**Supporting Information**

**Biodiversity survey and estimation for line-transect sampling**

**Mathematical foundation of the proposed statistical model**

*Some basic statistical properties*

Some key formulas, which are necessary for deriving the unbiased non-independent strength parameter, are provided below. For any integer , proofs for these key formulas can be easy, and thus omitted for brevity.

1)

2) , where

Note that Eqs. S1a and S1b are directly from transition probabilities of the Markov model in Eq. 2 of the main text.

3)

4) .

Eqs. S1c and S1d can be straightforwardly derived from Eqs. S1a and S1b. Because the sampling of the first individual along a line transect is random and proportional to species’ relative abundances, we have. For the second sampled individual, based on the above equalities (Eqs. S1a-S1d), we can derive the probability of the individual belonging to species as

5)

The result  for any integer  can be derived from mathematical induction.

*Derivation of sampling variance of the proposed model*

When sampling *m* individuals along a transect, the number of individuals being sampled for species *i* is denoted by , for which we can further express it in terms of species labels of the individuals as follows:

where if the *k*th sampled individual is from species *i*;, otherwise.

Its variance thus can be calculated by

As shown previously in Eq. S1e, , thus we have

(S4)

Further,

where the calculation of followed the standard Markov matrix property, after some tedious algebra, one can show that,

Therefore, substituting the results in Eqs. S4 and S5 into Eq. S3 leads to

in which, for a large *m*, we suggest an approximate and simple formula is

, (S7)

where we define for simplicity.

*Derivation of Eq. 8 of the main text*

To show the result in Eq. 8 of the main text, using

along with the result in Eq. S6 gives

as a consequence, we have

from which the desired result in Eq. 8 of the main text is evident and shown below

After removing the expectation operator, suppose we have found an estimator for the parameter or , the proposed estimator for becomes .

*Derivation of sampling covariance of the proposed model*

The covariance of abundance between two species *i* and *j* in the line-transect sampling under the Markov model is given by,

.

Using the property of Markov transition probabilities and after some algebra manipulation, one can show that

for, from which, for a large *m*, we have

The last equality of Eq. S13 is the result of Eq. 4 in the main text, for which it is also one of the key formulas in our study.

*Derivation of the proposed new Rao’s quadratic diversity index*

Based on the above derivation in Eq. S8, we can see that , from which we can derive a nearly unbiased estimator of the Rao’s quadratic diversity index, under the Markov model context, by

Therefore, after removing the expectation operator, if we can find an estimator for , we can derive the unbiased Rao’s quadratic diversity index under the non-independent sampling scenario as

The above equation is Eq. 10 in the main text. The associated Simpson index can be derived accordingly by assuming , which is actually the complement of Eq. 8 of the main text or the estimated as showed above in Eq. S10. Specifically, to show this, when in Eq. S15,

*Derivation of Eq. 9 of the main text*

Using some key formulas (Eqs. S1a-S1e) introduced in the first place, we can also easily prove Eq. 7 in the main text as follows. Taking expectation of Eq. 6 gives

Therefore, Solow (2000)’s estimator *v* tended to overestimate the true value, because of the confounding influence of . However, as remarked in the main text, Solow (2000)’s estimator is still useful for deriving the proposed unbiased estimator of the true and valid for some cases (*e.g*., is small).

To give the derivation of the proposed estimator of , we merge two estimating equations (Eq. 7 and 8 in the main text) into a single one as

that only gets the estimator involved. As a result, we can solve it as

where . This is Eq. 9 in the main text.

**Unification of Rao's quadratic diversity index and Nei's genetic diversity**

In molecular ecology and genetics, Nei's nucleotide diversity and haplotype diversity indices are two of the most widely applied indices (Nei 1973, 1987, Nei and Li 1979, Nei and Kumar 2000). Actually they are termed as Rao’s quadratic diversity index and unbiased Simpson index, respectively, as proved below (Simpson 1949, Rao 1982, Nayak 1986, Botta-Dukat 2005, Chen et al. 2018).

