Physical Models of Living Systems

Project

Inferring macro-ecological patterns from local presence/absence data

Filippo Pra Floriani - 2089902

WHY

- Extrapolating species richness from the local to the global scale is not straightforward, because it is not additive as a function of the area.

- Many analytical methods have been proposed to upscale species richness using as input the local Relative Species Abundance distribution (RSA):
  - need abundance data

- Some non-parametric approaches have been generalized to infer species richness from this presence to absence data
  - no explicit dependence of the observation scale

OBJECTIVE

- Implement an algorithm for inferring biodiversity patterns from local presence/absence data:

  Species Accumulation Curve (SAC)

  Relative Species Abundance (RSA)

  Relative Species Occupancy (RSO)

- Test the algorithm over in-silico forests and Barro Colorado Island BCI dataset
- Infer patterns on a new dataset: Bird abundance data

THEORY

- The framework exploits the use of the form-invariance property of the Negative Binomial NB:
  - analytical expression for how parameters change across scales

- Postulate: RSA at global scale A, p = 1

$$\mathcal{P}(n \mid r,\xi) = \binom{n+r-1}{n} \xi^n (1-\xi)^r, \quad c(r,\xi) = \frac{1}{1-(1-\xi)^r} \qquad \text{for } n \geq 1$$

$$r > 0 \text{ and } 0 \leq \xi < 1$$

- Consider a sub-sample of area a of the whole forest and define p = a/A the sample scale:
  - Conditional probability that species have k individuals in area a, given that there are n individuals in whole area A:

$$P_{\text{binom}}(k \mid n, p) = \binom{n}{k} p^k (1-p)^{n-k}, \quad k = 0,\ldots,n$$

Assumption:
WELL MIXED population

THEORY

- In general n is not know, integrate with RSA at global scale p = 1 and get another Negative Binomial at scale p:

$$P(k \mid p) = \begin{cases} c(r,\xi) \times \mathcal{P}(k \mid r, \xi_p) & k \geq 1 \\ 1 - c(r,\xi) / c(r,\xi_p) & k = 0 \end{cases}$$

Conversion:

$$r_p = r \qquad \xi_p = \frac{p\xi}{1-\xi(1-p)}$$

- Find parameters of NB and total number of species S at global scale, from a sample scale p*:

$$r = r_{p^*}$$

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

$$S \simeq \frac{S^*}{1 - P(k=0 \mid p^*)}$$

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

THEORY

- Furthermore, we can link the parameters of two given scales p, p*:

$$\xi_p = \frac{p\xi_{p^*}}{p^* + \xi_{p^*}(p - p^*)}$$

- Find the number of species at scale p, knowing parameters at scale p*:

$$S_p \simeq S^* \frac{1-\left(1-U(p \mid p^*, \xi_{p^*})\right)^r}{1-(1-\xi_{p^*})^r}$$

where

$$U(p \mid p^*, \xi_{p^*}) \overset{\text{def}}{=} \frac{p\xi_{p^*}}{p^* + \xi_{p^*}(p - p^*)}$$

THEORY

- **Algorithm:** reconstruct the SAC
  - Whole forest area A, p = 1
  - M* cells of area a, p* = M*a/A
  - Presence/absence data for each cell ⇒ S_pk at sub-sample scale p_k

$$S_{p_k} \simeq S^* \frac{1-\left(1-U(p_k \mid p^*, \xi_{p^*})\right)^r}{1-(1-\xi_{p^*})^r} \qquad \text{for } p_k = ka/A,\ k = 1, \ldots, M^*$$

$$= S^* \frac{1-\left(1-U(\tilde{p}_k \mid 1, \xi_{p^*})\right)^r}{1-(1-\xi_{p^*})^r}$$

$$\text{where} \quad \tilde{p}_k = p_k / p^* = k / M^* \qquad U(p \mid p^*, \xi_{p^*}) = U\left(\frac{p}{p^*} \mid 1, \xi_{p^*}\right)$$

- It is possible to know the number of species at each sub-sample scale and so estimate the parameters at sample scale p* shaping the RSA

THEORY

- Predict RSO pattern: number of occupied cells
  - Information on the fraction of species that occupy the same amount of area of the ecosystem

- Probability that a species occupies v over M cells at the global scale, given that is has abundance n:

$$Q_{occ}(v \mid n, M, 1) = \frac{\binom{M}{v}\binom{n-1}{v-1}}{\binom{n+M-1}{M-1}}$$

