Richness estimation in the existence of species identity error

Chun-Huo Chiu and Jai-Hua Yen

Department of Agronomy, National Taiwan University, Taipei, TAIWAN

Introduction

Species richness is widely used as biodiversity index due to its ecological intuitive concept and simplest form. However, completely species inventories in the wild field are almost unattainable goals. Therefore, the observed richness in the sample always underestimates the true species richness in the assemblage. Among the discussed estimation approaches of species richness, the nonparametric methods are widely used in practical application, which include first order Jackknife approach, second order jackknife approach and Chao1 (or Chao2) lower bound estimator. They all use the observed rare species in the sample (i.e. singletons and doubletons) to estimate the unseen richness in the sample. However, species identity error almost occurred in each survey especially in vegetation sampling was recently discussed in the literatures (Burg et al. 2015, Morrison 2015). Therefore, without error correction, the species richness estimation will be inaccurate based on original sampling data.

Models and Estimators

First Step: estimating mean species identity error rate

To estimate the species identity error, plant inventories from at least one subplot should be independently collected by each investigation team or observer before vegetation sampling. The mean species identity error rate of the investigation teams or observers can be statistically estimated by using the information of these plant inventories. Assume there are S species in the subplot and Y_i is the recorded abundance in the sample, then the expectation of recorded species richness S_r is

$$E(S_r) = \sum_{i=1}^{S} E[I(Y_i > 0)] = \sum_{i=1}^{S} 1 - E[I(Y_i = 0)] = S - \sum_{i=1}^{S} E[I(Y_i = 0)].$$

If plant inventories of the subplot is correct, then S_r should be equal to S species. However, when species identity error occur, S_r may be not equal to S species. Since, when the i-th species is misidentified and other species are also misidentified to the i-th species, then i-th species is not recorded.

Let e_i is the probability that i-th species is mis-identified to other species. And S_j^* is the number of the species set where j-th species be misidentified. Then the expectation of observed richness can be shown as

$$E(S_r) = S - \sum_{i=1}^{S} E[I(Y_i = 0)] = S - \sum_{i=1}^{S} e_i \prod_{j \neq i}^{S_j^*} (1 - \frac{e_j}{S_j^*}). \qquad \text{(A)}$$

To simply, assume e_i and S_j^* are random variables which separately following probability density function f(e) and $g(S^*)$, with mean e_E and S_E^*. Then *Eq.* (A) can be re-formulated as

$$E(S_r) = S - \sum_{i=1}^{S} E[I(X_i = 0)] = S - S \int\int e \times \left(1 - \frac{e}{S^*}\right)^{S^*} dG(S^*)dF(e)$$

$$\approx S + S[e_E \times \left(1 - \frac{e_E}{S_E^*}\right)^{S_E^*}] \approx S[1 - e_E \times exp(-e_E)]. \qquad \text{(B)}$$

Based on similar derivation, we have mean recorded richness of pooled richness of N observers:

$$E(S_{r.pooled}) \approx S(1 - [e_E \times \exp(-e_E)]^N).$$

Then we can estimate the mean species identity error rate by taking

$$\frac{E(S_{r.pooled})}{E(S_r)} = \frac{S(1 - \{e_E \times exp[-e_E]\}^N)}{S\{1 - e_E \times exp[-e_E]\}} = \frac{1 - \{e_E \times exp[-e_E]\}^N}{1 - e_E \times exp[-e_E]}$$

By solving the equation, we can get the estimate of mean species identity error rate $\widehat{e_E}$.

Second Step: Adjusting observed richness and species frequency counts of rare species

For the interesting region, the sample data were collected by different investigating teams or observers. However, due to sampling limitation, there are undetected richness in every survey. So, true richness is equal to observed richness plus unseen richness. Under the existence of species identity error, how to adjusted observed richness and frequency counts of rare species in the sample which are used to estimate undetected richness is the purpose of this step. Let X_i be the sample abundance of i-th species, i =1, 2,…, S. Thus, $S_{obs} = \sum_{i=1}^{S} I(X_i > 0)$ is the observed species richness, $f_1 = \sum_{i=1}^{S} I(X_i = 1)$ is singleton richness and $f_2 = \sum_{i=1}^{S} I(X_i = 2)$ is doubleton richness. With parallel derivations, the adjusted observed richness, singletons and doubletons can be separately obtained by

$$S_{obs.j} \approx \frac{S_{obs}}{1 - \widehat{e_E} \times exp(-\widehat{e_E})}$$

$$f_{1.j} \approx \frac{f_1}{1 - \widehat{e_E} \times exp(-\widehat{e_E})},$$

and

$$f_{2.j} \approx \frac{f_2}{1 - \widehat{e_E} \times exp(-\widehat{e_E})}.$$

Third Step: Estimating richness based on adjusted statistics

Therefore, the species richness in the interesting region can be estimated by using nonparametric approaches based on the adjusted observed richness, singleton richness and doubleton richness.

