How to predict newly found rare species in additional ecological samples using RSE package?

Youhua Chen[1], Tsung-Jen Shen[2]

1, CAS Key Laboratory of Mountain Ecological Restoration and Bioresource Utilization & Ecological Restoration and Biodiversity Conservation Key Laboratory of Sichuan Province, Chengdu Institute of Biology, Chinese Academy of Sciences, Chengdu, 610041, China

2, Institute of Statistics & Department of Applied Mathematics, National Chung Hsing University, 250 Kuo Kuang Road, Taichung 40227, Taiwan

2018-08-05

Introduction

Biodiversity survey or ecological sampling is usually time-consuming, scale-limited and labor-extensive. In most cases, ecologists will encounter the limiting-sampling problem when conducting field or lab experiments. As a consequence, practically, it becomes natural and important to use limited data derived from restricted sampling efforts to predict and explain biodiversity patterns as accurately as possible.

One of the unsolved practical problems confronted by conservationists and ecologists is: How many newly found species in additional ecological samples are likely to be rare (i.e., species with extremely either small populations or narrow range sizes)? The key reason for proving the necessity of addressing this question is that rare species are one of the most important components in biological diversity (Kunin & Gaston 1993; Gaston 2012), play essential roles in ecosystem functioning (Lyons et al. 2005; Mouillot et al. 2013; Jain et al. 2014; Leitao et al. 2016), and become a leading indicator in extinction risk assessment (Davies et al. 2000; Fattorini 2014; Chen et al. 2017). In conservation biology, if there are many rare species in the originally sampled communities (directly observed from the field data) and neighboring unsampled communities (which can be predicted using the estimators in RSE package), conservation priorities of these communities would be expected to stand on top-tier levels (Prendergast et al. 1993; Chen et al. 2017).

The *RSE* package is developed with the primary purpose of addressing this practical question resolved by rigorous but user-friendly statistical methods (detailed introduction of these methods are presented below). The data requirement of the package is very simple and common: species abundance or

incidence information. Additionally, the usage of the package is easy, but, as a get-started vignette presented here, some necessary installation steps and functions are introduced in the following sections.

Package installation

Before using the package, it is advised to download the source file (**RSE_1.1.tar.gz**) from the Github repository in the following link: https://github.com/ecomol/RSE, save the downloaded source file into a local directory of your computer.

To install the package, you need to open R software (if you do not have one yet, please download and install it from https://www.r-project.org/) and to type the following command to install the package:

```
install.packages("your_local_directory/RSE_1.1.tar.gz", repos=NULL,type
="source")
```

Note that, here, you need to replace "your_local_directory" by the exact computer path where you saved the installation package in your own computer. For instance, if you save the package file **RSE_1.1.tar.gz** in a default folder called "programs" of disk C in Windows operation system (that is the most common setting), then the exact path should be "C://programs/RSE_1.1.tar.gz", so you must replace "your_local_directory" by "C://programs". Note that, of course, you can save *RSE* package into any self-created folder of your computer, but the setting of "your_local_directory" should be altered accordingly.

If being successfully installed, you should be able to load the installed *RSE* package by typing:

```
library(RSE)
```

If no errors are reported after executing the command above, that means you have successfully installed and loaded the package already in your R environment.

Package usage

Data input requirement and data manipulation process for analyses

Two data types are accepted for implementing the statistical estimators available in *RSE* package. The first data type is species abundance data, while the second one is species incidence data. When applying *RSE* package, the input data format is extremely simple and flexible. For example, in a sampled plot, if you identified 8 species, collected the number of individuals of each species, and recorded them by a vector as (1, 1, 2, 3, 3, 5, 7, 10). Then, this is an appropriate data format that can be loaded into R for further analyses in *RSE* package.

```
dat = c(1,1,2,3,3,5,7,10)
```

Note that, for the two real conservation-directed examples demonstrated in *RSE* package, their input data formats actually are data matrices, but the specific analyses are conducted on a selected column (thus being a vector again); please refer to the manual of the package for testing the examples accompanied with the package, or alternatively see the demonstration about how to apply the methods to these two data at the end of this vignette.

