Inferring macro-ecological patterns from local presence/absence data

Quantitative Life Science exam

Sartore Marika

5th April 2022

Università degli Studi di Padova

Outline

1. Introduction
2. Model
3. In-silico forests
4. Real dataset application
5. Conclusion
6. Appendix

Outline

1 Introduction

2 Model

3 In-silico forests

4 Real dataset application

5 Conclusion

6 Appendix

Introduction

Problem: Biodiversity changes across spatial scales.
Aim: Translate local information on biodiversity into global ones.
Literature: Many proposed methods use species abundance data.
Data: Most datasets are composed of presence/absence data.
Solution: The method presented in [1].

Outline

1 Introduction

2 Model

3 In-silico forests

4 Real dataset application

5 Conclusion

6 Appendix

Model

The method proposed in [1]:

- is based on the **form-invariance** property of the Negative Binomial distribution;
- uses only **presence/absence data**;
- can infer **species richness** at larger scale and other relevant biodiversity patterns (SAC,RSA,RSO);
- has been developed for forest but it **can be generalized** to other ecological systems.

Model - Theory

From global to local:
Relative Species Abundance at scale 1:

$$P(n|1) \propto NB(n|r,\xi) = \binom{n+r-1}{n} \xi^n (1-\xi)^r$$

+ well mixed hypothesis:

$$P_{binom}(k|n,p) = \binom{n}{k} p^k (1-p)^{n-k}, \quad k = 0,..,n$$

→ RSA at scale p is a negative binomial [2]:

$$P(k|p) \propto NB(k|r,\xi_p), \quad \text{with} \quad \xi_p = \frac{p\xi}{1-\xi(1-p)}$$

Model - Theory

From local to global:

If we have information about a portion p^* of the whole forest, we can infer these quantities for the entire forest:

- RSA:

$$NB(n|r,\xi) \quad \text{with} \quad \xi = \frac{\xi_{p^*}}{p^* + \xi_{p^*}(1-p^*)}$$

- nr. of species:

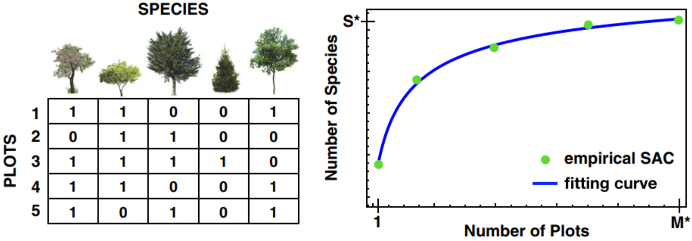
$$S = S^* \frac{1-(1-\xi)^r}{1-(1-\xi_{p^*})^r}$$

- RSO:

$$Q(v|M,1) = \sum_{n=v}^{\infty} Q_{occ}(v|n,M,1)P(n|1) \propto v^{r-1}$$

Model - Implementation

Data at the local scale p^*:

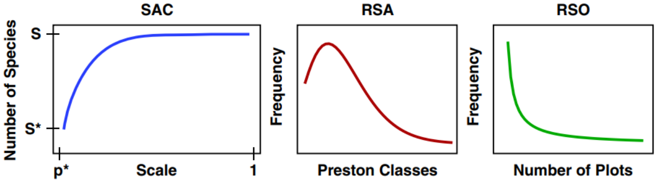


PLOTS	SPECIES				
1	1	1	0	0	1
2	0	1	1	0	0
3	1	1	1	1	0
4	1	1	0	0	1
5	1	0	1	0	1

(a) Presence/absence of S^* species in M^* plots, no abundances information. (b) For each subset of plots from 1 to M^* cells, it is computed the nr. of species observed in that portion of cells. The procedure is repeated 100 times and the green dots are the empirical averages. The fitting is done through $SAC_{th} = S\frac{1-(1-\xi_p)^r}{1-(1-\xi)^r}$.

Model - Implementation

Prediction at the global scale (p=1):



Using best-fit parameters found and the upscaling equations it is possible to predict the species richness S of the whole forest, the SAC, the RSA and the RSO.

