identified the significant negative slopes most of the time, but misidentified positive slopes when very few species were included. Random selection followed roughly the same trend as the even selection results. With the least common species, many lists
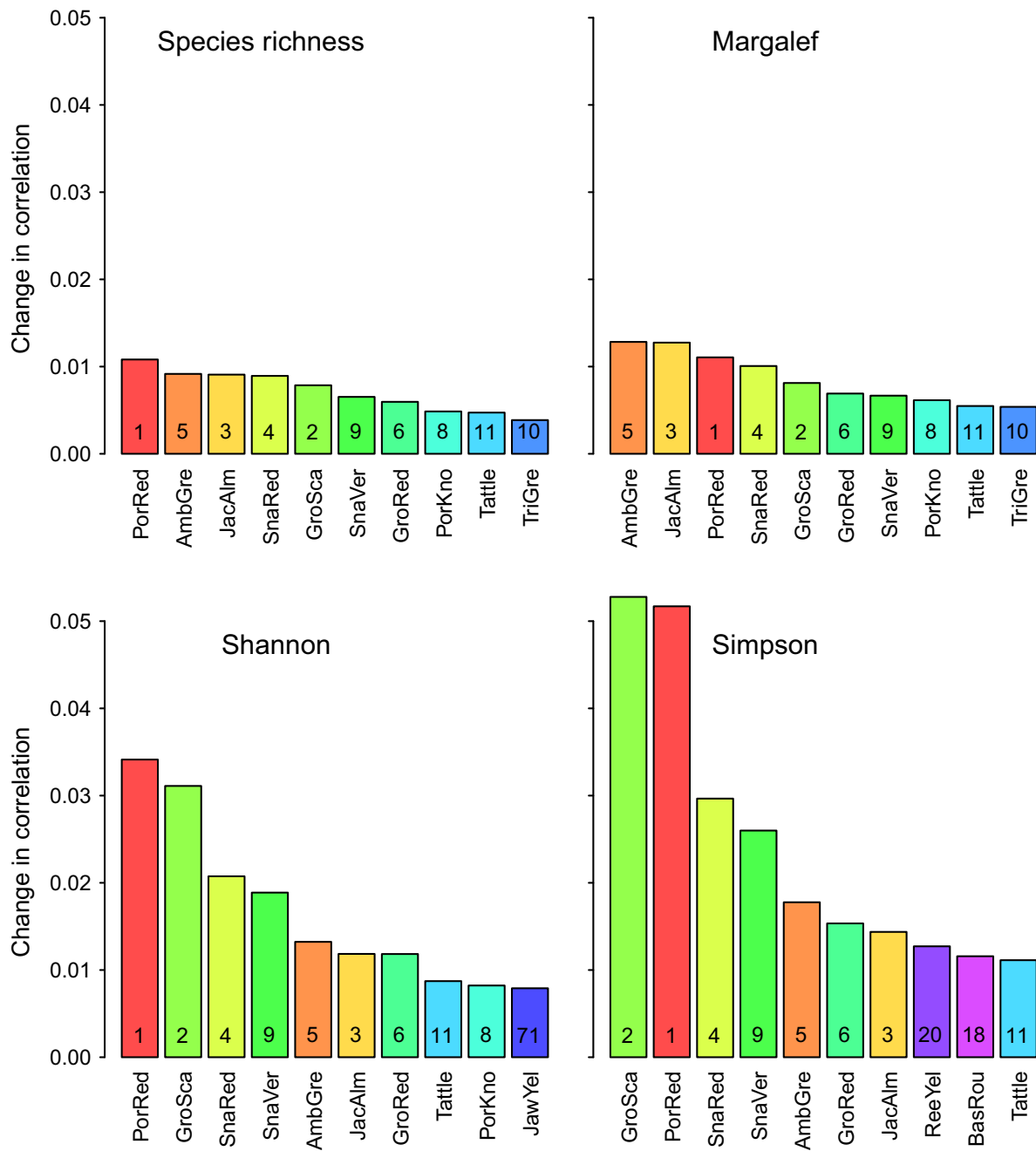
Fig. 5. Barplots of the 10 most influential species on each diversity metric. The change in correlation for species i is computed as $1 - \rho(i)$, where $\rho(i)$ is the correlation for the list with only species i removed. The numbers on each bar indicate the rank of each species in the commonness gradient with 1 being the most common. Bar colors are associated with each of the 13 species represented in this figure. PorRed = red porgy (*Pagrus pagrus*), GroSca = scamp (*Mycteroperca phenax*), JacAlm = almaco jack (*Seriola rivoliana*), SnaRed = red snapper (*Lutjanus campechanus*), AmbGre = greater amberjack (*Seriola dumerili*), GroRed = red grouper (*Epinephelus morio*), PorKno = knobbed porgy (*Calamus nodosus*), SnaVer = vermilion snapper (*Rhomboplites aurorubens*), TriGre = gray triggerfish (*Balistes capriscus*), Tattle = tattler (*Serranus phoebe*), BasRou = roughtongue bass (*Holanthias martinicensis*), ReeYel = yellowtail reef fish (*Chromis enchrysurus*), and JawYel = yellowhead jawfish (*Opistognathus aurifrons*).
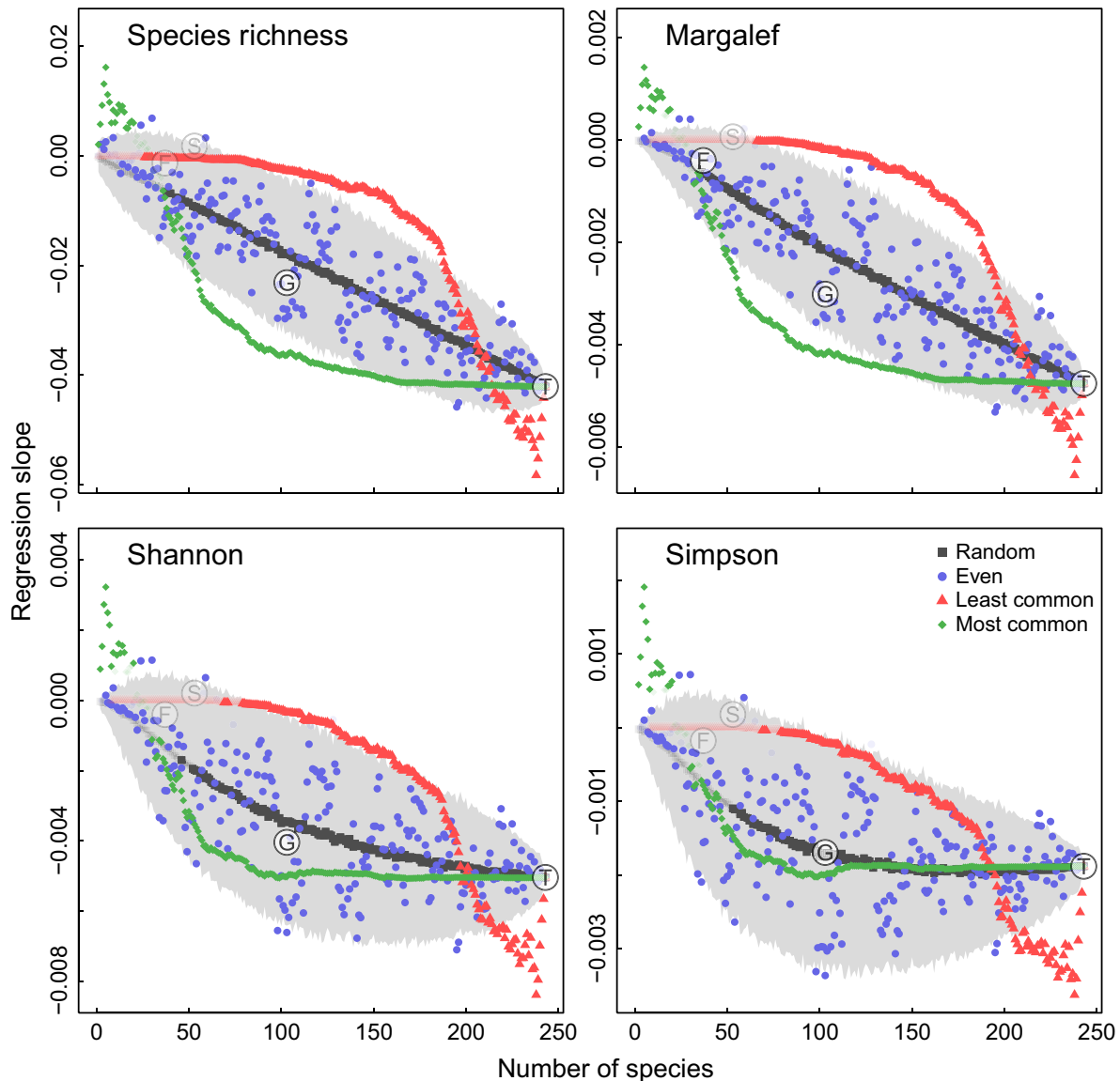
Fig. 6. Regression slopes between depth and diversity metrics estimated with reduced species lists and true values. Lists were generated by selecting *n* species from the set of all species ordered along a gradient of commonness in the dataset. Species were selected randomly (mean and 95% CI of 1000 random draws; gray shading), evenly spaced along the commonness gradient, or from the least, or most common species. Regression slopes for the empirical lists are plotted as the encircled first initial of the list name (Ⓕ = FSSI, Ⓢ = SERFS, Ⓖ = GOMVS, and Ⓣ = total). Faded symbols indicate non-significant regression slopes.

