Assessing performance and developing species richness estimators

Claire Vincent
Étudiante à l'Université Blaise Pascal Clermont-Ferrand, 41 rue Edmond Rostand 03100 Montluçon. clairevincent0310@gmail.com

In the field of biology, especially ecology, accurate assessing of the biodiversity is a recurring theme, which helps, among others, to improve knowledge about the various existing communities and the development of conservation plans. Moreover, the emergence of multidisciplinary sciences, such as bioinformatics or biostatistics, opens a new path for analyzing data generated, unthinkable until a few years ago.
Perform fauna or flora inventory may very quickly become complicated when it comes to sample a wide geographical area (region, country, earth) or an overrepresented taxon (Arthropods). Each ecosystem also has its own characteristics, which leads to having to accept a set of heterogeneous variables and different from one site to another.

To do this, many tools exist to assess the quality of the sampling. In the context of the use of nonparametric estimators, the most common are the bootstrap (Heck et al. 1975), the jackknife (Burham and Overton 1978, 1979) and the Chao (Chao, 1984). The latter two are used in the package Biodiversity Assessment Tools (BAT) (Cardoso et al. 2015) R. This algorithm realizes the alpha and beta diversity of sampling and through the creation of curves accumulation embodies the potential bias between the sampled species richness and species richness of the study area.

The objective of the study is to propose a classification of these different estimators according to their precision, accuracy and freedom from bias (Walther and Moore 2005). Classifications have been proposed (Poulin 1998 Hortal et al 2006. Chao et al., 2009), however none of them was made on large data sets (only two or three samples). The originality lies in the fact of using a relatively consistent data set (at least 100, grouping of regions and taxa of varying sizes).

Following the establishment of this classification, two applications seem possible. The first would be to evaluate the behavior estimator each area, the taxon, their size in order to establish a protocol for using each estimator according to the variables studied. The second would be to consider the implementation of a new tool that combines the strengths of existing estimators and would be applied globally.
Ultimately, the implementation of such a tool would allow both the development of conservation plans for the most endangered species, as well as eve of the most common species.

Assessing performance and developing species richness estimators

Claire VINCENT, Pedro CARDOSO

Estimer la biodiversité, pourquoi ?

→ Tâche complexe
(Smith & Belle, 1984), *(Walther & Moore, 2005)*, *(Lopez, De Aguiar Fracasso, Mesquita, Palma, & Riul, 2012)*

→ Applications en écologie, biogéographie et biologie de la conservation

Estimer la biodiversité, comment ?

→ Estimation de l'alpha-diversité

- Estimateurs de richesse spécifique
- Biodiversity Assessment Tools (package R)

Estimateurs

Jackknife (Burnham et Overton, 1978, 1979)		**Chao** (1984, 1987)	
Probabilité de capture variable d'un individu à l'autre		Espèces rares sont les plus informatives	
Abondance	Incidence	Abondance	Incidence
Jack1ab Jack2ab	Jack1in Jack2in	Chao1	Chao2

Jackknife

Jack1ab : $S_{JK1ab} = S_{obs} + s_1$

Jack1in : $S_{JK1in} = S_{obs} + q_1 \times \frac{q-1}{q}$

Jack2ab : $S_{JK2ab} = S_{obs} + 2s_1 - s_2$

Jack2in : $S_{JK2in} = S_{obs} + (q_1 \times \frac{2q-3}{q} - q_2 \times \frac{(q-2)^2}{q(q-1)})$ si $q > 1$

$S_{JK2in} = S_{obs} + 2q_1 - q_2$ pour $q = 1$

S_{JK} : richesse spécifique estimée
s_1 : nombre de singletons
q : nombre d'échantillons
q_2 : nombre d'échantillons doubles

S_{obs} : richesse spécifique observée
s_2 : nombre de doubletons
q_1 : nombre d'échantillons uniques

Chao

Chao1 : $S_{Chao1} = S_{obs} + \frac{{s_1}^2}{2\,s_2}$ si $s_2 \neq 0$

$S_{Chao1} = S_{obs} + \frac{s_1(s_1-1)}{2\,(s_2+1)}$ pour $s_2 = 0$

Chao2 : $S_{Chao2} = S_{obs} + \frac{{q_1}^2}{2\,q_2}$ si $q_2 \neq 0$

$S_{Chao2} = S_{obs} + \left(\frac{(m-1)}{m}\right)\left(q_1 \frac{(q_1-1)}{2(q_2+1)}\right)$ pour $q_2 = 0$