Outline

1 Introduction

2 Model

3 In-silico forests

4 Real dataset application

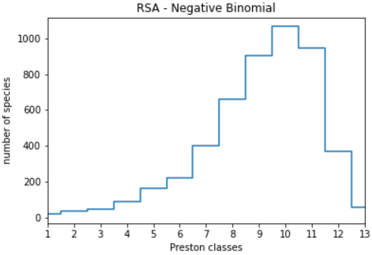
5 Conclusion

6 Appendix

In-silico forests

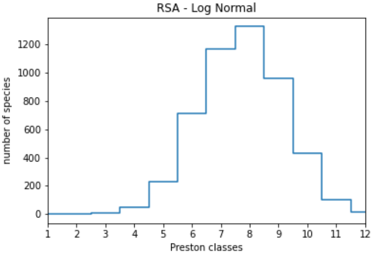
Generate in-silico forests according to two RSA distributions:

Negative Binomial:



$S = 4981$, $r = 0.8$, $\xi = 0.999$

Log Normal:



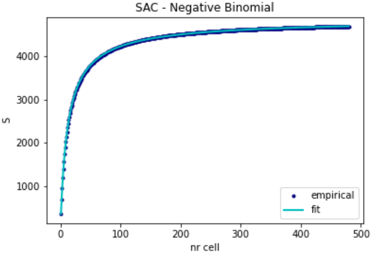
$S = 5000$, $\mu = 5$, $\sigma = 1$

Individuals are randomly distributed over 98*x*98 cells.

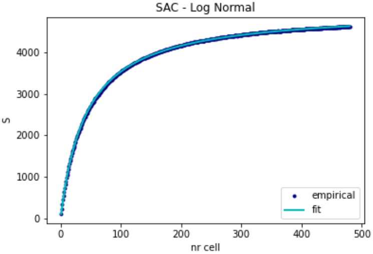
In-silico forests - results

Sample the $p^* = 0.05$ of the forest and infer species richness S at the whole scale:

Negative Binomial:



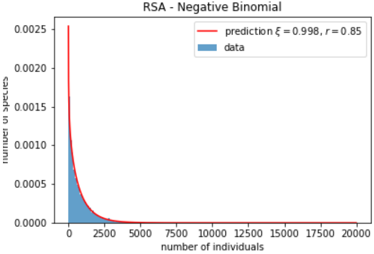
Log Normal:



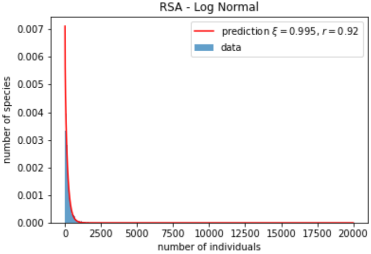
	S_{true}	S_{pred}	avg perc diff
NB	4981	4929 ±1	-1.04 ±0.02
LN	5000	5204 ±6	4 ±0.1

In-silico forests - results

Negative Binomial:



Log Normal:



Outline

1 Introduction

2 Model

3 In-silico forests

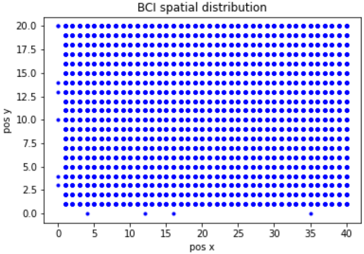
4 Real dataset application

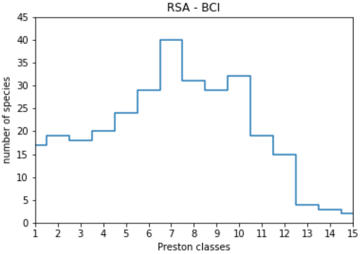
5 Conclusion

6 Appendix

Barro Colorado Island forest

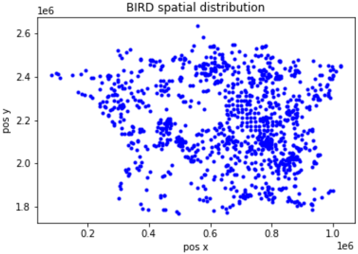
- Real dataset from the BCI forest
- Tot species = 302
- Tot cells = 800