of fewer than 60 or so species failed to identify the negative slope for all metrics; lists that included all but the most common species achieved the correct qualitative result, but tended to estimate more steeply negative slopes. Unexpectedly, the shortest lists including the most common species incorrectly identified a significant positive relationship

with depth for all metrics. Regarding the empirical lists, the GOMVS list correctly identified the negative slope for all metrics, while the SERFS list never did, and the FSSI list only yielded the correct result for Margalef's index.

In the total dataset, species richness and Margalef's index decreased significantly with latitude;

however, no significant relationship was observed for Shannon's or Simpson's index (Fig. 7). These correct results were obtained when using RSLs based on even selection of species, but many lists failed to identify a trend or incorrectly yielded the opposite trend. In these comparisons, incorrect results were common for long and short lists. For randomly selected lists, only the longest lists consistently identified the correct results for species richness and Margalef's index, but lists of all lengths tended to yield a non-significant slope for Shannon's and Simpson's indices.
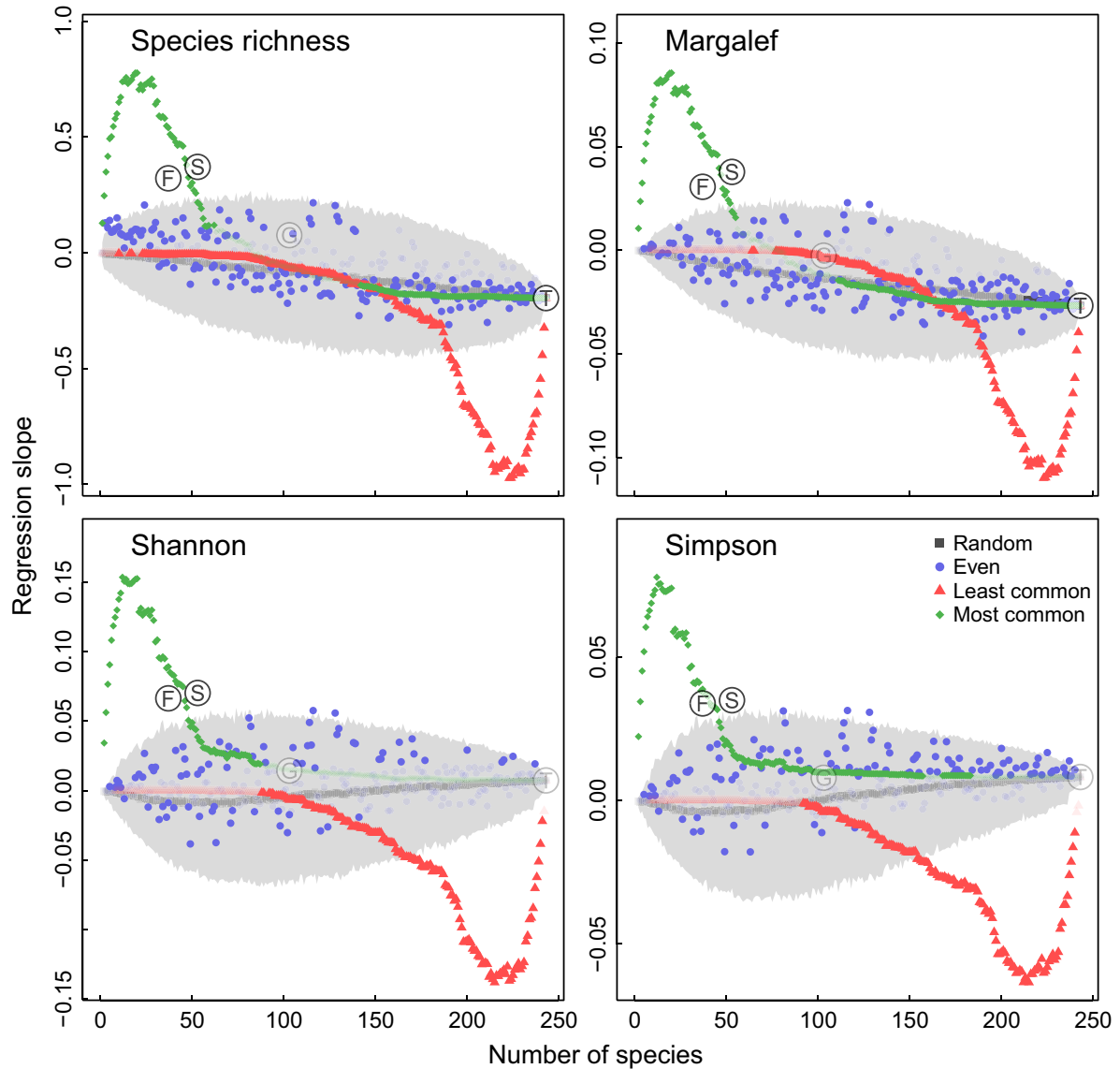