S_{Chao1} , S_{Chao2} : richesse spécifique estimée
s_1 : nombre de singletons
q_1 : nombre d'échantillons uniques
m : nombre total d'échantillons ($m \neq 0$)

S_{obs} : richesse spécifique observée
s_2 : nombre de doubletons
q_2 : nombre d'échantillons doubles

Correction des estimateurs

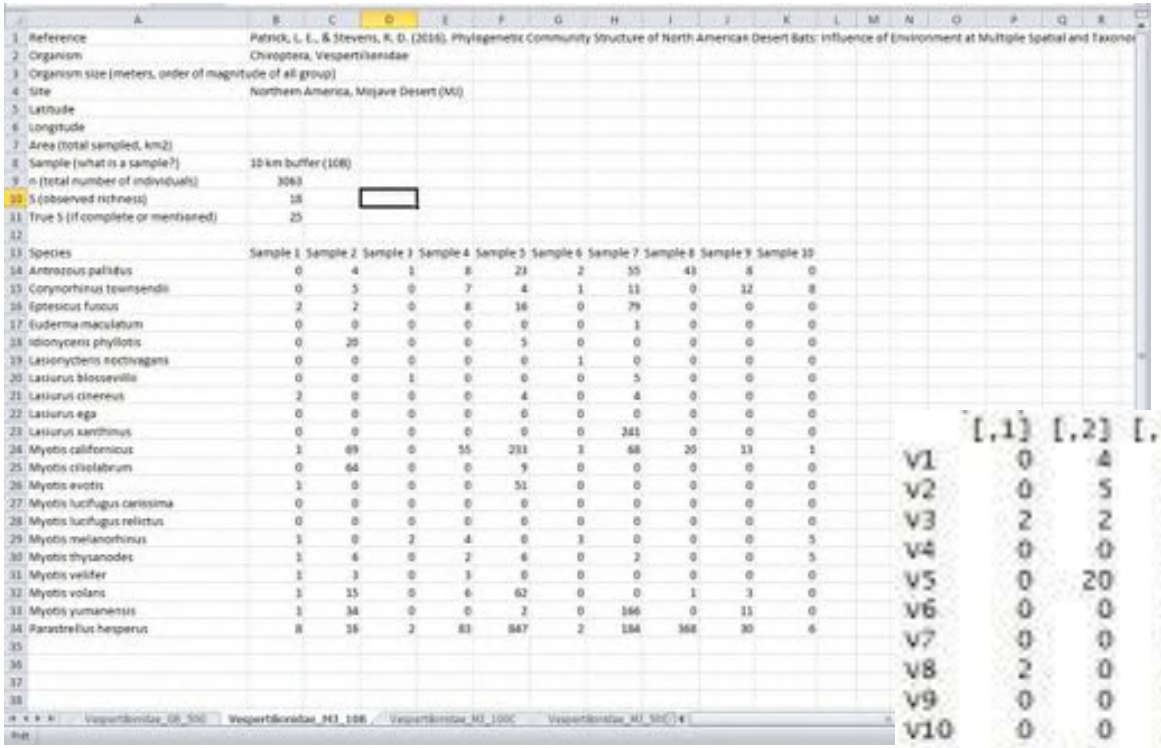
P-correction (Lopez et. al, 2012) :

→ prend en compte la proportion des singletons au sein d'un échantillonnage de taille réduite

$$S_{estP} = S_{est} \times (1 + P^2)$$

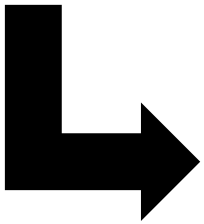
S_{estP} : estimation corrigée
S_{est} : S_{JK} ou S_{Chao}
P : proportion de singletons