When you have loaded your own data in the R environment and recall that you have loaded *RSE* package previously, it is ready to convert abundance or incidence data into frequency count data. To be specific, that can be accomplished by typing the following command with using *X.to.f()* function to convert a vector of sampled species abundance or incidence data to frequency data as follows:

```
f = X.to.f(dat)
print(f)

##  [1] 2 1 2 0 1 0 1 0 0 1
```

Also, you need to estimate two parameters that are necessary for predicting the true relative abundances of species using your sampled data. The estimation steps followed the method in Chao et al. (2015)'s paper. Readers interested in the derivation are encouraged to read the paper for further detailed information, here we did not discuss too much in the vignette. The estimation of true relative abundance-related parameters can be accomplished by using the function *DetAbu()* for species abundance data or *DetInc()* for species incidence data in *RSE* package. For the former case, you should type the following command:

```
b = DetAbu(x=dat, zero=FALSE)
print(b)

## [1] 6.057164e-02 6.610696e-05
```

The transformed data vector *f* and the estimated parameter vector *b* now are ready for subsequent analyses by using different estimators to predict the number

of newly found rare species. Relevant details are shown after that we have introduced some more about the statistical estimators available in *RSE* package.

Doing predictions based on observed frequency count data

In *RSE* package, there are three statistical methods currently available for predicting the number of newly found rare species in additional samples. Taking the Bayesian-weight estimator as an example, recall that you have converted your original 8-species abundance dataset to frequency count data. Now, if you are planning to sample an additional sample (suppose the sample size is 20), how many new rare species can you expect to observe in that sample before conducting a field survey? Using the statistical estimator in *RSE* package can tell you the answer immediately as follows:

```
Pred.Fk.BW(f=f, m=20, b=b, k.show=2)

## [1] 0.6888080 0.1613667
```

Note that here *f* and *b* respectively represent the transformed species frequency count vector and the estimated parameter vector used to estimate the true relative abundance of species in your original data (1, 1, 2, 3, 3, 5, 7, 10). Also, the argument *k.show* is to organize the display about the estimated result of the numbers of extremely rare species with abundance <= k.show in the additional sample. In the above case, the result returned is of two numeric values, of which the first one is the expected number of new singleton species in the additional sample, while the second value indicates the expected number of newly found doubleton species in the additional sample. Based on the result demonstrated, it indicates that, on average, there is around one new singleton species found in any additional sample of size 20 including the one you are planning to survey.

Doing predictions using two batch functions *Pred.abundance.rare* and *Pred.incidence.rare*

In practice, we did not need to conduct such detailed step-by-step analyses as shown above, we actually can use two batch-operating functions to get all the estimation results we need: *Pred.abundance.rare()* and *Pred.incidence.rare()*. The former is to conduct predictions for abundance data while the latter is for incidence data. These two functions are very easy to use and they return the predictions using all the available statistical methods in the package simultaneously; moreover, users do not need to fit the parameters related to the true relative abundance or species incidence probability like above. Finally, they are equipped with appropriate bootstrapping procedures to provide 95% bootstrap confidence intervals for the corresponding point estimates.

To showcase the usage of these two batch functions, for the hypothetical abundance example discussed here, suppose we want to predict the number of

singleton and doubleton species in an additional sample with a size of 30, we can type the following command:

```
Pred.abundance.rare(boot.rep=1000, xi=dat, m=30, k.show=2)
```

```
## $`Data information`
##   Original sample size (n) Additional sample size (m)
##                         32                         30
##
## $`Naive method`
##      k Estimate Estimated SE 95% lower limit 95% upper limit
## [1,] 1      0.3          0.2               0             0.6
## [2,] 2      0.2          0.1               0             0.3
##
## $`Bayesian-weight method`
##      k Estimate Estimated SE 95% lower limit 95% upper limit
## [1,] 1      0.8          0.6               0             1.9
## [2,] 2      0.3          0.1               0             0.5
##
## $`Unweighted method`
##      k Estimate Estimated SE 95% lower limit 95% upper limit
## [1,] 1      0.6          0.6               0             1.8
## [2,] 2      0.3          0.1               0             0.6
```

The estimation from all the three available statistical methods (which will be introduced in more details below) in the package are returned. Moreover, the 95% confidence intervals are provided. They are derived from a bootstrapping procedure with 1000 replicates, which can be realized by setting the parameter "*boot.rep=1000*" in the function *Pred.abundance.rare()*. Parameter "*k.show=2*"in the above script is to ask the function to return the estimate of singleton and doubleton species numbers. Parameter "*m=30*" is to tell the function that the additional sample has a size of 30. Finally, parameter "*xi=dat*" is to tell the function that the original sample has observed species frequency count presented in the "*dat*" variable.

