Species richness estimation by using machine learning algorithms

Chun-Huo Chiu* and Min-Shan Lin
chchiu2017@ntu.edu.tw
Department of Agronomy, National Taiwan University, Taipei, TAIWAN

National Taiwan University

Introduction

Accurate estimation of richness is always a challenge in statistics due to sampling resource limitations. Many richness estimators were proposed in the literature to address the underestimating problem of observed richness in the sample, where Chao1 and Jackknife estimators are most widely used due to no assumption on species composition. However, these estimators are seriously underestimated in the sample with small size or the community with high heterogeneity.

Since estimating richness given a random sample is basically a prediction question. In this study, we use machine learning(ML) algorithms to estimate the true richness in a defined area. First, we develop training datasets by computer simulation based on Chao1's 95% confidence interval and adjusted sample species relative composition by using sample coverage. Second, we select the important features based on the concept of the Good-Turing frequency formula.

We evaluate the statistical behaviors of four high frequently used ML algorithms including Ridge Regression, K Nearest Neighbors, Random Forest, and Boosting. The simulation results show that four ML methods have lower bias and RMSE than the nonparametric estimators, while there is no difference in statistical performance among these four ML algorithms. Hence, Ridge Regression and Random Forest are recommended for the reason of shorter computational time.

Methods

Assume there are S species in the community with relative composition $(p_1, p_2, \dots, p_S)$ and X_i is the abundance of i-th species in the sample. When n individuals are randomly sampled from community, then $(X_1, X_2, \dots, X_S) \sim Multinomial(n, p_1, p_2, \dots, p_S)$. Let $f_k = \sum_{i=1}^{S} I(X_i = k)$ be the number of species that exactly detected k individuals in the sample, where f_0 is undetected richness and $S_{obs} = \sum_{k=1} f_k$ is the observed richness.

To predict true richness given a sample by using ML algorithms, training dataset and important features are needed to develop the richness estimation machine. First, we generate training datasets by computer simulation based on a 95% confidence interval of Chao1 estimate and adjusted sample species relative composition. Second, we select the potential features based on the concept of the Good-Turing frequency formula.

Step 1: generate training data set

Given a species abundance sample with size n and S_{obs} observed richness, assume the undectected richness f_0 is ranged by 95% C.I. of Chao1 estimate $[\widehat{f_{0L}}, \widehat{f_{0,U}}]$, where

$$\widehat{f_0} = S_{obs} + \frac{n-1}{n}\frac{f_1^2}{2f_2}, \widehat{f_{0,U}} = \widehat{f_0} \times exp\left\{1.96\left[log\left(1+\frac{\widehat{var}(\hat{S})}{\widehat{f_0}^2}\right)\right]^{1/2}\right\} \text{ and } \widehat{f_{0,L}} = \widehat{f_0} / exp\left\{1.96\left[log\left(1+\frac{\widehat{var}(\hat{S})}{\widehat{f_0}^2}\right)\right]^{1/2}\right\}$$

(1) Random choose a value from this range denoted as f_0, then $S_{obs} + f_0 = S_m$ is the true richness of mth training data.
(2) Construct species composition $(p_1^*, \dots, p_{S_{obs}}^*, p_{S_{obs}+1}^*, \dots, p_{S_m}^*)$, where $\sum_{i=1}^{S_{obs}} p_i^* = 1 - f_1/n$ and $\sum_{i=S_{obs}+1}^{S_m} p_i^* = f_1/n$
(3) Random generate a species abundance random sample $(X_{1m}, X_{2m}, \dots, X_{S_m m})$ from $Multinomial(n, p_1^*, \dots, p_{S_{obs}}^*, \dots, p_{S_m}^*)$ and calculate the unseen richness as $S_m - \sum_{i=1}^{S_m} I(X_i^* > 0)$ denoted as $f_{0,m}$
(4) Repeat step(1)~step(3) M times to generate a training data set with size M.

Step 2: select the important features

According to the concept of Good-Turing frequency formula that imply the rare species contains most information about unseen species. We select the first k rarest species frequency counts $((f_{1,m}, f_{2,m}, \dots, f_{k,m})$ as the potential features(predictors).

Step 3: develop richness estimation machine

Based on Step1 and Step2, we reorganize the format of training dataset before training model.

$$\begin{bmatrix} X_{11}, X_{2,1} \dots, X_{S_1 1} \\ \vdots \\ X_{1M}, X_{2M} \dots, X_{S_M M} \end{bmatrix} \Rightarrow \overbrace{\begin{bmatrix} f_{1,1}, f_{2,1} \dots, f_{k,1} \\ \vdots \\ f_{1,m}, f_{2,m} \dots, f_{k,m} \end{bmatrix}}^{X} \overbrace{\begin{bmatrix} f_{0,1} \\ \vdots \\ f_{0,m} \end{bmatrix}}^{Y}$$

, where X are explanatory variables and Y is the response variable.