Assumption on of absence of spatial correlation:

Hyper-geometric distribution

- Marginalizing over n using RSA at global scale p = 1:

$$Q(v \mid M, 1) = \sum_{n=v}^{\infty} Q_{occ}(v \mid n, M, 1) P(n \mid 1)$$

FRAMEWORK

- Consider a sample at scale p*, covering M* cells and with S* species observed:
  - To each cell, associate the vector $\omega_s^i \in \{0,1\}, s \in \{1,\ldots,S^*\}, i \in \{1,\ldots,M^*\}$ giving information on presence/absence data of the species in the cell
  - A area of whole ecosystem
  - M* cells of area a, so that p* = M*a/A
  - At each sub-sampling scale $p_k = ka/A$ compute the average number of observed species:

$$S_{\text{emp}}(p_k) = \frac{\mathrm{I}}{\binom{M^*}{k}} \sum_{\substack{I \subseteq \{1,\ldots,M^*\} \\ |I|=k}} \sum_{s=1}^{S^*} \mathrm{I}\left(\sum_{i \in I} \omega_s^i \geq \mathrm{I} \right)$$

For every scale p_k, compute the empirical average of the number of the species observed in all subsets of k cells:

- unfeasible, make average over 100 randomly chosen subsets

FRAMEWORK

- Fit the empirical SAC and extract parameters of NB at sample scale p*.
- Upscale the parameters to global scale p = 1
- Find RSA and RSO distribution, starting from absence/presence data

NOTE:

This method works well in absence of spatial correlations:

- furthermore, in the average over the subsets we are neglecting any spatial information

RESULTS

The framework and algorithm is tested over 2 kind of datasets:

- In-silico generated forests:
  - NB RSA
  - LN RSA
- Real datasets:
  - Barro Colorado Island BCI trees
  - Birds abundance data from French Breeding Survey

RESULTS – IN-SILICO FORESTS

For both RSA:

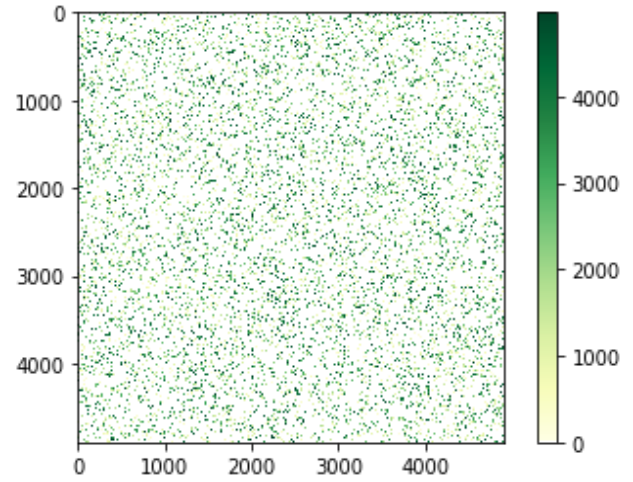
- Generate abundances

- Distribute individuals in a grid of 4900 x 4900 units, the forest area
- Two different process:
  - At random
  - According to a Thomas modified process with clustering radius of 15 units

- Divide the forest into M = 98×98 units cells
- Compute the M×S presence/absence matrix, thus forgetting the information about the species abundance
- Sub-sample the 5% of the cells (corresponding to a fraction p = 0.05 of the total forest area)
- Apply the algorithm