Statistical methods available in *RSE* package

Bayesian-weight estimator

The Bayesian-weight estimator allows ecologists to completely utilize the observed species abundance information from the original sample to predict the number of rare species newly found in an additional sample while skillfully dodging the usage of the information of unseen species that are totally unknown from the original sample. In *RSE* package, for species abundance data, the R function implementing the Bayesian-weight estimator is by *Pred.Fk.BW()*, for species incidence data, the R function implementing the Bayesian-weight estimator is by

Pred.Qk.BW(). Readers are encouraged to read the introductions about the two functions, related meanings of the parameters and data input requirement in the package manual. Or, if you have loaded the package in your R environment, you can type the following commands to check the detailed introduction of the two functions:

```
?Pred.Fk.BW
?Pred.Qk.BW
```

Unweighted estimator by Chao et al.'s

Chao et al.'s unweighted estimator can be derived from Chao et al. (2015)'s paper. The method is simple, but its performance is reasonably good. In RSE package, for species abundance data, the R function implementing the Chao et al.'s unweighted estimator is by *Pred.Fk.unweighted()*, for species incidence data, the R function implementing the Chao et al.'s unweighted estimator is by *Pred.Qk.unweighted()*. Readers are encouraged to read the introductions about the two functions, related meanings of the parameters and data input requirement in the package manual. Alternatively, if you have loaded the package in your R environment, you can type the following commands to check the detailed introduction of the two functions:

```
?Pred.Fk.unweighted
?Pred.Qk.unweighted
```

Naive unweighted estimator

Finally, we also presented a naive estimator, in which the mathematical foundation is to use an intuitive and direct maximum likelihood-based estimate of species' relative abundance to estimate the number of newly found rare species. In RSE package, for species abundance data, the R function implementing the naive estimator is by *Pred.Fk.Naive()*, while for species incidence data, the R function implementing the naive estimator is by *Pred.Qk.Naive()*. Readers are encouraged to read the introductions about the two functions, related meanings of the parameters and data input requirement in the package manual. Alternatively, if you have loaded the package in your R environment, you can type the following commands to check the detailed introduction of the two functions:

```
?Pred.Fk.Naive
?Pred.Qk.Naive
```

Important message: In addition to the above functions specifically designed for each estimator, as mentioned previously, we also provide two very helpful functions for users to use: *Pred.abundance.rare()* and *Pred.incidence.rare()*. They are very easy to use. More importantly, they are equipped with appropriate bootstrapping procedures to provide 95% bootstrap confidence intervals for the point estimates. Because we will also demonstrate them in the later part of this vignette, to be familiar with these two functions better, please type the following commands to see details:

```
?Pred.abundance.rare
?Pred.incidence.rare
```

Available datasets for demonstration in *RSE* package

In *RSE* package, we have provided two empirical datasets that have been tested in our paper (Predicting the number of newly discovered rare species: a Bayesian weight approach: under review in *Conservation Biology*). The users of the package are encouraged to use these two datasets for being familiar with the package. The two datasets can be loaded in the R environment by typing the following commands:

```
data(HerpetologicalData)
data(CanadaMite)
```

The first dataset contains abundance information of 62 amphibians (including anuran, lizard, snake, and turtle) in the conserved and human disturbed areas of Neotropical dry forests of Mexico (Suazo-Ortuno et al., 2008). In the conserved area, there are 50 species with 779 individuals, while in the disturbed areas, 48 amphibian species are identified from 876 individuals. To take a quick view on the first six rows of the data set, we can execute the following command in R:

```
head(HerpetologicalData)
```

```
##   Conserved Disturbed
## 1         1         3
## 2        18        48
## 3         0         1
## 4         7         0
## 5         2         1
## 6         1         7
```

The second dataset is about incidence (or presence-absence) information of 412 mite morphospecies found in the soil cones of 32 sampling plots (the moss carpets) of the costal areas of western Canada (Vancouver costal areas) (Chen 2013; Chen et al. 2015). As done with the dataset `HerpetologicalData`, for taking a quick view on a part of the data set, we can type the following command in R:

```
head(CanadaMite)
```

```
##      data1 data2
## [1,]    12     9
## [2,]    12    11
## [3,]     7     9
## [4,]     9     8
## [5,]     4     1
## [6,]     5     3
```

This dataset has two columns. The first column represents incidence counts of observed mite species in the first early-day 16 sampling plots, while the second column represents incidence counts of the observed species in the remaining late-day 16 sampling plots.