Fig. 7. Regression slopes between latitude and diversity metrics estimated with reduced species lists and true values. Lists were generated by selecting n species from the set of all species ordered along a gradient of commonness in the dataset. Species were selected randomly (mean and 95% CI of 1000 random draws; gray shading), evenly spaced along the commonness gradient, or from the least, or most common species. Regression slopes for the empirical lists are plotted as the encircled first initial of the list name (Ⓕ = FSSI, Ⓢ = SERFS, Ⓖ = GOMVS, and Ⓣ = total). Faded symbols indicate non-significant regression slopes.

Table 1. Median absolute relative errors [|(observed − true)/true|] for slopes of regression of diversity metrics on depth or latitude.

Predictor	Method	Species richness	Margalef	Shannon	Simpson
Depth	Random	0.5 (0.5)	0.47 (0.51)	0.24 (0.46)	0.04 (0.3)
Depth	Even	0.52 (0.56)	0.47 (0.57)	0.3 (0.5)	0.29 (0.45)
Depth	Low	0.9 (0.34)	0.93 (0.34)	0.87 (0.49)	0.8 (0.44)
Depth	High	0.1 (0.33)	0.09 (0.3)	0.02 (0.15)	0.02 (0.08)
Latitude	Random	0.5 (0.5)	0.42 (0.43)	1.23 (1.26)	0.77 (0.96)
Latitude	Even	0.67 (0.91)	0.5 (0.76)	1.15 (1.78)	0.63 (0.84)
Latitude	Low	0.92 (0.44)	0.98 (0.38)	2.86 (7.36)	2.05 (3.54)
Latitude	High	0.45 (1.5)	0.36 (1.19)	0.77 (2.64)	0.17 (0.64)

Notes: Interquartile ranges (3rd quartile–1st quartile) are in parentheses. Rows represent unique combinations of predictor and method.

Selecting the few most common species incorrectly yielded significant positive relationships with latitude, some of which were quite strong (Fig. 7). For species richness and Margalef's index, lists composed of an intermediate number of the most common species failed to identify any trend, but longer lists of approximately 100 or more species correctly identified the negative slope. For Shannon's index, the correct result of no trend was identified by lists longer than about 90 species, while for Simpson's index only the longest lists yielded the correct result. For all metrics, shorter lists of the least common species tended not to identify a trend, but most longer lists yielded significant negative slopes. These results were correct for species richness and Margalef's index, but incorrect for Shannon's and Simpson's indices, except for lists containing nearly all species. As in linear regressions with depth, lists containing all but the most common species produced the most steeply negative relationships with latitude, dramatically overestimating the strength of the true relationship.

For all metrics, both the FSSI and SERFS lists incorrectly identified significant positive slopes (Fig. 7). The GOMVS lists failed to identify the true negative relationships for species richness and Margalef's index, but yielded correct results for Shannon's and Simpson's indices.

None of the diversity metrics appeared to be more robust to the use of RSLs than any other. Absolute relative errors of regression slopes were quite variable for each metric, ranging from 0.02 to 2.86 (Table 1). However, absolute relative errors did show a clear ranking of selection methods, with selection of the most common species yielding the most accurate slope estimates (i.e., lowest absolute relative error), random and even selection methods demonstrating intermediate accuracy, and selection of the least common species exhibiting the lowest accuracy. This simple ranking agreed with our interpretation of Figs. 6 and 7, but did not reflect the complex relationships evident there. Error rates for tests of regression slopes were fairly low overall, with a median of 0.23 and 75th percentile of 0.33 (Table 2).

Table 2. Error rates of hypothesis tests for regression of diversity metrics on depth or latitude.