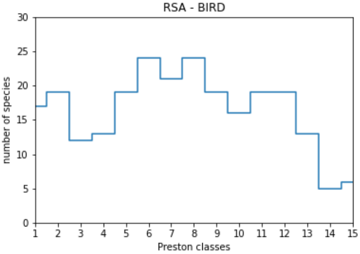




French birds species

- Real dataset of French bird species
- Tot species = 246
- Tot cells = 1067





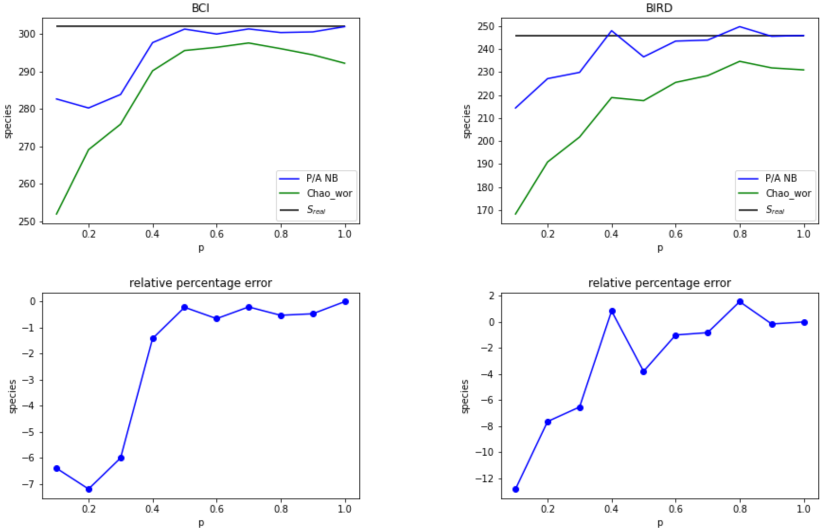
Model testing

- Sample each fraction $p^* = 0.1, .., 0.9$ of the P/A matrix and infer the nr. of species present at global scale.
- Relative percentage error $= \frac{S_{pred} - S_{true}}{S_{true}} 100$.
- Compare with CHAO estimator [3]:

$$S_{chao} = S^* + \frac{Q_1^2}{\frac{2Q_2 M^*}{M^* - 1} + \frac{Q_1 p^*}{1 - p^*}}$$

M^*= total nr. of sampled cells, S^*= total nr. of found species, Q_i= nr. of species detected in i plot at scale p^*

Results

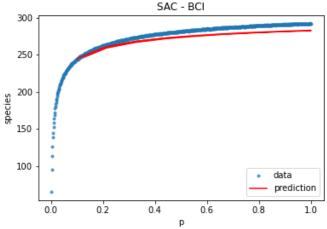

BCI
species
300
290
280
270
260
250
0.2
0.4
0.6
0.8
1.0
p
P/A NB
Chao_wor
S_{real}
BIRD
species
250
240
230
220
210
200
190
180
170
0.2
0.4
0.6
0.8
1.0
p
P/A NB
Chao_wor
S_{real}
relative percentage error
species
0
−1
−2
−3
−4
−5
−6
−7
0.2
0.4
0.6
0.8
1.0
p
relative percentage error
species
2
0
−2
−4
−6
−8
−10
−12
0.2
0.4
0.6
0.8
1.0
p

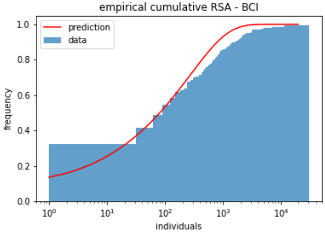
BCI results

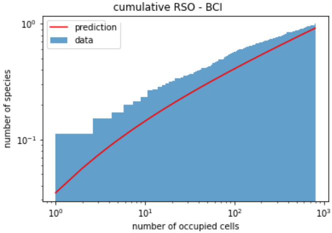
Using the $p^* = 0.1$ of the P/A matrix, infer SAC,RSA,RSO for the entire surveyed territory:

Parameters found:

r	ξ
0.37 ±0.01	0.998 ±0.004





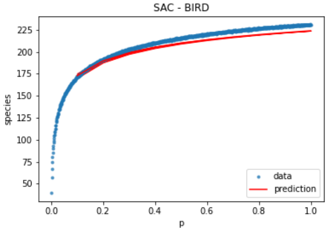


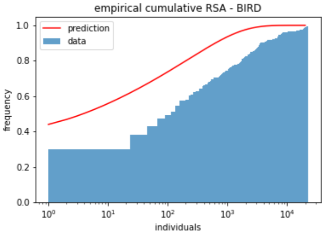
BIRD results

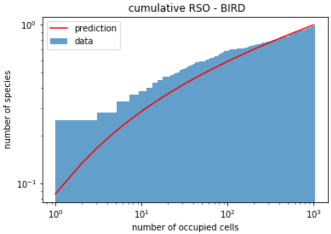
Using the $p^* = 0.1$ of the P/A matrix, infer SAC,RSA,RSO for the entire surveyed territory:

Parameters found:

r	ξ
0.11 $\pm$0.01	0.999 $\pm$0.004







Results

Infer species richness for the real entire territory.
BCI. This dataset represents the $p = 0.032$ of the whole BCI forest:

cell area [km^2]	p	S_{est}
25x25	0.032	337.0 ±0.2

BIRD. For this dataset we do not know which is the corresponding fraction of the French territory surveyed:

cell area [km^2]	p	S_{est}
5x5	0.05	312.0 ±0.6
10x10	0.2	284.0 ±0.3
15x15	0.44	266.0 ±0.2

Outline

1 Introduction

2 Model

3 In-silico forests

4 Real dataset application

5 Conclusion

6 Appendix

Conclusion

- The larger the sampled area the smaller the relative error.
- The more complete the dataset the more accurate the results.
- Results are compatible with real values.
- The method proposed in [1] is versatile.