The power of these two indices relies on the fact that they are unbiased estimators (Pons and Petit 1996, Hardy and Senterre 2007, Meirmans and Hedrick 2011), which is particularly true under large sample sizes. However, despite of its robust statistical property and wide application, a hidden issue is that organism sampling for genetic, molecular and ecological studies is that individuals are assumed to be independent. Both nucleotide and haplotype diversity indices are also established on the basis of this independent sampling assumption. This implies that, when field sampling is conducted in non-random ways, the conventional calculation of nucleotide diversity and haplotype diversity indices might become misleading or at least inaccurate.

In the calculation of traditional haplotype and nucleotide diversity indices, suppose among the *H* different haplotypes, their observed abundances in the sampled line transect with *m* individuals is given as, again we have . Then, the observed relative abundance of a specific haplotype *i* can be calculated as , forming a vector as . Then, the haplotype diversity index, under the context of random sampling of organisms, can be calculated as (Nei and Tajima 1981),

Suppose the observed site difference between a pair of haplotypes (*i* and *j*) is given by , then the traditional nucleotide diversity index, under the context of random sampling of organisms, is calculated as (Nei and Li 1979, Nei 1987),

This formula is identical to Rao's quadratic diversity metric as shown in Eq. 5 of the main text by setting the distance metrics , . By plugging observed relative abundance values into the formulae. In such a case, .

References

Botta-Dukat, Z. 2005. Rao’s quadratic entropy as a measure of functional diversity based on multiple traits. Journal of Vegetation Science 16:533–540.

Chen, Y., Y. Wu, and T. Shen. 2018. Evaluation of the estimate bias magnitude of the Rao’s quadratic diversity index. PeerJ:e5211.

Hardy, O., and B. Senterre. 2007. Characterizing the phylogenetic structure of communities by an additive partitioning of phylogenetic diversity. Journal of Ecology 95:493–506.

Meirmans, P., and P. Hedrick. 2011. Assessing population structure: FST and related measures. Molecular Ecology Resources 11:5–18.

Nayak, T. 1986. An analysis of diversity using Rao’s quadratic entropy. Sankhya Series B 48:315–330.

Nei, M. 1973. Analysis of gene diversity in subdivided populations. PNAS 70:3321–3323.

Nei, M. 1987. Molecular Evolutionary Genetics. Columbia University Press, New York.

Nei, M., and S. Kumar. 2000. Molecualr evolutoin and phylogenetics. Oxford University Press, Oxford.

Nei, M., and W. Li. 1979. Mathematical model for studying genetic variation in terms of restriction endonucleases. PNAS 76:5269–5273.

Nei, M., and F. Tajima. 1981. DNA polymorphism detectable by restriction endonucleases. Genetics 97:145–163.

Pons, O., and R. Petit. 1996. Measuring and testing genetic differentiation with ordered versus unordered alleles. Genetics 144:1237–1245.

Rao, C. 1982. Diversity and dissimilarity coefficients: a unified approach. Theoretical Population Biology 21:24–43.

Simpson, E. 1949. Measurement of diversity. Nature 163:688.

**Additional Tables and Figures**

Table S1. A comparison on the fitted Markov non-independent parameter using Solow’s original estimator and the proposed one in the semi-numerical test in which the biomass information of 26 plant species in ultramafic soils of central Italy was used.

