APPENDIX C: Statistical estimation issues

In the main article, all diversity measures and similarity/differentiation indices are in terms of true community parameters (species richness and species relative abundances). In practice, all these parameters are actually unknown and must be estimated from samples taken from the community. In diversity decomposition, statistical estimation issues include (1) how we can obtain reliable estimators for alpha, beta and gamma diversities based on sample data, and (2) how we can obtain reliable estimators for the Sørensen \((q = 0)\), Jaccard \((q = 0)\), Horn \((q = 1)\) and Morisita-Horn similarity measures \((q = 2)\) or their corresponding differentiation measures.

The observed species richness and sample species frequencies are usually substituted respectively for the true species richness and the community species frequencies in the theoretical formulas for diversity and similarity measures. The resulting estimator is called the maximum likelihood estimator (MLE). However, the MLEs are typically biased unless sample sizes are large so that all species are found. When sample size is relatively small so that a large fraction of rare species are missed in samples, it is well known that the observed species richness severely underestimates the true species richness. Also, since the population fraction of the unobserved species is ignored, the relative abundance of observed species, on average, is over-estimated. These are the main causes for biases associated with the MLEs especially in highly diverse communities (Gotelli and Chao 2013). Therefore, the MLEs are not reliable estimators in nearly all applications, and other statistical methods are needed.

Sampling limitations create challenges for making accurate estimates of various measures. An important statistical objective in diversity estimation is to reduce the under-sampling bias and to
adjust for the effect of under-sampled species on the estimation of diversity and similarity measures. Because sampling variation is an inevitable component of biodiversity studies, it is equally important to assess the variance (or standard error) of an estimator and provide a confidence interval that will reflect sampling uncertainty.

The statistical properties of diversities or similarity/differentiation measures depend on the order \( q \). Generally, statistical methods can be used to produce nearly unbiased estimators for diversity or similarity measures of \( q = 2 \) because these measures are dominated by very abundant species which will be detected in any samples (Chao et al. 2005). For \( q = 1 \) measures based on Shannon entropy, statistical methods can help produce low-bias estimator (Chao and Shen 2003; Beck and Schwanghart 2010) and an associated confidence interval, but total species richness or similarity based on species richness (\( q = 0 \)) is much more difficult to estimate because it is very sensitive to rare species that are often undetected, even in relatively large samples; see Gotelli and Chao (2013) for details.

Alternatively, rarefaction and extrapolation curves have been used by ecologists to compare species richesseses at a standardized sampling efforts or sample completeness; see Colwell et al. (2012) and Chao and Jost (2012) for a review. A worthwhile research topic is to extend their rarefaction and extrapolation methods to compare Hill numbers, alpha, beta and gamma diversities, as well as similarity/differentiation measures at standardized sampling efforts or standardized sample completeness.

LITERATURE CITED