To use the datasets for demonstration, taking the Canadian mite data as an example, we can use all the three estimators to predict the number of new rare species that will be present only one, two or three times in the remaining 16 sampling plots (but not present in the first 16 plots; the first column). We then can use the true observed data (the second column) to check the predictive power of each estimator. To do so, type the following commands in R:

```
# load the empirical dataset
data(CanadaMite)
# two columns represent two samples of incidence counts
X.merge = CanadaMite
# the first column is treated as the original sample
X.col1 = X.merge[,1]
# the number of quadrats in the first sample
nT = 16
# the number of quadrats in the additional sample (i.e., the second col
umn)
u = 16
# do the predictions and conduct the bootstrapping procedures
print(Pred.incidence.rare(boot.rep=100, Q=NULL, xi=X.col1, nT=nT, u=u,
k.show = 3))

## $`Data information`
##   Original sample size (t) Additional sample size (u)
##                         16                         16
##
## $`Naive method`
##      k Estimate Estimated SE 95% lower limit 95% upper limit
## [1,] 1     20.8          1.2            18.5            23.1
## [2,] 2     11.6          0.6            10.5            12.7
## [3,] 3      4.5          0.2             4.2             4.9
##
## $`Bayesian-weight method`
##      k Estimate Estimated SE 95% lower limit 95% upper limit
## [1,] 1     72.4          8.0            56.8            88.0
## [2,] 2     19.4          0.9            17.7            21.2
## [3,] 3      4.8          0.4             4.0             5.5
##
## $`Unweighted method`
##      k Estimate Estimated SE 95% lower limit 95% upper limit
## [1,] 1     54.8          1.1            52.7            57.0
## [2,] 2     22.1          0.6            20.8            23.4
## [3,] 3      6.3          0.3             5.8             6.8
```

For the newly found singleton rare species (that is, species present only in exactly one plot over the 16 remaining plots), the Bayesian-weight estimator predicted the highest number (72 species), followed by Chao et al.'s unweighted estimator (55 species). By contrast, the naive estimator predicted only 21 species. Are these predictions accurate? To check the predictive power of these estimators, we can compare the estimated values with the true observed value by executing the following command:

```
length(which(CanadaMite[,2]==1 & CanadaMite[,1]==0))
```

```
## [1] 102
```

The above script is to check how many newly found rare species in the additional sample (i.e., the remaining 16 plots) have only one-time presence in the additional 16 sampling plots. As there are 102 newly found species present in only a single plot over the 16 plots, one can see that the 95% bootstrapping confidence intervals of all the estimators did not cover the observed value. However, one should acknowledge that the Bayesian-weight estimator predicted the closest value to the true observed value.

Reversely, if users have an interest to use species incidence information from the 16 late-day sampling plots to predict the number of newly found rare species in the first 16 early-day sampling plots, the results are similar but also interesting:

```
# load the empirical dataset
data(CanadaMite)
# two columns represent two samples of incidence counts
X.merge = CanadaMite
# the second column is treated as the original sample
X.col1 = X.merge[,2]
# the number of quadrats in the first sample
nT = 16
# the number of quadrats in the additional sample (i.e., the first colu
mn)
u = 16
# do the predictions and conduct the bootstrapping procedures
print(Pred.incidence.rare(boot.rep=100, Q=NULL, xi=X.col1, nT=nT, u=u,
k.show = 3))
```

```
## $`Data information`
##   Original sample size (t) Additional sample size (u)
##                         16                         16
##
## $`Naive method`
##      k Estimate Estimated SE 95% lower limit 95% upper limit
## [1,] 1     25.1          1.6            22.0            28.1
## [2,] 2     13.5          0.8            12.0            15.0
## [3,] 3      4.9          0.3             4.4             5.4
##
## $`Bayesian-weight method`
```

```
##      k Estimate Estimated SE 95% lower limit 95% upper limit
## [1,] 1    107.8         11.8            84.7           130.9
## [2,] 2     20.4          1.1            18.2            22.7
## [3,] 3      3.9          0.5             2.9             4.8
##
## $`Unweighted method`
##      k Estimate Estimated SE 95% lower limit 95% upper limit
## [1,] 1     97.6          2.2            93.3           101.9
## [2,] 2     26.1          1.3            23.7            28.6
## [3,] 3      5.2          0.4             4.5             5.9
```