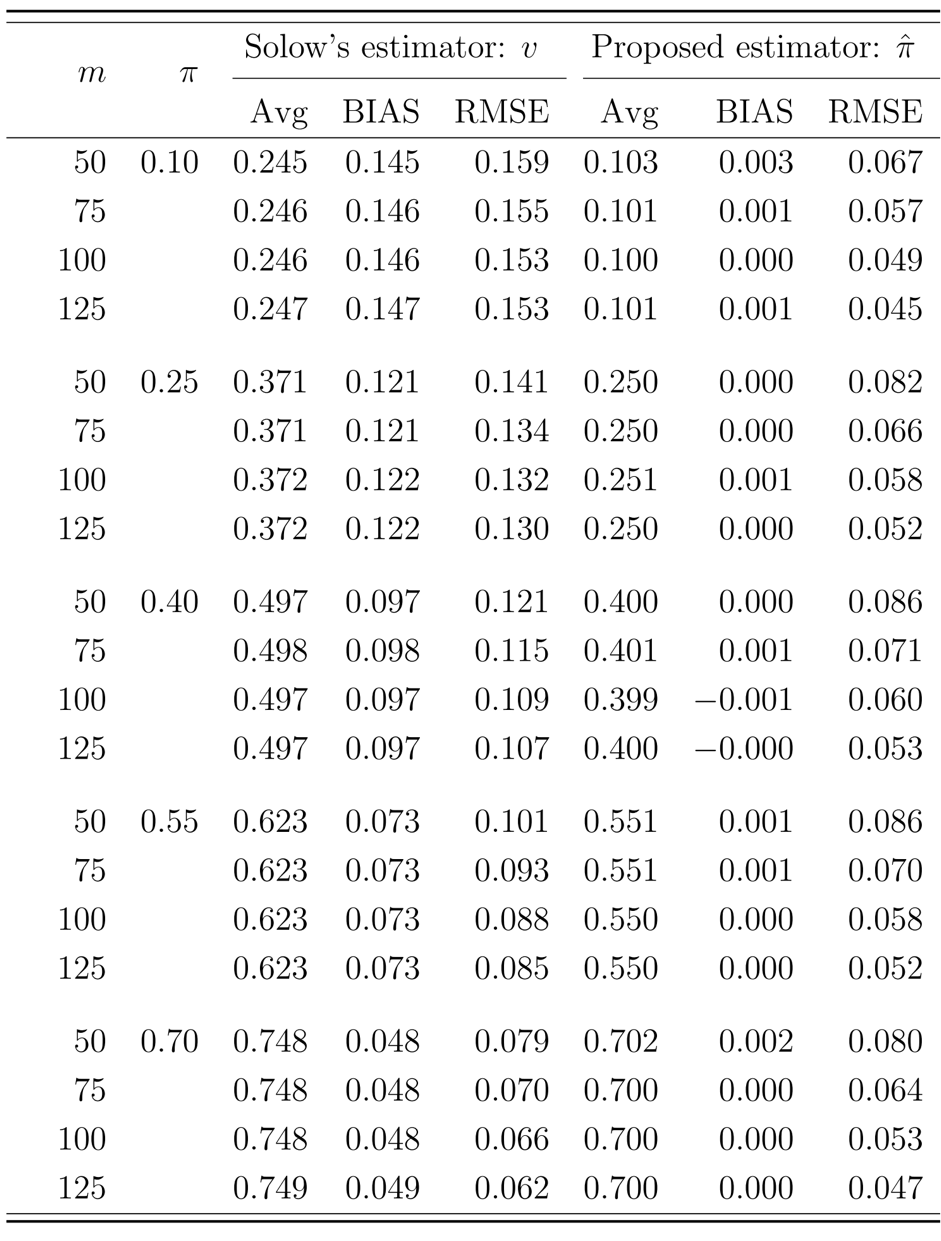


Table S2. Performance comparison between the proposed Rao’s estimator (Eq. 10) accounting for non-independent sampling of subsequent individuals, the original and unbiased Rao’s indices that only assume totally independent sampling of individuals if the sampling of individuals from the studied ecological community (i.e., 26 plant biomasses in ultramafic soils of central Italy) is assumed to be sequentially non-independent.