Données brutes



	A	B	C	D	E	F	G	H	I	J	K
1	Reference	Patrick, L. E., & Stevens, R. D. (2016). Phylogenetic Community Structure of North American Desert Bats: Influence of Environment at Multiple Spatial and Taxonor									
2	Organism	Chiroptera, Vespertilionidae									
3	Organism size (meters, order of magnitude of all group)										
4	Site	Northern America, Mojave Desert (MJ)									
5	Latitude										
6	Longitude										
7	Area (total sampled, km2)										
8	Sample (what is a sample?)	10 km buffer (10B)									
9	n (total number of individuals)	3063									
10	S (observed richness)	18									
11	True S (if complete or mentioned)	25									
12											
13	Species	Sample 1	Sample 2	Sample 3	Sample 4	Sample 5	Sample 6	Sample 7	Sample 8	Sample 9	Sample 10
14	Antrozous pallidus	0	4	1	8	23	2	55	43	8	0
15	Corynorhinus townsendii	0	5	0	7	4	1	11	0	12	8
16	Eptesicus fuscus	2	2	0	8	16	0	79	0	0	0
17	Euderma maculatum	0	0	0	0	0	0	1	0	0	0
18	Idionycteris phyllotis	0	20	0	0	5	0	0	0	0	0
19	Lasionycteris noctivagans	0	0	0	0	0	1	0	0	0	0
20	Lasiurus blossevillii	0	0	1	0	0	0	5	0	0	0
21	Lasiurus cinereus	2	0	0	0	4	0	4	0	0	0
22	Lasiurus ega	0	0	0	0	0	0	0	0	0	0
23	Lasiurus xanthinus	0	0	0	0	0	0	241	0	0	0
24	Myotis californicus	1	69	0	55	233	3	68	20	13	1
25	Myotis ciliolabrum	0	64	0	0	9	0	0	0	0	0
26	Myotis evotis	1	0	0	0	51	0	0	0	0	0
27	Myotis lucifugus carissima	0	0	0	0	0	0	0	0	0	0
28	Myotis lucifugus relictus	0	0	0	0	0	0	0	0	0	0
29	Myotis melanorhinus	1	0	2	4	0	3	0	0	0	5
30	Myotis thysanodes	1	6	0	2	6	0	2	0	0	5
31	Myotis velifer	1	3	0	3	0	0	0	0	0	0
32	Myotis volans	1	15	0	6	62	0	0	1	3	0
33	Myotis yumanensis	1	34	0	0	2	0	166	0	11	0
34	Parastrellus hesperus	8	16	2	83	847	2	184	368	30	6
35											
36											
37											
38											