Predictor	Method	Species richness	Margalef	Shannon	Simpson
Depth	Random	0.18	0.15	0.21	0.23
Depth	Even	0.06	0.05	0.05	0.06
Depth	Low	0.11	0.25	0.29	0.29
Depth	High	0.06	0.06	0.06	0.06
Latitude	Random	0.95	0.88	0	0
Latitude	Even	0.28	0.26	0.33	0.47
Latitude	Low	0.08	0.28	0.64	0.63
Latitude	High	0.33	0.23	0.36	0.71

Notes: Numerical values represent the proportion of hypothesis test results that do not match the true result, when testing the significance of slopes of regressions of diversity metrics on each predictor. Rows represent unique combinations of predictor and method.

Time series of total species richness and rarefied species richness showed that trends based on RSLs reflected many of the main features of the true time series, but trends in the true time series of asymptotic species richness were much less evident in the RSL time series. Variation among years was lower in lists with fewer species (Fig. 8B, D, and F), which had an overall effect of reducing apparent changes in diversity over time (Fig. 8A, C, and E). In some cases,
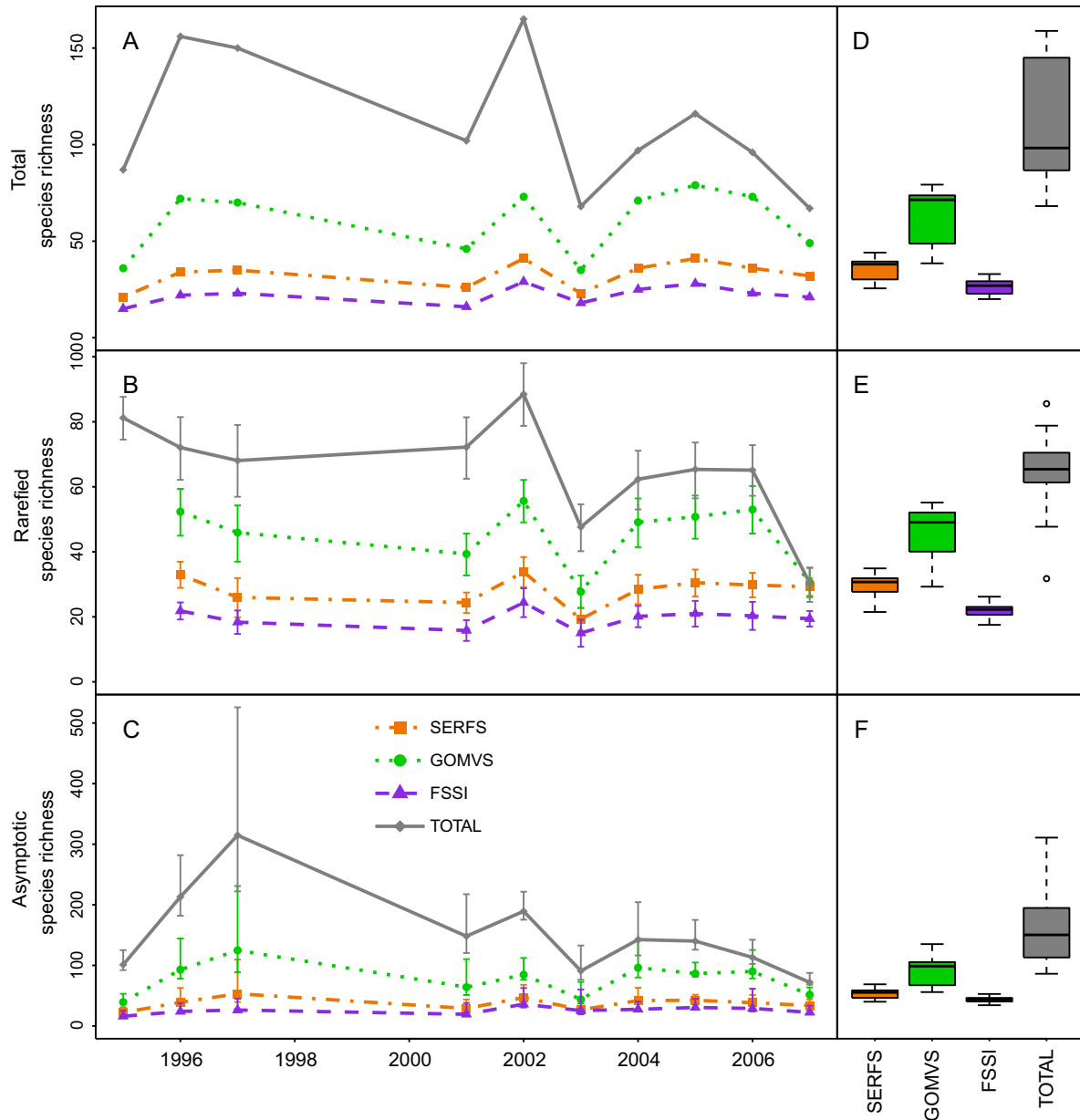

