When to use species richness estimators to infer about diversity losses or gains
Version: 2024/03/29

Author: Gabriel Arellano
ORCID: https://orcid.org/0000-0003-3990-5344
E-mail: gabriel.arellano.torres@gmail.com

Affiliations:

- Ecology and Evolutionary Biology, University of Michigan, Ann Arbor, MI 48109, U.S.A.
- Oikobit LLC, www.oikobit.com, 2105 Vista Oeste St NW, Albuquerque, NM 87120, U.S.A.

Observed richness, S_{obs}, underestimates true richness, S_{true}. Richness estimators have the form $S_{est} = S_{obs} + a$, where a is a non-negative number. Using S_{est} is a general recommendation, as most richness estimators are closer to the true richness than S_{obs} (Reese et al. 2014). However, it is not immediately obvious if using richness estimators for the estimation of changes in richness ($\frac{S_{true}\,final - S_{true}\,initial}{S_{true}\,initial}$) is also a general recommendation. We can think, naively, that we can use observed richness to do this calculation without any bias, because the bias is cancelled when dividing by "initial". This is, in general, not true. It is easy to simulate realistic processes of diversity loss where the estimate of change in richness based on observed richness will be biased. On the other hand, richness estimators can be sensitive to certain aspects of the sample that can change abruptly. For example, Chao1 is sensitive to the number of singletons: small changes in the number of singletons can result in large changes in the estimated richness. Furthermore, Chao1 assumes that singletons result from the process of increasing sample size and incorporating new species into the sample, which may or may not be the case. For example, in a permanent tree plot being measured repeatedly, some new singletons will be species that reduced their abundances from the previous census. It is unclear if such differences in process or assumptions matter. Here, I study when to use richness estimators for the estimation of richness gains/losses, and when we could be better off by simply using the observed richness.

I use the notation I_{obs} for initial observed richness, and F_{obs} for final observed richness. All reasonable richness estimators can be expressed as $I_{est} = I_{obs} + a_i$ for the initial richness estimated by any richness estimator ($a_i \geq 0$) and $F_{est} = F_{obs} + a_f$ for the final richness ($a_f \geq 0$). The true initial richness is $I_{true} = I_{est} + \varepsilon_i$, and the final richness is $F_{true} = F_{est} + \varepsilon_f$, where ε_i and ε_f are errors. We also define $\Delta_{true} = \frac{F_{true} - I_{true}}{I_{true}}$, $\Delta_{obs} = \frac{F_{obs} - I_{obs}}{I_{obs}}$, and $\Delta_{est} = \frac{F_{est} - I_{est}}{I_{est}}$.

The bias in the estimation of change of richness when using observed richness is $b_{obs} = \Delta_{obs} - \Delta_{true}$. The bias in the estimation of change of richness when using estimated richness is $b_{est} = \Delta_{est} - \Delta_{true}$. We want to know when it is better to use observed richness over estimated richness, *i.e.*, when $b_{obs}^2 < b_{est}^2$. The problem is unsolvable in general, as true diversities are unknown and ε_i and ε_f could, in theory, take any value (positive or negative). However, the inequality can be evaluated for a range of realistic scenarios, assuming:

- The user will follow general recommendations (Reese et al. 2014). For the tool that I present here, I will assume that the user will pick one among the three best performing species richness estimators for their sample and community properties, as in Table 4 of Reese et al. (2014), or any similarly-biased estimator.
- The communities are "reasonable" in the sense that they fall within the scenarios evaluated by Reese et al. (2014). My approach may not work if the communities are experimental or extreme in certain aspects.

In general, we ignore ε_i and ε_f, but Reese et al. (2014) provide an estimate of the expected ranges for these errors, in realistic conditions. I suggest evaluating $b_{obs}^2 < b_{est}^2$ homogeneously within the space defined by realistic boundaries of ε_i and ε_f. This approach is conservative, in the sense that we would cover from the worst-case to the best-case, and we would do it homogeneously (*i.e.*, the exploration in the tails of the distribution of errors has much weight, even if those extreme errors are unlikely). However, it is true that our reasonable boundaries are taken from the three best performing estimators in just one study that covered a certain number of scenarios. If the richness estimators are suspected to be more biased than in Reese et al. (2014), one can expand our suggested boundaries and evaluate $b_{obs}^2 < b_{est}^2$ in a much larger error space. This should not be done just for the sake of it and it is not clearly a "conservative" decision in all cases: assuming that the richness estimator has huge error implies assuming real communities with very low or very high numbers of species.