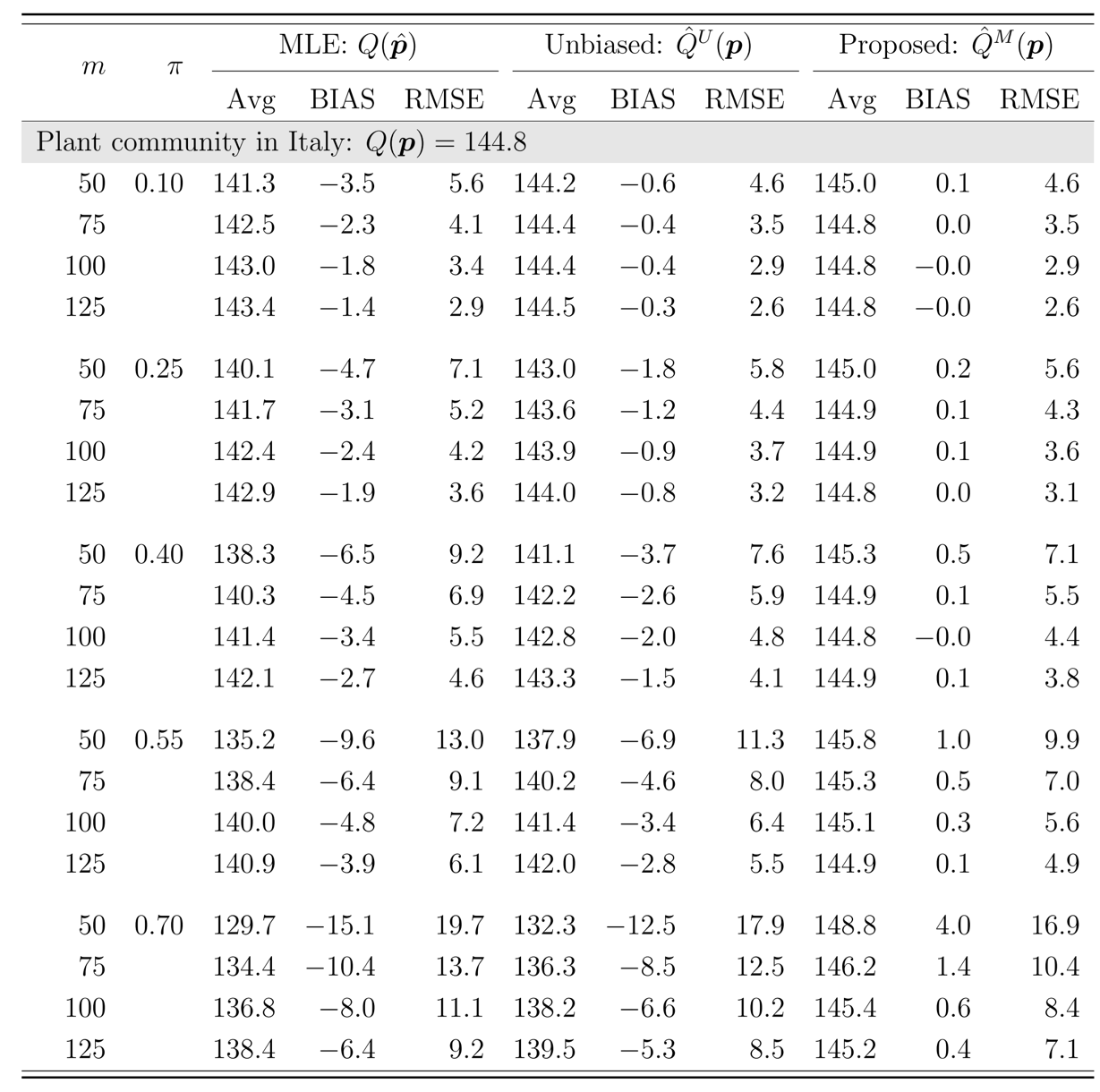


Table S3. Performance comparison between the proposed Gini-Simpson estimator (Eq. 8) accounting for non-independent sampling of subsequent individuals, the original and unbiased Gini-Simpson indices that only assume totally independent sampling of individuals if the sampling of individuals from the studied ecological community (i.e., 26 plant biomasses in ultramafic soils of central Italy) is assumed to be sequentially non-independent.

