World_insects_estimate

Load libraries and ../Data

```r
library(SpadeR)
library(dplyr)
library(stringr)
library(pairwiseCI)
#library(remotes)
#install_version("CARE1", "1.1.0")
library(CARE1)

## Load Data

#Microgastrinae

reared <- read.csv("../Data/Microgastrinae ACG Reared.csv") %>%
  select(BOLD.code, Frequency)

core <- read.csv("../Data/Microgastrinae ACG Core Malaise.csv") %>%
  select(BOLD.code, Frequency)

peri <- read.csv("../Data/Microgastrinae ACG Peripheral Malaise.csv") %>%
  select(BOLD.code, Frequency)

# All insecta core

insc <- read.csv("../Data/Insecta ACG Core Malaise.csv") %>%
  select(BOLD.code, Frequency)

## highlight estimator

estimator <- "    iChao1 (Chiu et al. 2014)"
```

```
## adding in the trees

#TA = number of tree species in ACG
TA <- 1500

#Tc = number of tree species in CR

TC <- 2500
#Tw = number of tree species on Earth

## estimated tree species

TW <- 73274
TW_CI <- c(73271, 73276)

## described number of tree species

TW_d <- 60065

### Number of Saturniid moths

SA <- 106
SW <- 4535
```

Approach 1: Calculate insect richness worldwide using the CARE estimator for Microgastrinae

```
### Total number of BINS

reared %>% nrow()
```

[1] 886

```
core %>% nrow()
```

[1] 388

```
peri %>% nrow()
```

[1] 576

```
###
```

Step 1: Combining Malaise and Reared samples of Microgastrinae and estimating the true richness of ACG Microgastrinae

```
full_micro_table <- reared %>%
  rename(reared = Frequency) %>%
  full_join(rename(core, core = Frequency)) %>%
  full_join(rename(peri, peri = Frequency)) %>%
  mutate(reared = ifelse(is.na(reared), 0, 1)) %>%
  mutate(core = ifelse(is.na(core), 0, 1)) %>%
  mutate(peri = ifelse(is.na(peri), 0, 1))
```

Joining with `by = join_by(BOLD.code)`
Joining with `by = join_by(BOLD.code)`

```
head(full_micro_table)
```

```
    BOLD.code reared core peri
1 BOLD:AAA1266      1    0    0
2 BOLD:AAA1634      1    1    1
3 BOLD:AAA2143      1    0    0
4 BOLD:AAA2511      1    0    0
5 BOLD:AAA2668      1    0    0
6 BOLD:AAA3229      1    0    1
```

```
nrow(full_micro_table)
```

[1] 1411

3

```
# Transform observed ../Data to CARE1 input format
micro_care <- as.record(full_micro_table[,-1])

micro_care
```

```
001 010 011 100 101 110 111
246 133 146 663 114  39  70
```

```
# Main step to obtain population size estimates
CARE1.print(micro_care)
```

Warning in ni/nhat: Recycling array of length 1 in vector-array arithmetic is deprecated.
  Use c() or as.vector() instead.

Warning in ni/nhat: Recycling array of length 1 in vector-array arithmetic is deprecated.
  Use c() or as.vector() instead.

Warning in ni/nhat: Recycling array of length 1 in vector-array arithmetic is deprecated.
  Use c() or as.vector() instead.

(1) NUMBER OF IDENTIFIED CASES IN EACH LIST:
 n1 n2 n3
886 388 576

(2) ESTIMATES BASED ON ANY PAIR OF SAMPLES:
     Petersen Chapman se cil ciu
pa12   3154   3136 236 2725 3655
pa13   2774   2765 149 2501 3087
pa23   1035   1033 37 970 1115

Note1: Refer to Seber(1982,pages 59 and 60) for Petersen estimator
       and Chapman estimators as well as s.e formula.
Note2: A log-transformation is used is used to obtain the confidence
       interval so that the lower limit is always greater than the
       number of ascertained. Refer to Chao(1987,Biometrics,43,783-791)
       for the construction of the confidence interval.