Fig. 8. Time series of (A) total, (B) rarefied, and (B) asymptotic species richness, by year, and corresponding boxplots of values for all years combined (D, E, and F). Error bars extend to 95% confidence limits in panels (B) and (C). Whiskers in boxplots extend to the most extreme data point no more than 1.5 times the interquartile range out from the box in (D), (E), and (F).

differences in diversity between consecutive years (i.e., non-overlapping confidence intervals) present in the total dataset were not evident in shorter RSLs, such as in rarefied species richness from 2006 to 2007 (Fig. 8C) and in asymptotic species richness from 1995 to 1996, 2002 to 2003, and 2006 to 2007 (Fig. 8E). Somewhat surprisingly, variation within years also tended to decrease for RSLs with fewer species. Thus in some cases, differences in diversity between consecutive years were evident in one or more of the RSL-based trends, but not in the trend based on the total dataset, such as in rarefied species richness from 2001 to 2002 and 2003 to 2004 (Fig. 8C) and in asymptotic species richness from 2003 to 2004 (Fig. 8E). These were false differences representing statistical type I errors (Quinn and Keough 2003).

DISCUSSION

We found that including more species in a reduced species list generally resulted in more accurate predictions of biodiversity. The pattern of species selection also had a large effect on RSL performance. Based on correlations alone, the best way to select species for an RSL seemed to be including at least the most common species. We also observed that the relationships between RSL length and performance were nonlinear, differing substantially between selection methods. For example, when preferentially selecting common species, increases in correlation were rapid for the most common species and then diminished quickly once these were included.

In the Gulf of Mexico reef fish data, including only the 49 most common species yielded correlations of at least 0.9 for all metrics. Thus, it might seem that the remaining 194 species could be excluded when describing trends in biodiversity. However, the results of our hypothesis tests indicated the need to include even more species. Considering relationships between biodiversity and depth, an RSL with the 49 most common species was able to detect the significant decrease in diversity with depth, but the estimated relationship (i.e., slope) was not as strong as the true relationship. The situation was substantially worse when investigating latitudinal trends in biodiversity. For all metrics, an RSL with the 49 most common species resulted in the incorrect

result of a significant positive slope. In fact, it was not until 184 of the most common species were counted that all metrics consistently yielded the true result, with Simpson's index requiring the longest list of species to agree with the estimate from the total dataset. These results were very consistent among diversity metrics, as none appeared to be more robust to the use of RSLs than any other.

These results suggest that it is generally preferable to avoid using RSLs when collecting data intended for biodiversity research, because analyses based on the filtered data may not represent true patterns in the entire community. However, when resource limitations require the use of RSLs, as in reef fish surveys in the SEUS, it becomes necessary to assess the utility of the filtered dataset, and devise efficient ways to improve the data if necessary. It seems clear that neither the FSSI nor the SERFS list captures enough of the variability in the video data to be reliable for use in studying true biodiversity of reef fish communities in the Gulf of Mexico. However, the list currently used on this survey (GOMVS) fared considerably better. It did fail to detect decreases in species richness and Margalef's D with latitude, but agreed with the true results on all other tests, and followed temporal trends fairly well.

Performance in biodiversity studies would likely be improved by adding common species to the list. As tests of trends with depth and latitude showed, RSLs that excluded only a few of the most common species dramatically underestimated the true slopes of these relationships (Figs. 6 and 7). Unfortunately, the most common species tend to be the most time-consuming to count, so there is a clear conflict between accuracy and efficiency.

Implications for other sampling methods

Every sampling method tends to have a unique species selectivity profile (e.g. Sorensen et al. 2002, McClanahan and Mangi 2004), where the probability of observing an individual with that method differs for each species. Species selectivity of a sampling method is often directly dependent on the ecology or morphology of a species. For instance, aerial hand collecting of forest spiders favors understory species (Sorensen et al. 2002), and small mesh mist nets catch more

small- than large-bodied birds (Heimerdinger and Leberman 1966, Pardieck and Waide 1992). But such traits often correlate with the commonness of species and therefore so does species selectivity. A study on insects shows that rare species are less likely to be caught in small pitfall traps (Work et al. 2002); research on fish shows that the most common species (*Leptoscarus vaigiensis*) in pooled samples was completely missed by one of the six methods used (McClanahan and Mangi 2004). In another study, inexperienced volunteers were shown to be less likely than experienced observers to detect uncommon insect species (Fitzpatrick et al. 2009).