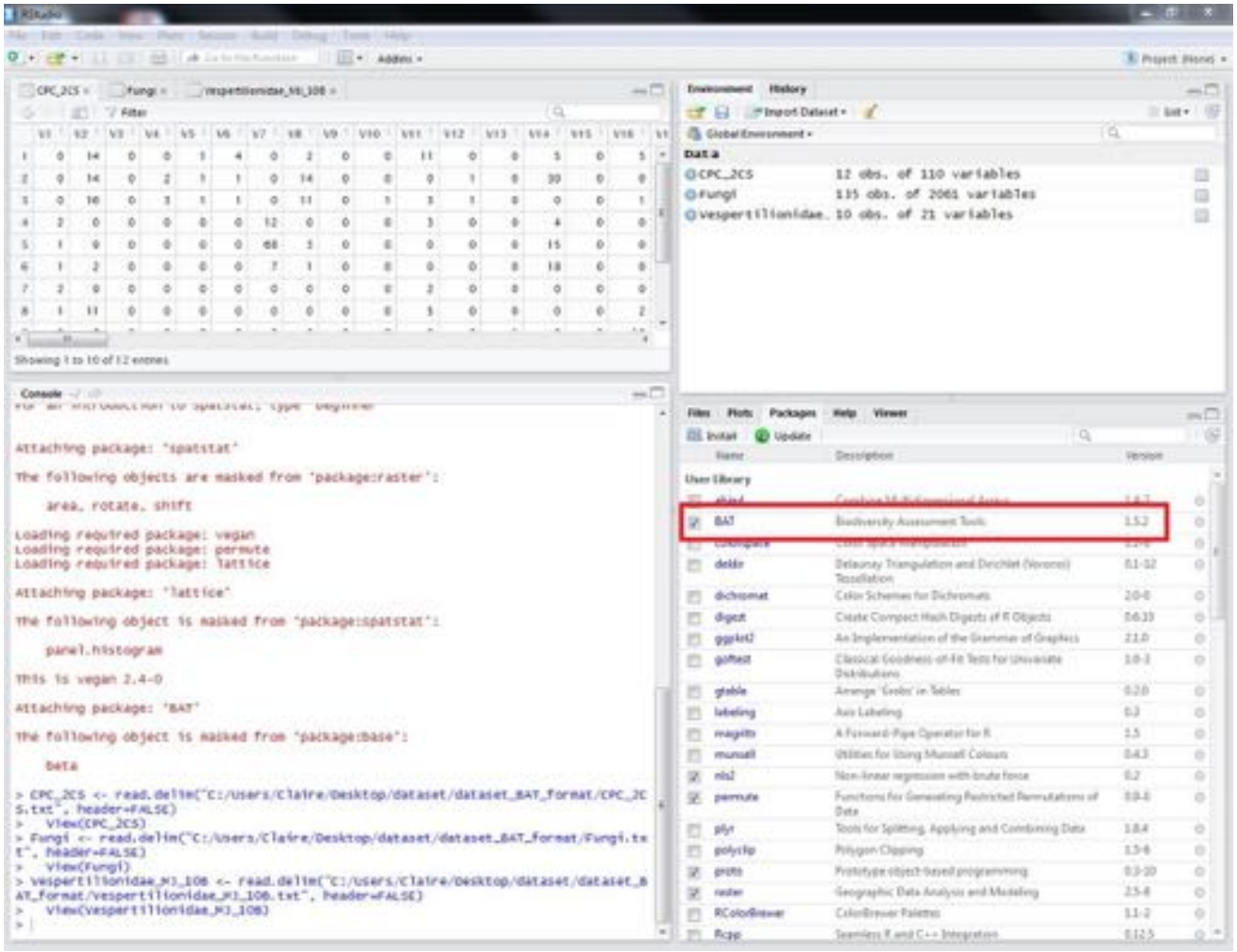
fichier Excel



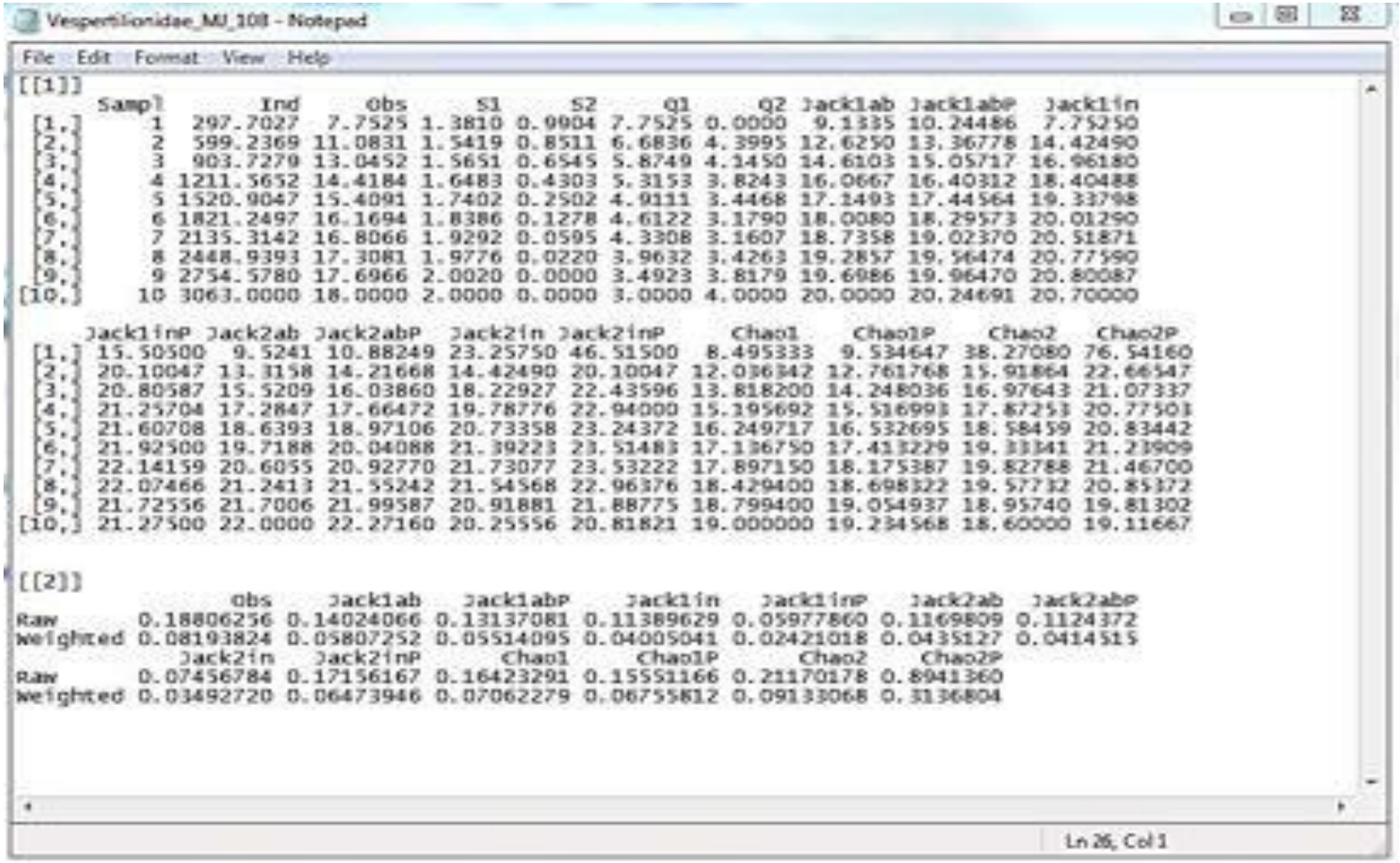
	[,1]	[,2]	[,3]	[,4]	[,5]	[,6]	[,7]	[,8]	[,9]	[,10]
V1	0	4	1	8	23	2	55	43	8	0
V2	0	5	0	7	4	1	11	0	12	8
V3	2	2	0	8	16	0	79	0	0	0
V4	0	0	0	0	0	0	1	0	0	0
V5	0	20	0	0	5	0	0	0	0	0
V6	0	0	0	0	0	1	0	0	0	0
V7	0	0	1	0	0	0	5	0	0	0
V8	2	0	0	0	4	0	4	0	0	0
V9	0	0	0	0	0	0	0	0	0	0
V10	0	0	0	0	0	0	241	0	0	0
V11	1	69	0	55	233	3	68	20	13	1
V12	0	64	0	0	9	0	0	0	0	0
V13	1	0	0	0	51	0	0	0	0	0
V14	0	0	0	0	0	0	0	0	0	0
V15	0	0	0	0	0	0	0	0	0	0
V16	1	0	2	4	0	3	0	0	0	5
V17	1	6	0	2	6	0	2	0	0	5
V18	1	3	0	3	0	0	0	0	0	0
V19	1	15	0	6	62	0	0	1	3	0
V20	1	34	0	0	2	0	166	0	11	0
V21	8	16	2	83	847	2	184	368	30	6