(3) SAMPLE COVERAGE APPROACH:
          M   D Chat est se cil ciu
```

```
Nhat-0 1411 1063.667 0.494 2153  67 2033 2297
Nhat   1411 1063.667 0.494 3480 575 2624 4942
Nhat-1 1411 1063.667 0.494 2402 101 2224 2621
```

Warning: The estimated sample coverage(overlapping information)is too
        low so that Nhat is unstable. Recommend the use of Nhat-0 or Nhat-1.
Parameter estimates:

```
         u1   u2   u3   r12    r13  r23
Nhat-0 0.41 0.18 0.27 -0.32 -0.22 1.08
Nhat   0.25 0.11 0.17  0.10  0.25 2.36
Nhat-1 0.37 0.16 0.24 -0.24 -0.13 1.32
```

Definitions for the sample coverage approach:
M: number of individuals ascertained in at least one list.
D: the average of the number of invididuals listed in the combination
   of any two lists omitting the other one.
C^: sample coverage estimate, see Equation (14) of Chao and Tsay(1998).
    est: population size estimate.
se: estimated standard error of the population size estimation based on
    bootstrap replications.
cil: 95% confidence interval lower limit(using a log-transformation).
ciu: 95% confidence interval upper limit(using a log-transformation).
Nhat: Population size estimate for sufficiently high sample coverage
      cases, see Equation (20) of Chao and Tsay (1998).
Nhat-1: One-step population size estimate for low sample coverage cases;
        see Equation (2.21) of Chao et al. (1996). This estimator is
        suggested for use when the estimated se of Nhat is relatively large.
u1,u2,u3: estimated mean probabilities depending on the estimate of N.
r12,r13,r23 etc.: estimated coefficient of covariation(CCV) depending on the
                  estimate of N.

```r
  # Print the CARE1 input format

  ## n1 = reared = 886
  ## n2 = core = 388
  ## n3 = peripheral = 576

  ## Using Nhat-1

  Nhat1 <- round(estN(micro_care, method = "LSC", se = TRUE, nboot = 200))
```

```r
micro_all_estimate <- Nhat1[1]
micro_all_CI <- c(Nhat1[3], Nhat1[4])
```

Step 2. Estimating the undersampling ratio for Core Microgastrinae.

```r
## proportion of core malaise of the ichao total

proportion_core <- nrow(core)/micro_all_estimate

proportion_core
```

```
[1] 0.1615321
```

```r
CI_proportion_core <- nrow(core)/micro_all_CI

CI_proportion_core
```

```
[1] 0.1746175 0.1479222
```

Step 3 Calculate total number of insects in ACG

```r
## Total number of insecta bins in ACG core malaise
n_insects_core <- nrow(insc)

n_insects_core
```

```
[1] 53945
```

```r
total_insec_acg <- n_insects_core*(1/proportion_core)

total_insec_acg
```

```
[1] 333958.5
```

```r
CI_total_insect_acg <- n_insects_core*(1/round(CI_proportion_core,3))

n_insects_core*(1/round(CI_proportion_core,3))
```

[1] 308257.1 364493.2

```r
n_insects_core*(1/CI_proportion_core)
```

[1] 308932.4 364684.9

```r
CI_total_insect_acg
```

[1] 308257.1 364493.2

Step 4. Estimating global insect richness – based on trees

```r
##### Total insects
print("Total world insects estimated tree richness")
```

[1] "Total world insects estimated tree richness"

```r
total_insect_world <- (TW*total_insec_acg)/TA
total_insect_world
```

[1] 16313649

```r
CI_total_insect_world <- (TW*CI_total_insect_acg)/TA

CI_total_insect_world
```

[1] 15058156 17805252

```r
print("Total world insects described trees")
```

[1] "Total world insects described trees"

```r
total_insect_world <- (TW_d*total_insec_acg)/TA
total_insect_world
```

[1] 13372811

```r
CI_total_insect_world <- (TW_d*CI_total_insect_acg)/TA

CI_total_insect_world
```

[1] 12343644 14595524

Step 4. Estimating global insect richness – based on saturniid moths

```r
##### Total insects

print("Total world insects")
```

[1] "Total world insects"

```r
total_insect_world <- (SW*total_insec_acg)/SA
total_insect_world
```

[1] 14287752

```r
CI_total_insect_world <- (SW*CI_total_insect_acg)/SA

CI_total_insect_world
```

[1] 13188171 15594121

Approach 2: Calculate insect richness worldwide using the Chao estimator for Microgastrinae and for ACG malaise insects

Step 1: Combining Malaise and Reared samples of Microgastrinae and estimating the true richness of ACG Microgastrinae

```
full_micro_table <- reared %>%
  rename(reared = Frequency) %>%
  full_join(rename(core, core = Frequency)) %>%
  full_join(rename(peri, peri = Frequency)) %>%
  mutate(reared = ifelse(is.na(reared), 0, 1)) %>%
  mutate(core = ifelse(is.na(core), 0, 1)) %>%
  mutate(peri = ifelse(is.na(peri), 0, 1))
```