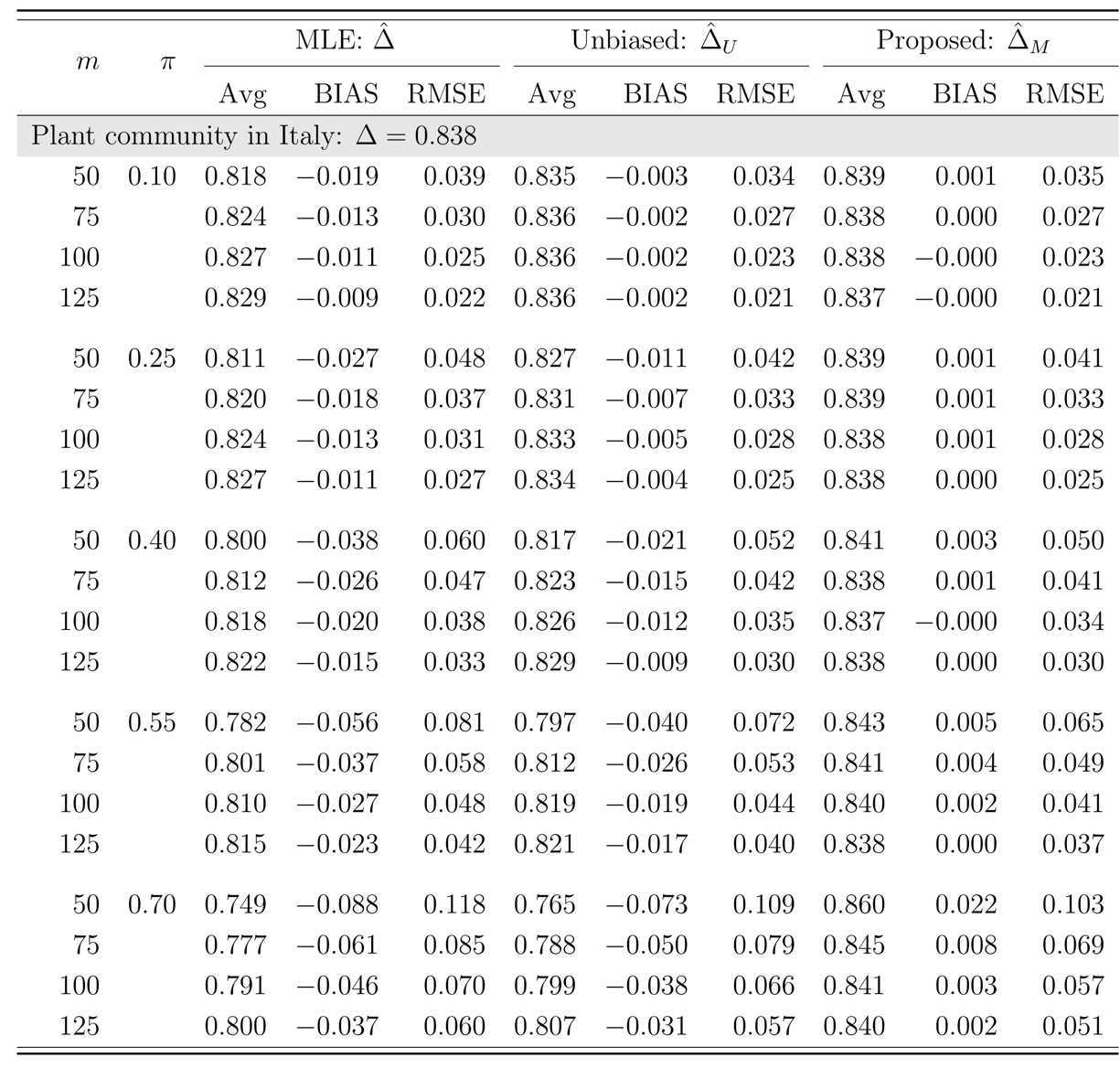


Table S4. Performance comparison between the proposed Gini-Simpson estimator (Eq. 8) accounting for non-independent sampling of subsequent individuals, the original and unbiased Gini-Simpson indices that are only valid for totally independent sampling of individuals if the sampling of individuals from the studied ecological community (i.e., 34 bat Phyllostomid genus abundances in Selva lacandona habitats) is assumed to be sequentially non-independent.