RESULTS – IN-SILICO FORESTS

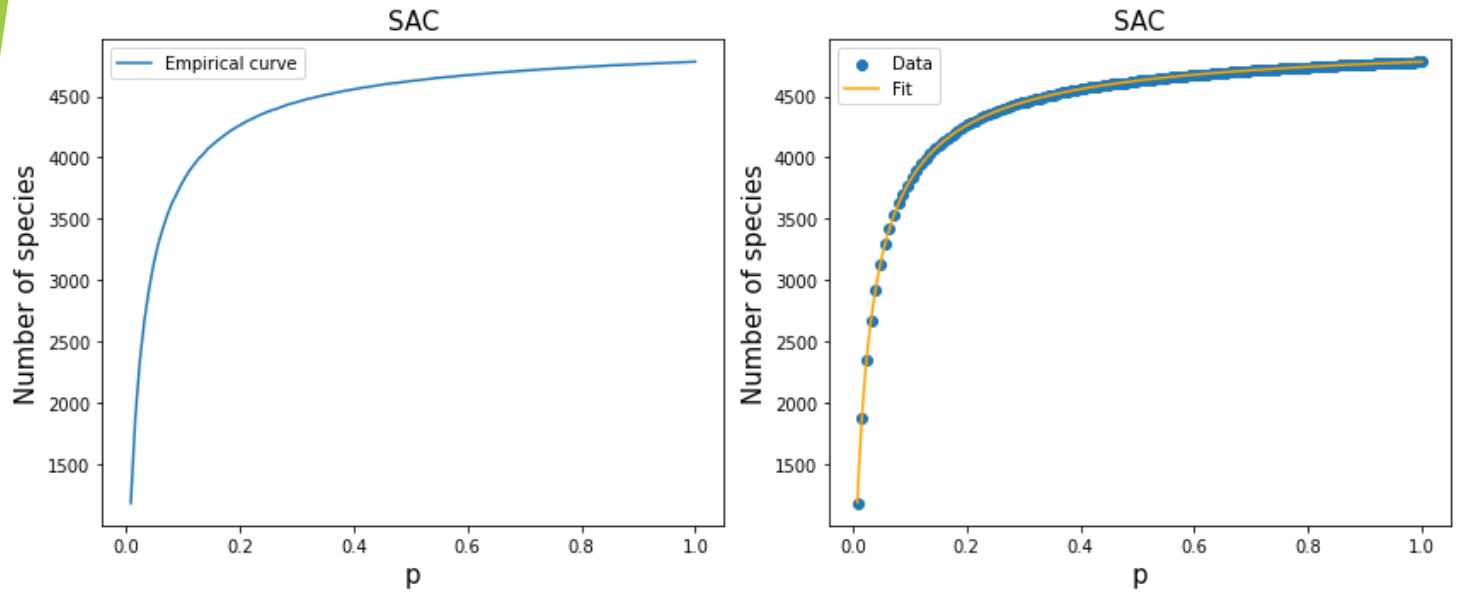
NEGATIVE BINOMIAL - RANDOM DISTRIBUTION

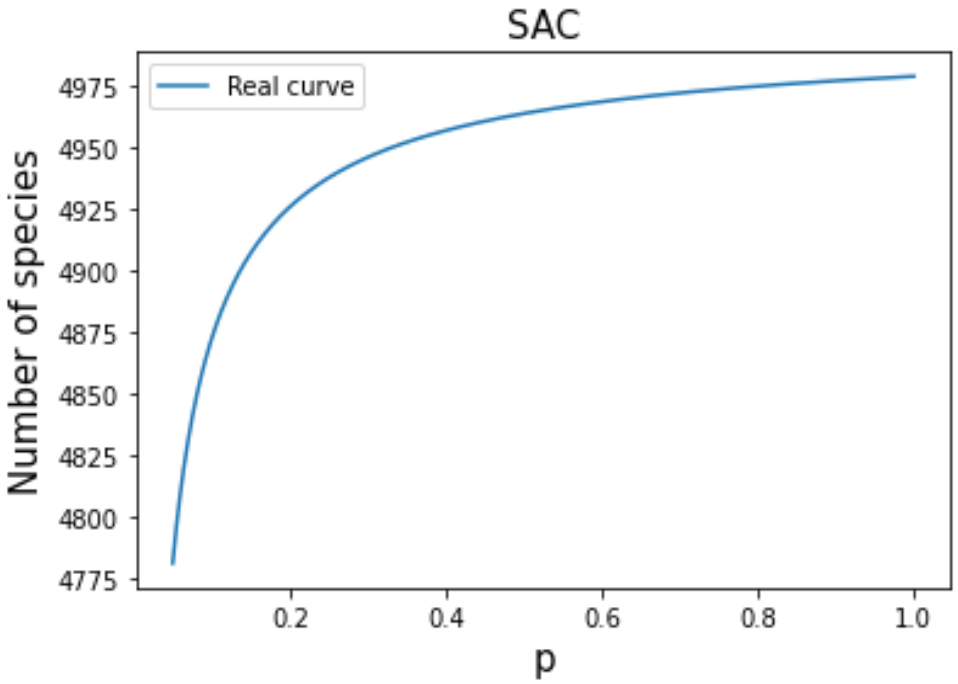


- Generate abundances from a NB RSA with parameters r = 0.8 and ξ = 0.999.
- Set the number of species S = 4974.