```
Joining with `by = join_by(BOLD.code)`
Joining with `by = join_by(BOLD.code)`
```

```
head(full_micro_table)
```

```
    BOLD.code reared core peri
1 BOLD:AAA1266      1    0    0
2 BOLD:AAA1634      1    1    1
3 BOLD:AAA2143      1    0    0
4 BOLD:AAA2511      1    0    0
5 BOLD:AAA2668      1    0    0
6 BOLD:AAA3229      1    0    1
```

```
nrow(full_micro_table)
```

```
[1] 1411
```

```
# Transform observed ../Data to CARE1 input format
micro_care <- as.record(full_micro_table[,-1])
```

```
micro_care
```

```
001 010 011 100 101 110 111
246 133 146 663 114  39  70
```

9

```
# Main step to obtain population size estimates
CARE1.print(micro_care)
```

Warning in ni/nhat: Recycling array of length 1 in vector-array arithmetic is deprecated.
  Use c() or as.vector() instead.

Warning in ni/nhat: Recycling array of length 1 in vector-array arithmetic is deprecated.
  Use c() or as.vector() instead.

Warning in ni/nhat: Recycling array of length 1 in vector-array arithmetic is deprecated.
  Use c() or as.vector() instead.

(1) NUMBER OF IDENTIFIED CASES IN EACH LIST:
 n1 n2 n3
886 388 576

(2) ESTIMATES BASED ON ANY PAIR OF SAMPLES:
     Petersen Chapman se   cil ciu
pa12   3154   3136 236 2725 3655
pa13   2774   2765 149 2501 3087
pa23   1035   1033 37 970 1115

Note1: Refer to Seber(1982,pages 59 and 60) for Petersen estimator
       and Chapman estimators as well as s.e formula.
Note2: A log-transformation is used is used to obtain the confidence
       interval so that the lower limit is always greater than the
       number of ascertained. Refer to Chao(1987,Biometrics,43,783-791)
       for the construction of the confidence interval.

(3) SAMPLE COVERAGE APPROACH:
          M   D Chat est se cil ciu
Nhat-0 1411 1063.667 0.494 2153 68 2032 2299
Nhat   1411 1063.667 0.494 3480 506 2701 4730
Nhat-1 1411 1063.667 0.494 2402 93 2236 2603

Warning: The estimated sample coverage(overlapping information)is too
         low so that Nhat is unstable. Recommend the use of Nhat-0 or Nhat-1.
Parameter estimates:
         u1   u2   u3   r12   r13 r23
Nhat-0 0.41 0.18 0.27 -0.32 -0.22 1.08

```
Nhat    0.25 0.11 0.17  0.10  0.25 2.36
Nhat-1 0.37 0.16 0.24 -0.24 -0.13 1.32
```

Definitions for the sample coverage approach:
M: number of individuals ascertained in at least one list.
D: the average of the number of invididuals listed in the combination
   of any two lists omitting the other one.
C^: sample coverage estimate, see Equation (14) of Chao and Tsay(1998).
    est: population size estimate.
se: estimated standard error of the population size estimation based on
    bootstrap replications.
cil: 95% confidence interval lower limit(using a log-transformation).
ciu: 95% confidence interval upper limit(using a log-transformation).
Nhat: Population size estimate for sufficiently high sample coverage
      cases, see Equation (20) of Chao and Tsay (1998).
Nhat-1: One-step population size estimate for low sample coverage cases;
        see Equation (2.21) of Chao et al. (1996). This estimator is
        suggested for use when the estimated se of Nhat is relatively large.
u1,u2,u3: estimated mean probabilities depending on the estimate of N.
r12,r13,r23 etc.: estimated coefficient of covariation(CCV) depending on the
                  estimate of N.

```
  # Print the CARE1 input format

  ## n1 = reared = 886
  ## n2 = core = 388
  ## n3 = peripheral = 576


  ## Using Nhat-1

  Nhat1 <- round(estN(micro_care, method = "LSC", se = TRUE, nboot = 200))

  micro_all_estimate <- Nhat1[1]
  micro_all_CI <- c(Nhat1[3], Nhat1[4])
```

Step 2. Estimating the undersampling ratio for Core Microgastrinae.