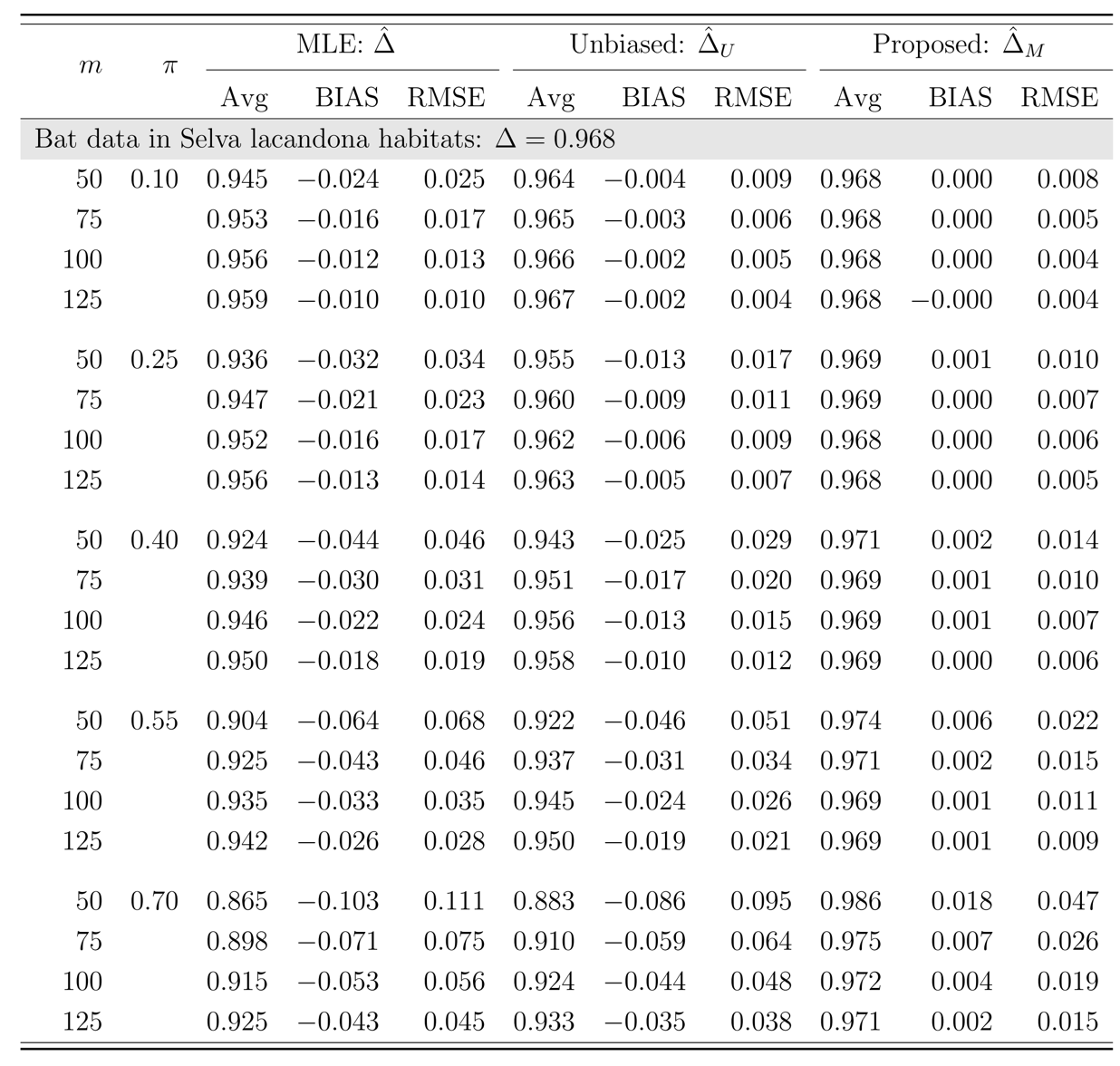


Table S5. Estimate average, averaged bias (BIAS), and RMSE of the ML (maximum likelihood), unbiased Rao’s index, and the proposed estimators in the tree data sampled from the line transects in the 50-ha BCI forest plot.