Then, we applied 4 common machine learning techniques: Ridge Regression , K Nearest Neighbor, Random Forest and adaptive Boosting to develop richness estimation machine to predict the richness of undetected species.

Develop Richness estimation Machine

Different widely used ecological species-abundance models are used to develop richness estimation machine. The number of species in each model was fixed at $S = 200$. Four sample size (200, 400, 800, 1200) were considered, resulting in total 16 model–size combinational scenarios. For each scenario, 500 simulated data are generated, using bias and RMSE as selection criteria and using cross-validation for variables selection, decision of the size of training data, and selection of ML algorithm.

The Influential Factors of the Machine Learning Performance

The Candidate Set of Variables	Training Data Size(M)	Machine Learning Model
$f_1, f_2, \dots, f_5$	500	Ridge regression
$f_1, f_2, \dots, f_{10}$	1000	Random forest
$f_1, f_2, \dots, f_{15}$	2000	K nearest neighbor
$f_1, f_2, \dots, f_5, n$	5000	Adaptive boosting
$f_1, f_2, \dots, f_5, \hat{C}$		
$f_1, f_2, \dots, f_5, \widehat{f_0}$		

(shaded): the recommended setting for machine learning method

Simulation Study

To investigate the statistical behavior of four learning machine algorithms and compare them with two nonparametric methods: the Chao1 and 1st Jackknife estimator.