```
## proportion of core malaise of the ichao total

proportion_core <- nrow(core)/micro_all_estimate

proportion_core
```

[1] 0.1615321

```
CI_proportion_core <- nrow(core)/micro_all_CI

CI_proportion_core
```

[1] 0.1739910 0.1486021

Step 3 Estimate using iChao1 total number of insects in ACG

```
## Total number of insecta bins in ACG core malaise -- ran the Chao estimator before becau

#insc_acg_estimate <- ChaoSpecies(insc$Frequency,"abundance",k=10,conf=0.95)

insc_acg_estimate <-readRDS("../Data/insecta_acg_chao.rds")

n_insects_core <- insc_acg_estimate$Species_table[estimator,]
n_insects_core
```

```
 Estimate      s.e.   95%Lower   95%Upper
84414.957    357.209 83722.847 85123.154
```

```
## Calculate total number of insects in ACG
total_insec_acg <- MOVERR(theta1 = n_insects_core['Estimate'], ci1 = c(n_insects_core['95%

total_insec_acg$estimate
```

```
Estimate
522589.5
```

```
total_insec_acg$conf.int
```

```
lower.Estimate upper.Estimate
      483639.7       566503.3
```

Step 4. Estimating global insect richness – based on trees

```
##### Total insects

print("Total insects estimated tree richness ")
```

```
[1] "Total insects estimated tree richness "
```

```
total_insect_world <- (TW*total_insec_acg$estimate)/TA
total_insect_world
```

```
Estimate
25528149
```

```
CI_total_insect_world <- (TW*total_insec_acg$conf.int)/TA

CI_total_insect_world
```

```
lower.Estimate upper.Estimate
      23625478       27673310
```

```
print("Total insects described tree richness")
```

```
[1] "Total insects described tree richness"
```

```
total_insect_world <- (TW_d*total_insec_acg$estimate)/TA
total_insect_world
```

```
Estimate
20926226
```

```
CI_total_insect_world <- (TW_d*total_insec_acg$conf.int)/TA

CI_total_insect_world
```

```
lower.Estimate upper.Estimate
      19366546       22684682
```

Step 4. Estimating global insect richness – based saturniidae moths

```
##### Total insects

total_insect_world <- (SW*total_insec_acg$estimate)/SA
total_insect_world
```

```
Estimate
22357957
```

```
CI_total_insect_world <- (SW*total_insec_acg$conf.int)/SA

CI_total_insect_world
```

```
lower.Estimate upper.Estimate
      20691567       24236723
```

Additional Information

For Pooled Microgastrinae CV and Sample Coverage: 1-[number singletons]/[number specimens].

```
### Combine all microgastrinae

micro_all <- rbind(reared, core, peri) %>%
  group_by(BOLD.code) %>%
  summarise(new_freq = sum(Frequency)) %>%
  select(BOLD.code, new_freq)

## number of pooled micros
```

```r
nrow(micro_all)
```

```
[1] 1411
```

```r
## use chao on microgastrinae

micro_all_estimate <- ChaoSpecies(micro_all$new_freq,"abundance",k=10,conf=0.95)

micro_all_estimate$Species_table[estimator,]
```

```
Estimate     s.e. 95%Lower 95%Upper
1911.739   35.582 1846.715 1986.466
```

```r
## from spade R
micro_all_estimate$Basic_data_information[c("    CV for entire ../Dataset",
                                            "    Coverage estimate for entire ../Dataset")
```

```
     Variable Value
NA       <NA>  <NA>
NA.1     <NA>  <NA>
```

```r
## based on coverage spreadsheet

total_bins <- nrow(micro_all)
total_bins
```

```
[1] 1411
```

```r
total_singletons_micro_all <- sum(micro_all$new_freq ==1)
total_singletons_micro_all
```

```
[1] 406
```

```r
total_doubletons_micro_all <- sum(micro_all$new_freq ==2)
total_doubletons_micro_all
```

```
[1] 202
```

```
# total samples
sum(micro_all$new_freq)
```

[1] 21708

```
coverage_micro_all <- 1-(total_singletons_micro_all/sum(micro_all$new_freq))
coverage_micro_all
```

[1] 0.9812972

```
CV_micro_all <- sd(micro_all$new_freq)/mean(micro_all$new_freq)
CV_micro_all
```

[1] 2.990786

Sample Coverage For Core Malaise Microgastrinae.