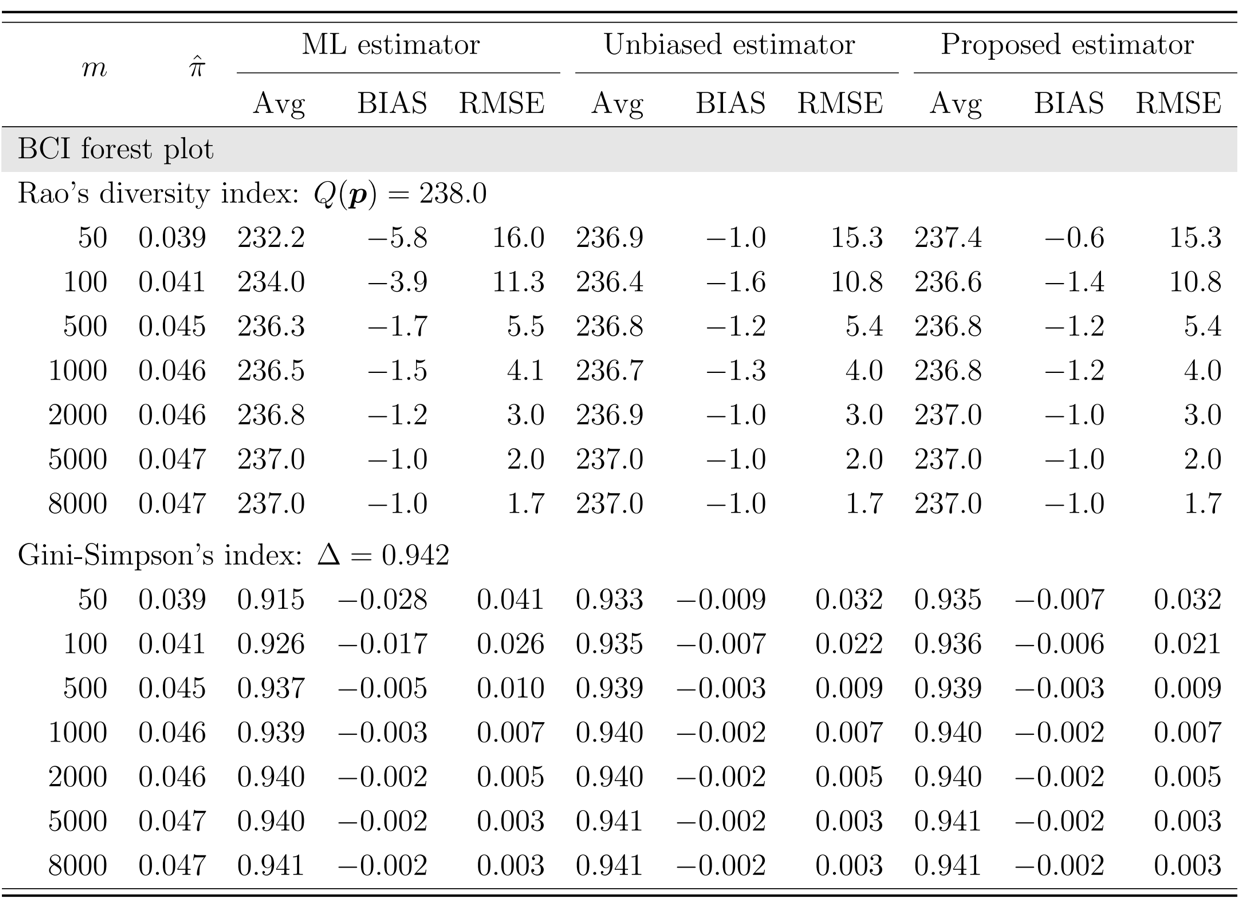


Table S6. Estimate average, averaged bias (BIAS), and RMSE of the ML (maximum likelihood), unbiased Rao’s index, and the proposed estimators in the tree data sampled from the line transects in the 50-ha HSD forest plot.