Thanks for the attention!

Outline

1 Introduction

2 Model

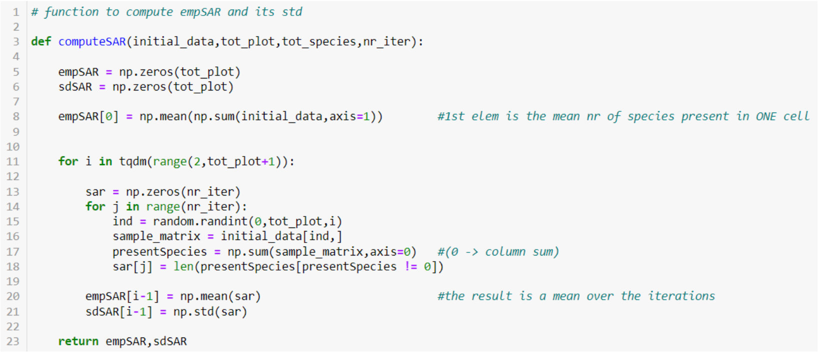
3 In-silico forests

4 Real dataset application

5 Conclusion

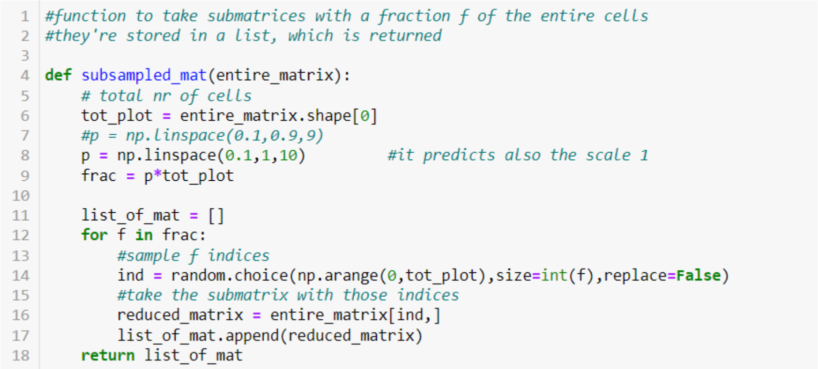
6 Appendix

Appendix



```python
# function to compute empSAR and its std

def computeSAR(initial_data,tot_plot,tot_species,nr_iter):

    empSAR = np.zeros(tot_plot)
    sdSAR = np.zeros(tot_plot)

    empSAR[0] = np.mean(np.sum(initial_data,axis=1))       #1st elem is the mean nr of species present in ONE cell


    for i in tqdm(range(2,tot_plot+1)):

        sar = np.zeros(nr_iter)
        for j in range(nr_iter):
            ind = random.randint(0,tot_plot,i)
            sample_matrix = initial_data[ind,]
            presentSpecies = np.sum(sample_matrix,axis=0)  #(0 -> column sum)
            sar[j] = len(presentSpecies[presentSpecies != 0])

        empSAR[i-1] = np.mean(sar)                         #the result is a mean over the iterations
        sdSAR[i-1] = np.std(sar)

    return empSAR,sdSAR
```

Appendix



```python
#function to take submatrices with a fraction f of the entire cells
#they're stored in a list, which is returned

def subsampled_mat(entire_matrix):
    # total nr of cells
    tot_plot = entire_matrix.shape[0]
    #p = np.linspace(0.1,0.9,9)
    p = np.linspace(0.1,1,10)          #it predicts also the scale 1
    frac = p*tot_plot

    list_of_mat = []
    for f in frac:
        #sample f indices
        ind = random.choice(np.arange(0,tot_plot),size=int(f),replace=False)
        #take the submatrix with those indices
        reduced_matrix = entire_matrix[ind,]
        list_of_mat.append(reduced_matrix)
    return list_of_mat
```

References

[1] Anna Tovo et al. "Inferring macro-ecological patterns from local presence/absence data". In: *Oikos* 128.11 (2019), pp. 1641–1652. DOI: https://doi.org/10.1111/oik.06754. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/oik.06754.

[2] Anna Tovo et al. "Upscaling species richness and abundances in tropical forests". In: *Science Advances* 3.10 (2017), e1701438. DOI: 10.1126/sciadv.1701438. URL: https://www.science.org/doi/abs/10.1126/sciadv.1701438.

[3] Anne Chao and Chun-Huo Chiu. "Species Richness: Estimation and Comparison". In: Aug. 2016, pp. 1–26. ISBN: 9781118445112. DOI: 10.1002/9781118445112.stat03432.pub2.