```
## from Spade R
micro_core_estimate <- ChaoSpecies(core$Frequency,"abundance",k=10,conf=0.95)

micro_core_estimate$Basic_data_information[c("    CV for entire ../Dataset",
                                             "    Coverage estimate for entire ../Dataset")
```

```
     Variable Value
NA        <NA>  <NA>
NA.1      <NA>  <NA>
```

```
## calculating from the spreadsheet

total_bins <- nrow(core)
total_bins
```

[1] 388

```r
total_singletons_core <- sum(core$Frequency ==1)
total_singletons_core
```

[1] 176

```r
total_doubletons_core <- sum(core$Frequency ==2)
total_doubletons_core
```

[1] 54

```r
# total samples
sum(core$Frequency)
```

[1] 3781

```r
coverage_core <- 1-(total_singletons_core/sum(core$Frequency))
coverage_core
```

[1] 0.9534515

```r
CV_core <- sd(core$Frequency)/mean(core$Frequency)
CV_core
```

[1] 3.322479

Sample Coverage For Core Insecta

```r
## from spade R

insc_acg_estimate$Basic_data_information[c("    CV for entire ../Dataset",
                                           "    Coverage estimate for entire ../Dataset")
```

```
      Variable Value
NA        <NA>  <NA>
NA.1      <NA>  <NA>
```

```
insecta <- read.csv("../Data/Insecta ACG Core Malaise.csv")

## from the spreadsheet

total_bins <- nrow(insecta)
total_bins
```

[1] 53945

```
total_singletons_insecta <- sum(insecta$Frequency ==1)
total_singletons_insecta
```

[1] 19374

```
total_doubletons_insecta <- sum(insecta$Frequency ==2)
total_doubletons_insecta
```

[1] 7402

```
coverage_insecta <- 1-(total_singletons_insecta/sum(insecta$Frequency))
coverage_insecta
```

[1] 0.9881422

```
CV_insecta <- sd(insecta$Frequency)/mean(insecta$Frequency)
CV_insecta
```

[1] 8.956866

Number of singletons and sample coverage for Core Malaise for each taxonomic level (Braconidae, Ichneumonoidea, and Hymenoptera).

Braconidae

```
braconidae <- read.csv("../Data/Braconidae ACG Core Malaise.csv")

## from the spreadsheet

total_bins <- nrow(braconidae)
total_bins
```

[1] 2160

```
total_singletons_braconidae <- sum(braconidae$Frequency ==1)
total_singletons_braconidae
```

[1] 1022

```
total_doubletons_braconidae <- sum(braconidae$Frequency ==2)
total_doubletons_braconidae
```

[1] 342

```
## total number of samples

sum(braconidae$Frequency)
```

[1] 15108

```
coverage_braconidae <- 1-(total_singletons_braconidae/sum(braconidae$Frequency))
coverage_braconidae
```

[1] 0.9323537

```
CV_braconidae <- sd(braconidae$Frequency)/mean(braconidae$Frequency)
CV_braconidae
```

[1] 3.270035

Ichneumonoidea

```r
ichneumonoidea <- read.csv("../Data/Ichneumonoidea ACG Core Malaise.csv")

## from the spreadsheet

total_bins <- nrow(ichneumonoidea)
total_bins
```

[1] 3402

```r
total_singletons_ichneumonoidea <- sum(ichneumonoidea$Frequency ==1)
total_singletons_ichneumonoidea
```

[1] 1496

```r
total_doubletons_ichneumonoidea <- sum(ichneumonoidea$Frequency ==2)
total_doubletons_ichneumonoidea
```

[1] 530

```r
## total number of samples

sum(ichneumonoidea$Frequency)
```

[1] 33207

```r
coverage_ichneumonoidea <- 1-(total_singletons_ichneumonoidea/sum(ichneumonoidea$Frequency
coverage_ichneumonoidea
```

[1] 0.9549493

```r
CV_ichneumonoidea <- sd(ichneumonoidea$Frequency)/mean(ichneumonoidea$Frequency)
CV_ichneumonoidea
```

[1] 3.862408

Hymenoptera

```r
hymenoptera <- read.csv("../Data/Hymenoptera ACG Core Malaise.csv")

## from the spreadsheet

total_bins <- nrow(hymenoptera)
total_bins
```

```
[1] 13934
```

```r
total_singletons_hymenoptera <- sum(hymenoptera$Frequency ==1)
total_singletons_hymenoptera
```

```
[1] 5689
```

```r
total_doubletons_hymenoptera <- sum(hymenoptera$Frequency ==2)
total_doubletons_hymenoptera
```

```
[1] 2061
```

```r
## total number of samples

sum(hymenoptera$Frequency)
```

```
[1] 235248
```

```r
coverage_hymenoptera <- 1-(total_singletons_hymenoptera/sum(hymenoptera$Frequency))
coverage_hymenoptera
```

```
[1] 0.975817
```

```r
CV_hymenoptera <- sd(hymenoptera$Frequency)/mean(hymenoptera$Frequency)
CV_hymenoptera
```

```
[1] 7.55354
```