matrice espèce × échantillon

R version 3.3.1 « Bug in Your Hair »



Résultats bruts



Vespertilionidae_MU_108 - Notepad

File Edit Format View Help

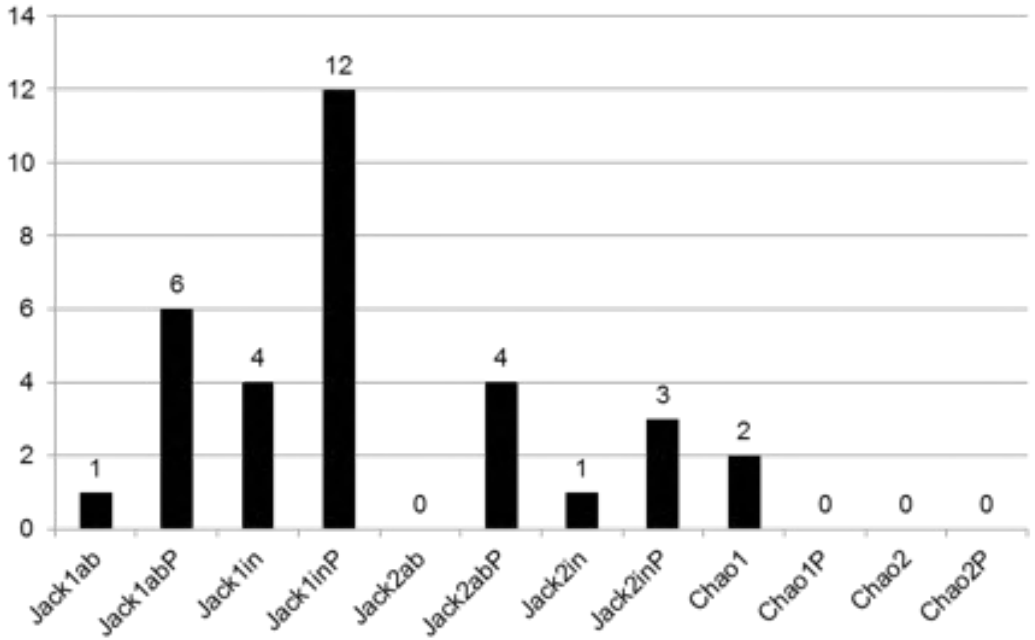
```
[[1]]
      Sampl       Ind      Obs      S1      S2      Q1      Q2 Jack1ab Jack1abP  Jack1in
 [1,]     1  297.7027   7.7525 1.3810 0.9904 7.7525 0.0000  9.1335 10.24486  7.75250
 [2,]     2  599.2369 11.0831 1.5419 0.8511 6.6836 4.3995 12.6250 13.36778 14.42490
 [3,]     3  903.7279 13.0452 1.5651 0.6545 5.8749 4.1450 14.6103 15.05717 16.96180
 [4,]     4 1211.5652 14.4184 1.6483 0.4303 5.3153 3.8243 16.0667 16.40312 18.40488
 [5,]     5 1520.9047 15.4091 1.7402 0.2502 4.9111 3.4468 17.1493 17.44564 19.33798
 [6,]     6 1821.2497 16.1694 1.8386 0.1278 4.6122 3.1790 18.0080 18.29573 20.01290
 [7,]     7 2135.3142 16.8066 1.9292 0.0595 4.3308 3.1607 18.7358 19.02370 20.51871
 [8,]     8 2448.9393 17.3081 1.9776 0.0220 3.9632 3.4263 19.2857 19.56474 20.77590
 [9,]     9 2754.5780 17.6966 2.0020 0.0000 3.4923 3.8179 19.6986 19.96470 20.80087
[10,]    10 3063.0000 18.0000 2.0000 0.0000 3.0000 4.0000 20.0000 20.24691 20.70000

      Jack1inP Jack2ab Jack2abP  Jack2in Jack2inP     Chao1    Chao1P    Chao2   Chao2P
 [1,] 15.50500  9.5241 10.88249 23.25750 46.51500  8.495333  9.534647 38.27080 76.54160
 [2,] 20.10047 13.3158 14.21668 14.42490 20.10047 12.036342 12.761768 15.91864 22.66547
 [3,] 20.80587 15.5209 16.03860 18.22927 22.43596 13.818200 14.248036 16.97643 21.07337
 [4,] 21.25704 17.2847 17.66472 19.78776 22.94000 15.195692 15.516993 17.87253 20.77503
 [5,] 21.60708 18.6393 18.97106 20.73358 23.24372 16.249717 16.532695 18.58459 20.83442
 [6,] 21.92500 19.7188 20.04088 21.39223 23.51483 17.136750 17.413229 19.33341 21.23909
 [7,] 22.14159 20.6055 20.92770 21.73077 23.53222 17.897150 18.175387 19.82788 21.46700
 [8,] 22.07466 21.2413 21.55242 21.54568 22.96376 18.429400 18.698322 19.57732 20.85372
 [9,] 21.72556 21.7006 21.99587 20.91881 21.88775 18.799400 19.054937 18.95740 19.81302
[10,] 21.27500 22.0000 22.27160 20.25556 20.81821 19.000000 19.234568 18.60000 19.11667


[[2]]
              Obs    Jack1ab   Jack1abP    Jack1in   Jack1inP   Jack2ab  Jack2abP
Raw      0.18806256 0.14024066 0.13137081 0.11389629 0.05977860 0.1169809 0.1124372
Weighted 0.08193824 0.05807252 0.05514095 0.04005041 0.02421018 0.0435127 0.0414515
            Jack2in   Jack2inP      Chao1     Chao1P      Chao2     Chao2P
Raw      0.07456784 0.17156167 0.16423291 0.15551166 0.21170178 0.8941360
Weighted 0.03492720 0.06473946 0.07062279 0.06755812 0.09133068 0.3136804
```