$$\hat{S} = S_{obs.j} + \frac{{f_{1.j}}^2}{2f_{2.j}} = \frac{S_{obs}}{1 - \widehat{e_E} \times exp(-\widehat{e_E})} + \frac{[\frac{f_1}{1 - \widehat{e_E} \times exp(-\widehat{e_E})}]^2}{\frac{2f_2}{1 - \widehat{e_E} \times exp(-\widehat{e_E})}}.$$

$$= \frac{S_{obs} + \frac{{f_1}^2}{2f_2}}{1 - \widehat{e_E} \times exp(-\widehat{e_E})}.$$

Simulation Study

To test the performance of the estimator proposed, we present the simulation results into two parts.

A. Mean error rate estimation

Set species richness in the subplot (S_{sub}) as 30, and assume species identity error rate (e) follows from random uniform with mean equal to 0.1, 0.2, and 0.3, respectively.

The average error rate estimate of 200 simulated data and its variance estimator estimated by bootstrapping approach data are given in Table 1 and Table 2.

Table 1. Estimation of species identity error rate, 200 simulation trials with bootstrapped 200 times $S_{sub} = 30$, $N = 5$.

True e	Estimated e	Sample s.e.	Estimated s.e.
0.095	0.098	0.024	0.026
0.198	0.199	0.035	0.039
0.291	0.298	0.048	0.052

Table 2. Estimation of species identity error rate, 200 simulation trials with bootstrapped 200 times $S_{sub} = 30$, $N = 10$.

True e	Estimated e	Sample s.e.	Estimated s.e.
0.095	0.096	0.016	0.018
0.198	0.197	0.024	0.027
0.291	0.292	0.032	0.035

Figure 1 and Figure 2 presents the bias performance of the estimator under different richness setting. Figure 3 and Figure 4 presents the bias performance of the estimator under different setting of number of observers.

Figure 1. Species identity error rate estimation bias with different species richness. 200 trials, $N = 5$, species identity error rate follows Uniform(0.05,0.15).

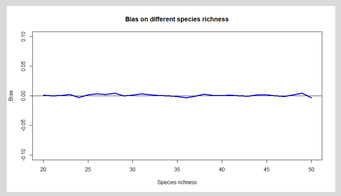


Figure 2. Species identity error rate estimation bias with different species richness. 200 trials, $N = 5$, species identity error rate follows Uniform(0.25,0.35).

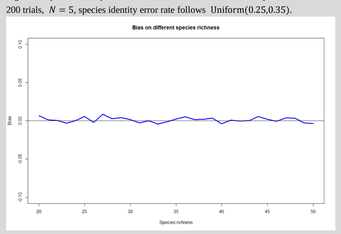


Figure 3. Species identity error rate estimation bias with different number of observers. 200 trials, $S_{sub} = 30$, species identity error rate follows Uniform(0.05,0.15).

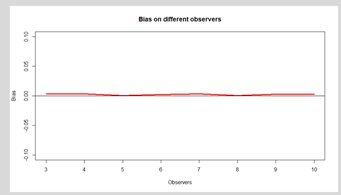
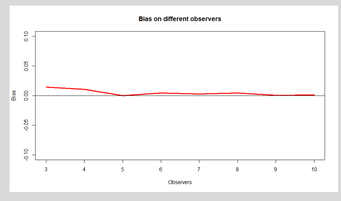


Figure 4. Species identity error rate estimation bias with different number of observers. 200 trials, $S_{sub} = 30$, species identity error rate follows Uniform(0.25,0.35).



B. Species richness estimation

When the mean species identity error rate was estimated, we used the species identity error rate to adjust the error species richness. We define some notations:

$\hat{S}_{true}$: "species richness estimator without species identity error"

$S_{obs.True}$: "observed species richness without species identity error"

$f_{1.True}$: "singleton richness without species identity error"

$f_{2.True}$: "doubleton richness without species identity error"

Table 3. Estimate species richness after estimating mean species identity error rate in subplot, 200 simulation trials, $S = 100$ with detection rate follows Uniform(0.01,0.3), $S_{sub} = 30$, $N = 5$, $\hat{S}_{True} = 91.1$, $S_{obs.True} = 55$, $f_{1.True} = 33$, and $f_{2.True} = 16$.