Each RSL we generated can be thought of as the selectivity profile of a hypothetical sampling method. The video survey data we used also have an underlying selectivity profile, but it is constant across all comparisons and therefore does not affect our results. Our results showed that it was important for RSLs to contain a lot of species, suggesting that other sampling methods in biodiversity studies should be as unselective as possible. To the degree that they are selective, methods that miss very common species will be the most problematic, regardless of the diversity metric. This is especially evident given that lists including all but the two or three most common species sometimes failed to detect the true relationships between diversity metrics and latitude (Fig. 7). Methods that fail to detect rare species should tend to be less problematic. However, failure to detect rare species would be a critical problem for asymptotic species richness estimators such as the Chao2 index (Chao 1989), which rely heavily on observations of rare species.

Our results for lists generated based on random selection of species should be especially informative to studies using other sampling methods, because they are only defined by the number of species included. Therefore, any method that captures a given number of reef fish species in the region we studied should be roughly represented in our plots of simulated lists. For example, by sampling with gear such as longlines or trawls capable of catching any 200 Gulf of Mexico reef fish species, it seems reasonable to expect that the strong negative relationship between depth and biodiversity would be detected. At the same time, because so many lists including 200 species did not identify the true

relationship between biodiversity metrics and latitude, one may not be very confident in that dataset for identifying weaker latitudinal trends.

Considering species selectivity when interpreting results

When a dataset has already been collected, or it is otherwise not possible to use a unselective sampling method, biodiversity estimates from different sampling methods should be treated independently. One may then continue to interpret trends in biodiversity for that sampling method alone, as an index of a particular part of a community, but not expect it to represent true biodiversity. This is similar to the caveat from Gotelli (2008) that comparing diversity estimates from rarefaction curves based on different sampling methods is invalid. Problems comparing diversity estimates from different methods are also reflected in observations by Abele and Walters (1979) in a reevaluation of earlier studies.

Reconsidering our comparisons of biodiversity and latitude from this alternative perspective, we observe that for sampling methods that capture only small numbers (<50) of the most common species, biodiversity increases with latitude. This is not the same trend that exists in the broader community, but might be the relevant trend in certain contexts. For instance, this might suggest that higher latitudes are also more valuable to human user groups attracted to both abundance and diversity of species, such as recreational divers. By contrast, reefs with high diversity may be of low value to fishers trying to avoid bycatch.

Conclusions

When attempting to estimate true diversity of an ecological community, filtering data through the use of RSLs may obscure true patterns and should therefore be avoided. When such lists must be used, they should include as many species as possible, especially common species. By extension, even when not explicitly filtering out species in a dataset, sampling methods should be as unselective of species as possible. When sampling methods are very species selective, results should not be extrapolated to the broader ecological community, but should only be interpreted to represent the list of species sampled.

Acknowledgments

Literature Cited

Abele, L. G., and K. Walters. 1979. The stability-time hypothesis: reevaluation of the data. American Naturalist 114:559–568.

Aguzzi, J., C. Costa, Y. Fujiwara, R. Iwase, E. Ramirez-Llorda, and P. Menesatti. 2009. A novel morphometry-based protocol of automated video-image analysis for species recognition and activity rhythms monitoring in deep-sea fauna. Sensors 9:8438–8455.

Bacheler, N. M., and K. W. Shertzer. 2015. Estimating relative abundance and species richness from video surveys of reef fishes. Fishery Bulletin 113:15–26.

Bibby, C. J., S. Marsden, and M. Jones. 1998. Bird surveys. Expedition Advisory Centre, London, UK.

Boulinier, T., J. D. Nichols, J. R. Sauer, J. E. Hines, and K. Pollock. 1998. Estimating species richness: the importance of heterogeneity in species detectability. Ecology 79:1018–1028.

Buckland, S. T., D. L. Borchers, A. Johnston, P. A. Henrys, and T. A. Marques. 2007. Line transect methods for plant surveys. Biometrics 63:989–998.

Campbell, M. D., A. G. Pollack, C. T. Gledhill, T. S. Switzer, and D. A. DeVries. 2015. Comparison of relative abundance indices calculated from two methods of generating video count data. Fisheries Research 170:125–133.