Ln 26, Col 1

Weighted accuracy

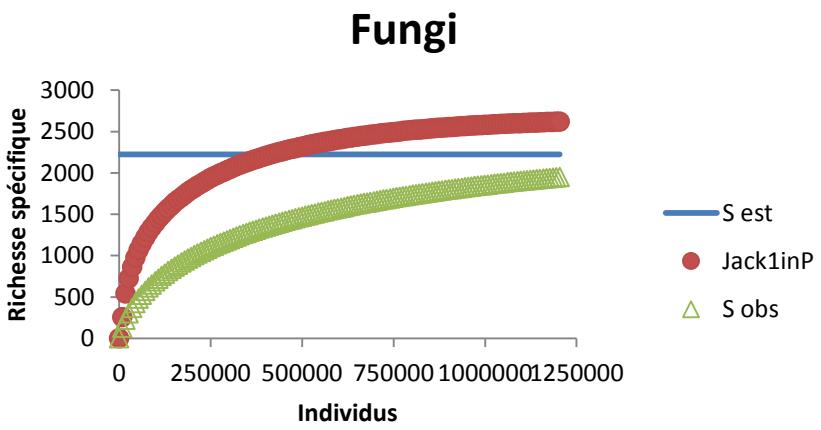
→ Accuracy *(Walter et al., 2005)* : différence entre S_{est} et S_{obs}

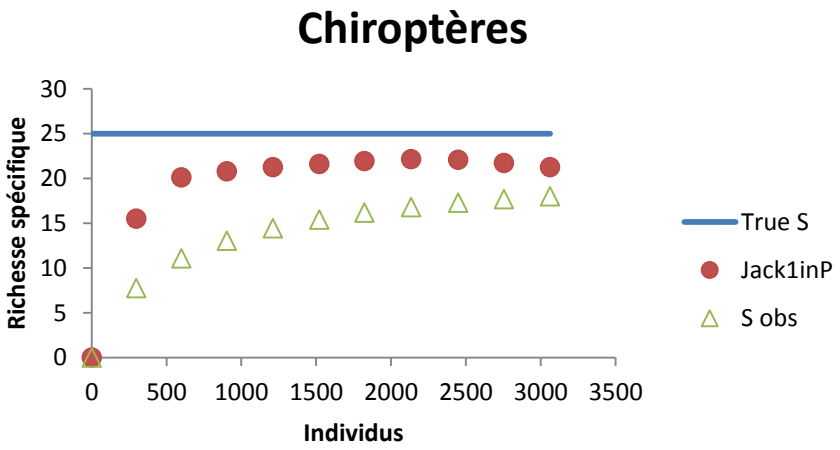
→ Weighted accuracy : pondère les résultats finaux

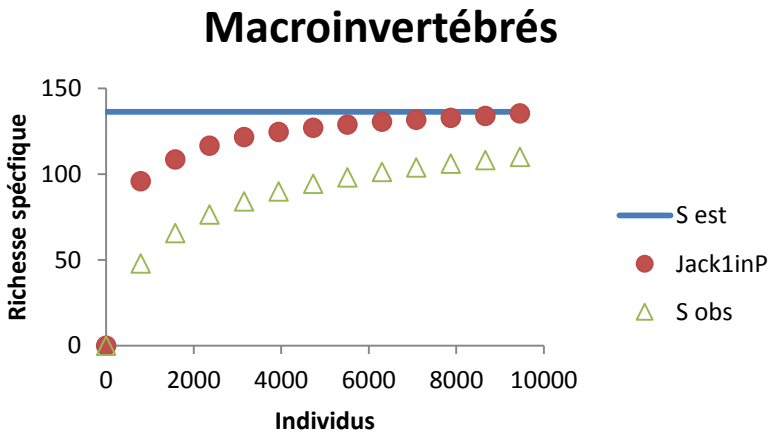


Performance des estimateurs selon leur weighted accuracy

BAT – Courbes d'accumulation







Conclusion

- Aucun estimateur ne fonctionne mieux que tous les autres de manière générale. L'impact des caractéristiques inhérentes à chaque jeu de données reste en cours d'étude.

- En considérant le meilleur estimateur pour chaque jeu de données, un haut taux de complétude de l'inventaire est nécessaire (> 80%).

- Nécessité de nouveaux outils moins biaisés et applicables de manière universelle.

Merci de votre attention

Claire Vincent[1], Pedro Cardoso[2]

[1] Blaise Pascal University of Clermont-Ferrand, 63000 Clermont-Ferrand, France, Claire.Vincent1@etudiant.univ-bpclermont.fr

[2] Finnish Museum of Natural History, University of Helsinki, 00014 Helsinki, Finland, pedro.cardoso@helsinki.fi