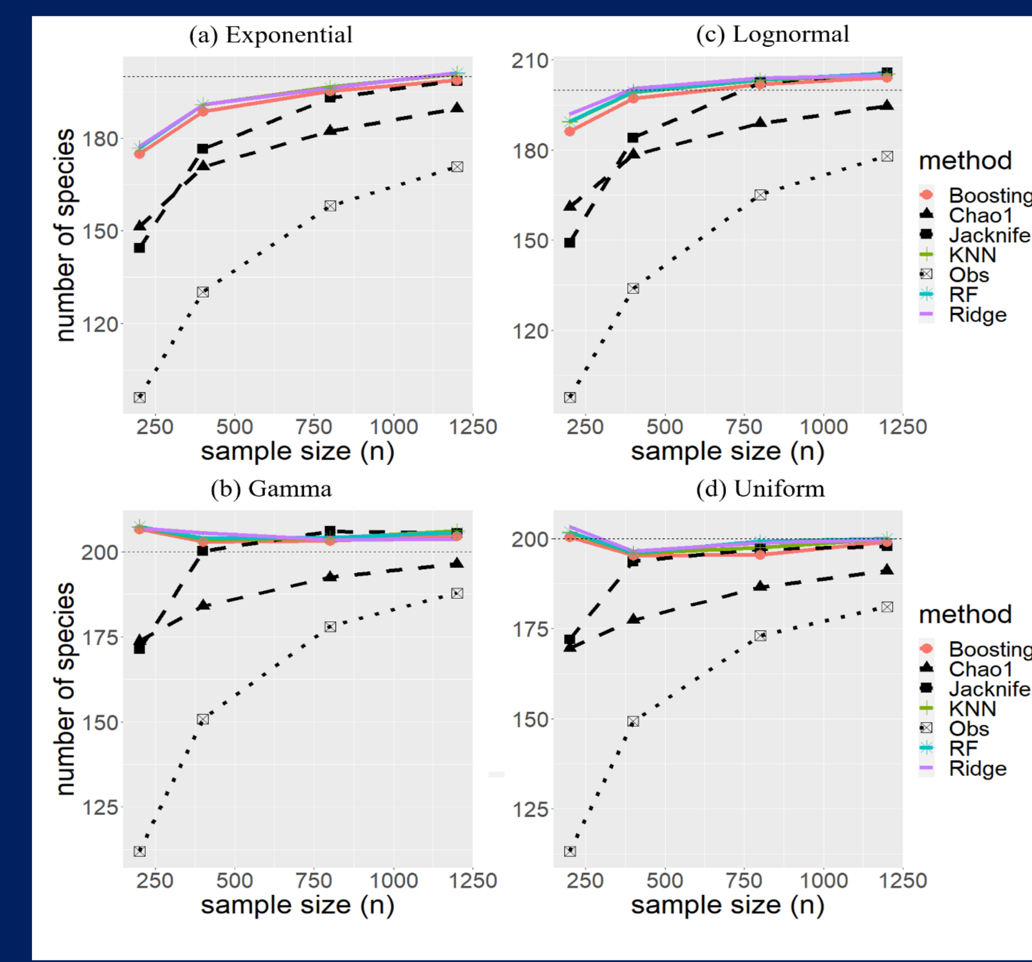

Figure 1. the averaged estimate over 500 datasets

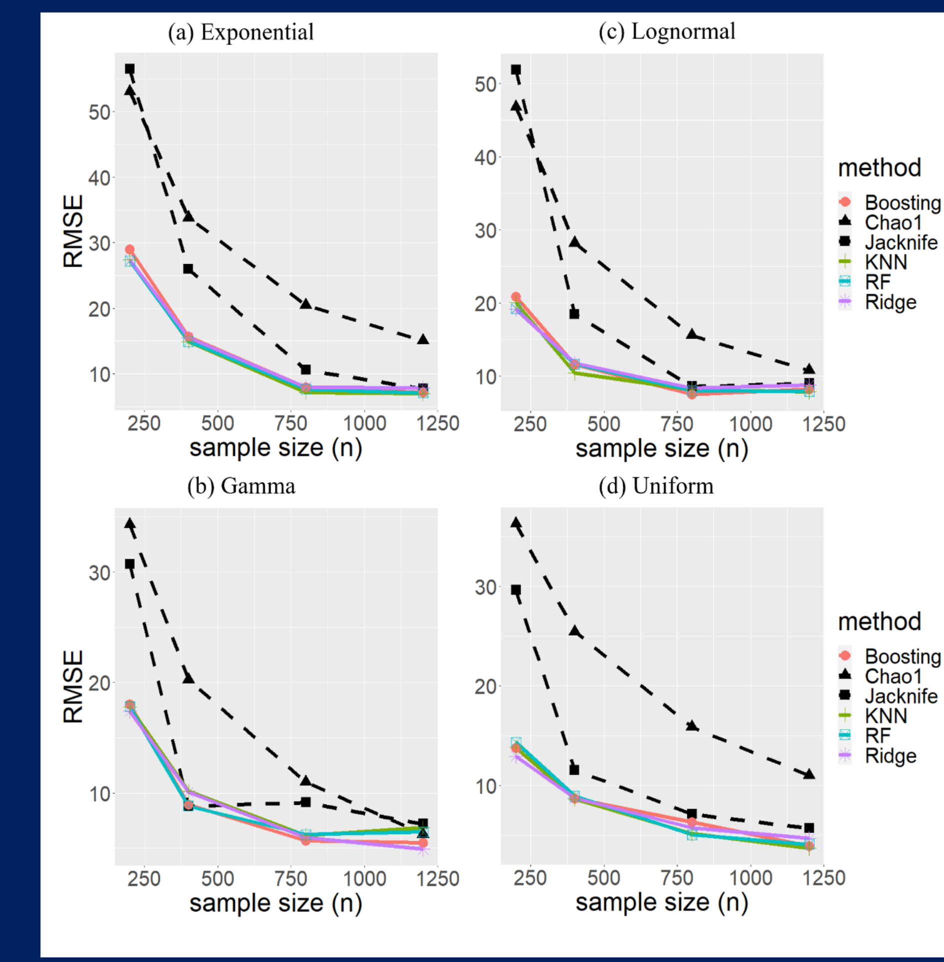

Figure 2. the RMSE of estimator over 500 datasets

Conclusions and Discussions

1. The first fifteen rare species frequency counts $(f_1, f_2, \dots, f_{15})$ as an explanatory variables are recommended as the features(predictors) for richness estimation machine and training dataset with size 500 is sufficient.
2. The Bias and RMSE of the discussed ML algorithms are similar. Ridge Regression and Random Forest are recommended for the reason of efficiency.
3. The developed ML methods perform well over traditional nonparametric estimators especially for the sample with small size.
4. The standard error of ML method could be estimated by using bootstrapping method.

Reference

1. Breiman, L. 2001. Random forests. *Machine learning* 45:5-32.
2. Chao, A. (1984). Nonparametric estimation of the number of classes in a population. *Scandinavian Journal of Statistics*, 11, 265-270.
3. Freund, Y., R. Schapire, and N. Abe. 1999. A short introduction to boosting. Journal-Japanese Society For Artificial Intelligence 14:1612.
4. Friedman, J. H., F. Baskett, and L. J. Shustek. 1975. An algorithm for finding nearest neighbors. IEEE Transactions on computers 100:1000-1006.
5. Good, I. J. 1953. The population frequencies of species and the estimation of population parameters. Biometrika 40: 237-264.
6. Marquardt, D. W., and R. D. Snee. 1975. Ridge regression in practice. The American Statistician 29:3-20.