Chao, A. 1989. Estimating population-size for sparse data in capture recapture experiments. Biometrics 45:427–438.

Colwell, R. K. 2013. EstimateS User's Guide. http://viceroy.eeb.uconn.edu/estimates/EstimateS Pages/EstSUsersGuide

Colwell, R. K., C. X. Mao, and J. Chang. 2004. Interpolating, extrapolating, and comparing incidence-based species accumulation curves. Ecology 85:2717–2727.

Dobzhansky, T. H. 1950. Evolution in the tropics. American Scientist 38:209–221.

Fitzpatrick, M. C., E. L. Preisser, A. M. Ellison, and J. S. Elkinton. 2009. Observer bias and the detection of low-density populations. Ecological Applications 19:1673–1679.

Gotelli, N. J. 2008. A primer of ecology. Sinauer Associates Inc, Sunderland, Massachusetts, USA.

Heimerdinger, M. A., and R. C. Leberman. 1966. The comparative efficiency of 30 and 36 mm. Mesh in mist nets. Bird-Banding 37:280–285.

Kenchington, T., and E. Kenchington. 2013. Biodiversity metrics for use in the ecosystem approach to oceans management. Canadian Technical Report of Fisheries and Aquatic Sciences 3059:vi+188.

Kumar, N., P. N. Belhumeur, A. Biswas, D. W. Jacobs, W. J. Kress, I. C. Lopez, and J. V. Soares. 2012. Leafsnap: a computer vision system for automatic plant species identification. Computer Vision–ECCV 2012:502–516.

Longino, J. T., and R. K. Colwell. 1997. Biodiversity assessment using structured inventory: capturing the ant fauna of a tropical rain forest. Ecological Applications 7:1263–1277.

Magurran, A. E. 2004. Measuring biological diversity. Blackwell Science Ltd., Malden, Massachusetts, USA.

Mallet, D., and D. Pelletier. 2014. Underwater video techniques for observing coastal marine biodiversity: a review of sixty years of publications (1952–2012). Fisheries Research 154:44–62.

McClanahan, T., and S. Mangi. 2004. Gear-based management of a tropical artisanal fishery based on species selectivity and capture size. Fisheries Management and Ecology 11:51–60.

Pardieck, K., and R. B. Waide. 1992. Mesh size as a factor in avian community studies using mist nets (Orificios de Tamano Diferente, Como Factor en Estudios de Comunidades de Aves en Donde se Utilizan Redes). Journal of Field Ornithology 63:250–255.

Parker, D., H. Winker, A. Bernard, and A. Götz. 2016. Evaluating long-term monitoring of temperate reef fishes: a simulation testing framework to compare methods. Ecological Modelling 333:1–10.

Quinn, G. P., and M. J. Keough. 2003. Experimental design and data analysis for biologists. Cambridge University Press, Cambridge, UK.

Schobernd, Z. H., N. M. Bacheler, and P. B. Conn. 2014. Examining the utility of alternative video monitoring metrics for indexing reef fish abundance. Canadian Journal of Fisheries and Aquatic Sciences 71:464–471.

Sorensen, L. L., J. A. Coddington, and N. Scharff. 2002. Inventorying and estimating sub-canopy spider

diversity using semiquantitative sampling methods in an Afromontane forest. Environmental Entomology 31:319–330.

Stevens, G. C. 1989. The latitudinal gradient in geographical range: how so many species coexist in the tropics. American Naturalist 133:240–256.

Topping, C., and K. Sunderland. 1992. Limitations to the use of pitfall traps in ecological studies exemplified by a study of spiders in a field of winter wheat. Journal of Applied Ecology 29:485–491.

White, D., C. Svellingen, and N. Strachan. 2006. Automated measurement of species and length of fish by computer vision. Fisheries Research 80:203–210.

Willig, M., D. Kaufman, and R. Stevens. 2003. Latitudinal gradients of biodiversity: pattern, process, scale, and synthesis. Annual Review of Ecology, Evolution, and Systematics 34:273–309.

Work, T. T., C. M. Buddie, L. M. Korinus, and J. R. Spence. 2002. Pitfall trap size and capture of three taxa of litter-dwelling arthropods: implications for biodiversity studies. Environmental Entomology 31:438–448.

Yu, X., J. Wang, R. Kays, P. A. Jansen, T. Wang, and T. Huang. 2013. Automated identification of animal species in camera trap images. EURASIP Journal on Image and Video Processing 2013:52.

SUPPORTING INFORMATION

Additional Supporting Information may be found online at: http://onlinelibrary.wiley.com/doi/10.1002/ecs2.1842/full