Now check the observed numbers of rare species that are present one, two and three times in the first 16 sampling plots by respectively executing the following commands:

```
#newly found species that are present in only a single plot of the firs
t 16 plots
length(which(CanadaMite[,1]==1 & CanadaMite[,2]==0))

## [1] 78

#newly found species that are present in only two plots of the first 16
 plots
length(which(CanadaMite[,1]==2 & CanadaMite[,2]==0))

## [1] 21

#newly found species that are present in only three plots of the first
16 plots
length(which(CanadaMite[,1]==3 & CanadaMite[,2]==0))

## [1] 7
```

Given that the observed value of the newly found rare singleton species is 78, one can see that the Bayesian estimator has the most accurate predictive power, because it had a point estimate as 107.8 and the corresponding 95% bootstrap confidence interval (85.8, 129.9). The lower bound of this confidence interval was very close to the true value. The performance of Chao et al.'s unweighted estimator was also reasonably well; the 95% confidence interval was (93.3, 101.9). By contrast, the prediction of the naive unweighted estimator was not desirable, as the upper bound of the confidence interval was only 28.2, being too far away from the observed value!

Moreover, when comparing the predicted values of the newly found rare species that are present in exactly two plots, we can see that the Bayesian-weighted estimator was completely accurate: the point estimate was 20.4 with the 95% confidence interval as (18.1, 22.7), covering the true observed value 21 well.

Warming: The statistical methods we developed in *RSE* package are not risk-free when applied to surveyed data. They are developed based on the statistical assumptions that species' individuals are sampled randomly according to species' relative abundance (thus follow multinomial distribution) and the original sample data and predicting sample are assume to be homogeneous. If species' individuals are believed to derive from different species pools with heterogeneous environmental conditions, or if species' individuals are sampled dependently (e.g., biodiversity surveys are conducted using clustering or adaptive sampling method), the estimation of newly found rare species using our methods might be deviated from the true value to a great extent.

Note: Other than the vignette shown here, users are also encouraged to visit the Github repository (https://github.com/ecomol/RSE) to read more information about the package. They are also suggested to read the manual of the package for better understanding the meaning of a specific function or parameter.

References

Chao A, Hsieh T, Chazdon R, Colwell R, Gotelli N. 2015. Unveiling the species-rank abundance distribution by generalizing the Good-Turing sample coverage theory. Ecology 96:1189-1201.

Chen Y. 2013. Microarthropod diversity and distribution in Southwestern Canada. Master thesis, University of British Columbia, Vancouver, Canada.

Chen Y, Amundrud SL, Srivastava DS. 2015. Spatial variance in soil microarthropod communities: Niche, neutrality, or stochasticity? Ecoscience 21:1-14.

Chen Y, Zhang J, Jiang J, Nielsen S, He F. 2017. Assessing the effectiveness of China's protected areas to conserve current and future amphibian diversity. Diversity and Distributions 23:146-157.

Davies K, Margules C, Lawrence J. 2000. Which traits of species predict population declines in experimental forest fragments? Ecology 81:1450-1461.

Fattorini S. 2014. Relations between species rarity, vulnerability, and range contraction for a beetle group in a densely populated region in the Mediterranean biodiversity hotspot. Conservation Biology 28:169-176.

Gaston K. 2012. The importance of being rare. Nature 487:46-47.

Jain M et al. 2014. The importance of rare species: a trait-based assessment of rare species contributions to functional diversity and possible ecosystem function in tall-grass prairies. Ecology and Evolution 4:104-112.

Kunin W, Gaston K. 1993. The bioogy of rarity: patterns, causes and consequences. Trends in Ecology and Evolution 8:298-301.

Leitao R, Zuanon J, Villeger S, Williams S, Baraloto C, Fortunel C, Mendonca F, Mouillot D. 2016. Rare species contribute disproportionately to the functional structure of species assemblages. Proceedings of Royal Society B: Biological Sciences 283:20160084.

Lyons K, Brigham C, Traut B, Schwartz M. 2005. Rare species and ecosystem functioning. Conservation Biology 19:1019-1024.

Mouillot D et al. 2013. Rare species support vulnerable functions in high-diversity ecosystems. PLoS Biology 11:e1001569.

Suazo-Ortuno I, Alvarado-Diaz J, Martines-Ramos M. 2008. Effects of conversion of dry tropical forest to agricultural mosaic on herpetofaunal assemblages. Conservation Biology 22:362-374.

Prendergast J, Quinn R, Lawton J, Eversham B, Gibbons D. 1993. Rare species, the coincidence of diversity hotspots and conservation strategies. Nature 365:335-337.