The R function below implements the suggested approach. The main input to the function are I_{obs}, I_{est}, F_{obs} and F_{est}, for any number of samples. It is the responsibility of the user to obtain I_{est} and F_{est} following reasonable methods. The function explores a uniform grid of possible values of ε_i and ε_f within certain boundaries and arbitrary resolution. These boundaries, by default, are the minimum and maximum values in the "bias" column of Table 4 of Reese et al. (2014), plus/minus 2 standard deviations taken from the "precision" column of Table 4 of Reese et al. (2014). Using these defaults, the function explores biases from -62% to $+21\%$ for the richness estimator. The user can expand those boundaries by any factor, but large expansions will get to extreme diversities, so this is in general not recommended. The function evaluates b_{obs} and b_{est} at each combination of values of ε_i and ε_f. It returns one recommendation per sample about using observed richness or estimated richness, plus intermediate relevant results.

Here is the code:

```
obs_or_est_when_looking_at_changes <- function(
    Iobs = NULL,
    Fobs = NULL,
    Iest = NULL,
    Fest = NULL,
    lower.bias = -0.56,
    upper.bias = +0.05,
    sd.lower.bias = 0.03,
    sd.upper.bias = 0.08,
    factor.to.expand.boundaries = 1,
    n.grid = 1e+4) {

  # Differences between the observed and estimated richness:
  ns <- c(length(Iobs), length(Iest), length(Fobs), length(Fest))
  n = unique(ns)
  if(length(n) > 1) error("all inputs must be of the same length")
```

```
  ai <- Iest - Iobs
  if(any(ai <  0)) warning("initial observed > estimated (!)")

  af <- Fest - Fobs
  if(any(af <  0)) warning("final observed > estimated (!)")

  # Define the grid of possible biases of the estimated richness:
  L = lower.bias - 2*abs(sd.lower.bias)
  U = upper.bias + 2*abs(sd.upper.bias)
  extra = abs(L-U)*factor.to.expand.boundaries - abs(L-U)
  L = L - extra/2
  U = U + extra/2
  s <- seq(from = L, to = U, length.out = ceiling(sqrt(n.grid)))
  grid <- expand.grid(simulated.bias.Iest = s, simulated.bias.Fest = s)

  # Do the evaluation at each scenario of possible biases:
  out <- lapply(1:n, function(j) {
    Itrue <- Iest[j]/(1 + grid[,"simulated.bias.Iest"]) # assumed "true"
    Ftrue <- Fest[j]/(1 + grid[,"simulated.bias.Fest"]) # assumed "true"
    change.true <- (Ftrue - Itrue)/Itrue # assumed "true"
    change.obs = (Fobs[j] - Iobs[j])/Iobs[j]
    change.est = (Fest[j] - Iest[j])/Iest[j]
    bias.obs <- change.true - change.obs
    bias.est <- change.true - change.est
    obs.better = sum(bias.est^2 > bias.obs^2)/nrow(grid)
    est.better = sum(bias.est^2 < bias.obs^2)/nrow(grid)

    # Generate advice, initialized with the observed diversity
    use = "observed"
    confidence = obs.better/(obs.better + est.better)
    if(est.better > obs.better)
    {
      use = "estimated"
      confidence = est.better/(obs.better + est.better)
    }
    grid <- cbind(grid, bias.obs, bias.est)
    list(use = use, confidence = confidence, grid = grid)
  })

  out
}
```

Good luck!

Acknowledgements: This idea was inspired by conversations with Belén Fadrique, who is leading an exciting project that involves estimation of species gains or losses. I wrote this note while being funded by the US National Science Foundation award number 2020424: "AccelNet: International Tropical Forest Science Alliance (ITFSA): A multi-network science and training initiative to accelerate understanding of the role of tropical forests in the Earth System".

References:

Reese, G.C., Wilson, K.R., & Flather, C.H. 2014. Performance of species richness estimators across assemblage types and survey parameters. *Global Ecology and Biogeography* 23: 585–594.