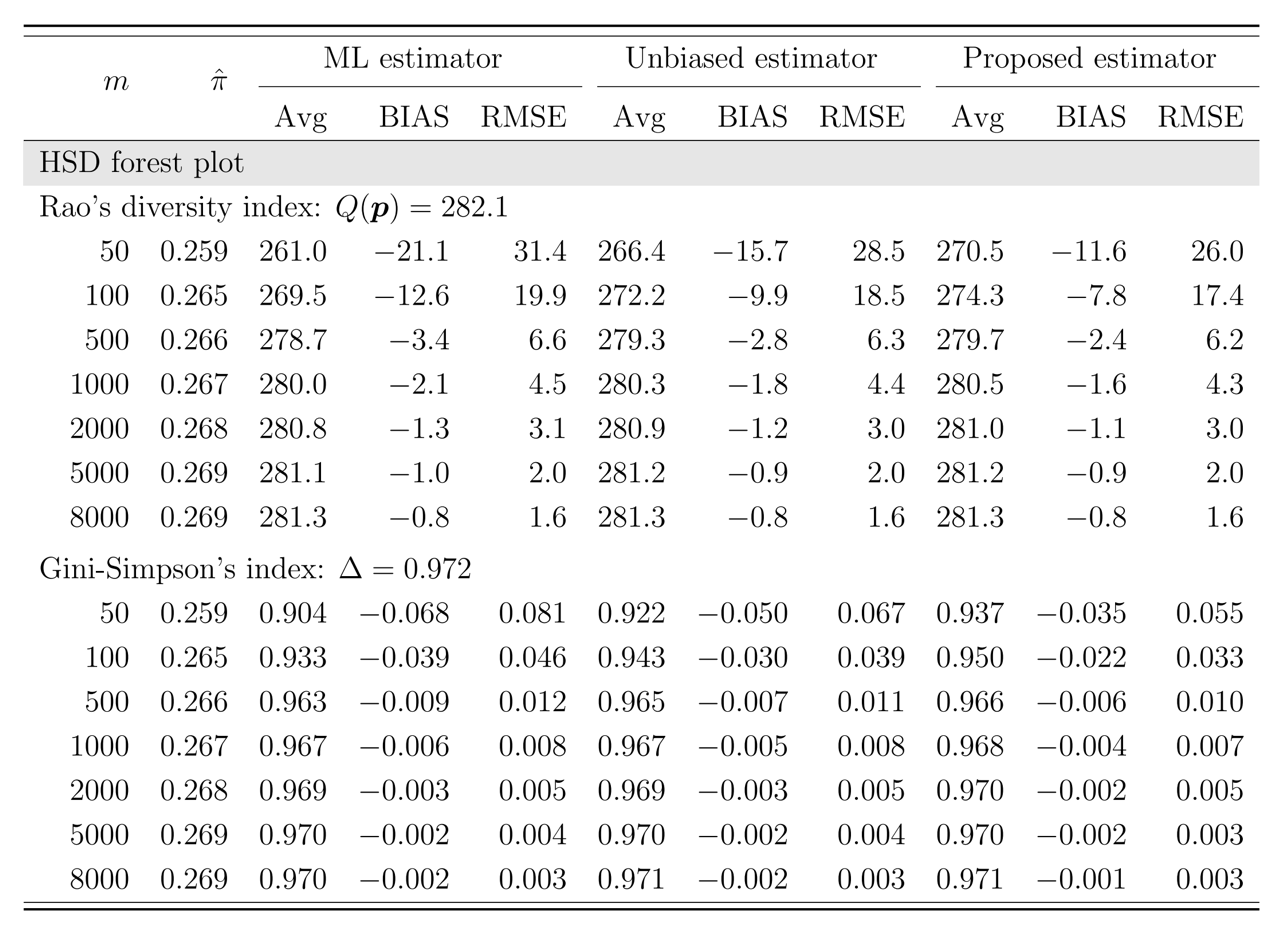


Table S7. Estimate average, averaged bias (BIAS), and RMSE of the ML (maximum likelihood), unbiased Rao’s index, and the proposed estimators from the line transects in the territory of Australia for Acacia species.