True e	Estimated e	Method	$\hat{S}$	S_{obs}	f_1	f_2
0.097	0.099	Adjusted	84	57	32	17
		Uncorrected	76.5	52	29	16
0.203	0.204	Adjusted	85.9	59	32	18
		Uncorrected	71.7	50	27	15
0.295	0.296	Adjusted	86.3	61	32	18
		Uncorrected	67.3	48	25	14

Table 4. Estimate species richness after estimating mean species identity error rate in subplot, 200 simulation trials, $S = 100$ with detection rate follows Uniform(0.01,0.3), $S_{sub} = 30$, $N = 10$, $\hat{S}_{True} = 92$, $S_{obs.True} = 74$, $f_{1.True} = 26$, and $f_{2.True} = 21$.

True e	Estimated e	Method	$\hat{S}$	S_{obs}	f_1	f_2
0.097	0.097	Adjusted	90.5	76	25	21
		Uncorrected	82.6	69	23	19
0.203	0.206	Adjusted	91.6	78	24	21
		Uncorrected	76.4	65	20	18
0.295	0.296	Adjusted	91.8	79	23	21
		Uncorrected	71.6	61	18	16

Based on the results of simulation study, we have following short conclusions:

1. The estimator of species identity error rate is nearly unbiased in different richness of subplots or different number of observers.
2. The observed richness, singletons and doubletons with species identity error are always lower than their corresponding statistics without species identity error. The adjusted observed richness, singletons and doubletons are nearly unbiased.
3. The estimator based on adjusted observed richness and species frequency counts can efficiently reduce the negative bias of species richness estimator based on the raw data.

Real Data Example

The real data set of alpine meadows in the Swiss Alps was released in the published paper (Vittoz and Guisan 2007). There are total six observers participated in this wild field work. The original study is about to test the reliable in different survey design. Here, we use the four inventories of subplot (0.4 m^2) shown in Figure 5 to estimate the mean species identity error rate among six observers.

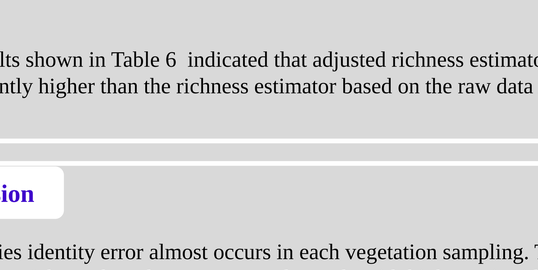


Figure 5. Plots used in this survey.

The dissimilarity of pooled subplots for each pair observers are shown in Table 5. the estimate of species identity error rate is 0.204 shown in Table 6. Based on the sampling data in the region D, we compare the estimates based on raw data and adjusted observed species richness, singletons and doubletons.

Table 5. Dissimilarity (1 − Jaccard index) for the pooled data in A_1 to A_4 subplot (0.4 m^2 each) from Vittoz (2007).

OBSERVERS	A	B	D	E	F	H
A	0	0.182	0.088	0.147	0.212	0.152
B		0	0.206	0.212	0.226	0.219
D			0	0.171	0.182	0.182
E				0	0.188	0.286
F					0	0.194
H						0

Table 6. Use species identity error rate from subplot A to adjust species richness in plot D. $\widehat{e_E} = 0.204$.

	$\hat{S}$ (95% CI)	S_{obs}	f_1	f_2
Adjusted	62.7	61	5	6
	(61.249,72.588)			
Uncorrected	52.6	51	4	5
	(51.260,60.812)			

The results shown in Table 6 indicated that adjusted richness estimator has significantly higher than the richness estimator based on the raw data in the region D.

Conclusion

1. Species identity error almost occurs in each vegetation sampling. The estimated species richness based on uncorrected sample with high species identity error will seriously underestimate the true species richness of the assemblage.
2. In this study, we proposed an nearly unbiased estimator to estimate mean species identity error rate based on pre-inventories survey work of subplot.
3. Based on similar deviation, we adjusted the observed richness, singletons and doubletons in the sample with identity error.
4. Using adjusted statistics, the richness estimator could efficiently reduce the negative bias of richness estimator based on raw data.

Reference

Burg, S., Rixen, C., Stöckli, V. & Wipf, S. (2015) Observation bias and its causes in botanical surveys on high-alpine summits. *Journal of Vegetation Science* **26**, 191–200.

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

Morrison, LW. (2015) Observer error in vegetation surveys: a review. *Journal of Plant Ecology* **9**, 367–379.

Vittoz, P. & Guisan, A. (2007) How reliable is the monitoring of permanent vegetation plots? A test with multiple observers. *Journal of Vegetation Science* **18**, 413–422.