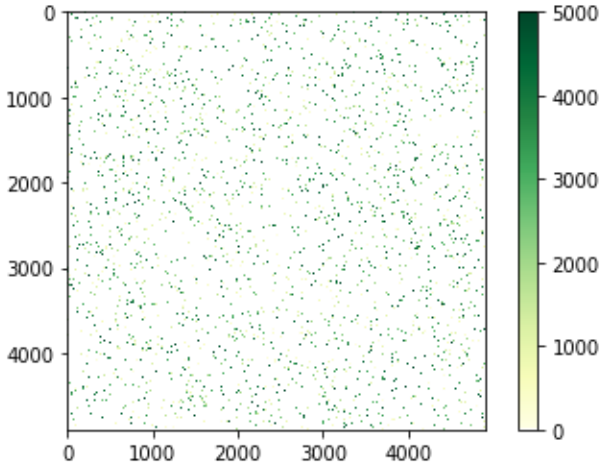
Results relative to 10 simulations:

- S = 4972 +- 0.2
- Relative Error = -0.041 +- 0.009





RESULTS – IN-SILICO FORESTS

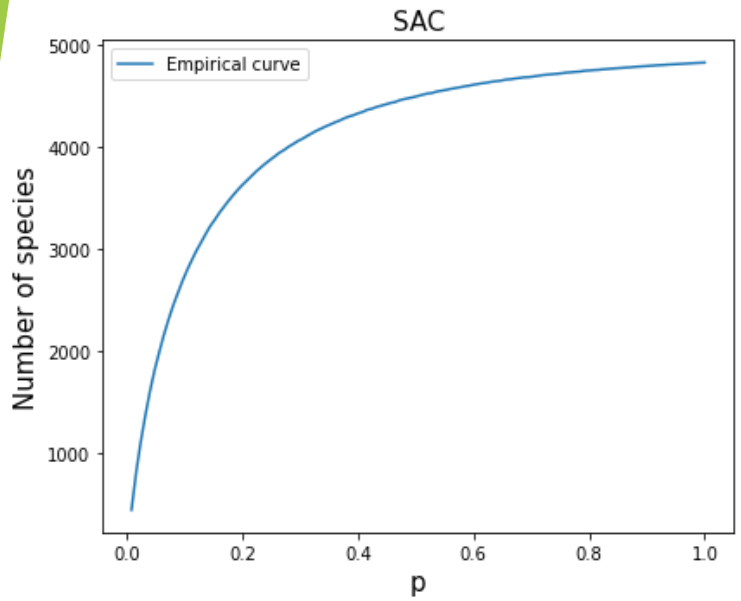


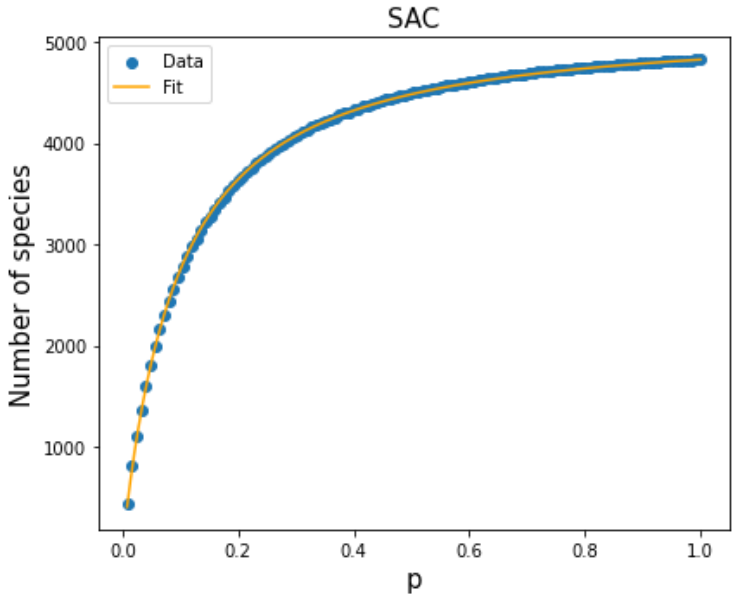
LOG NORMAL- RANDOM DISTRIBUTION

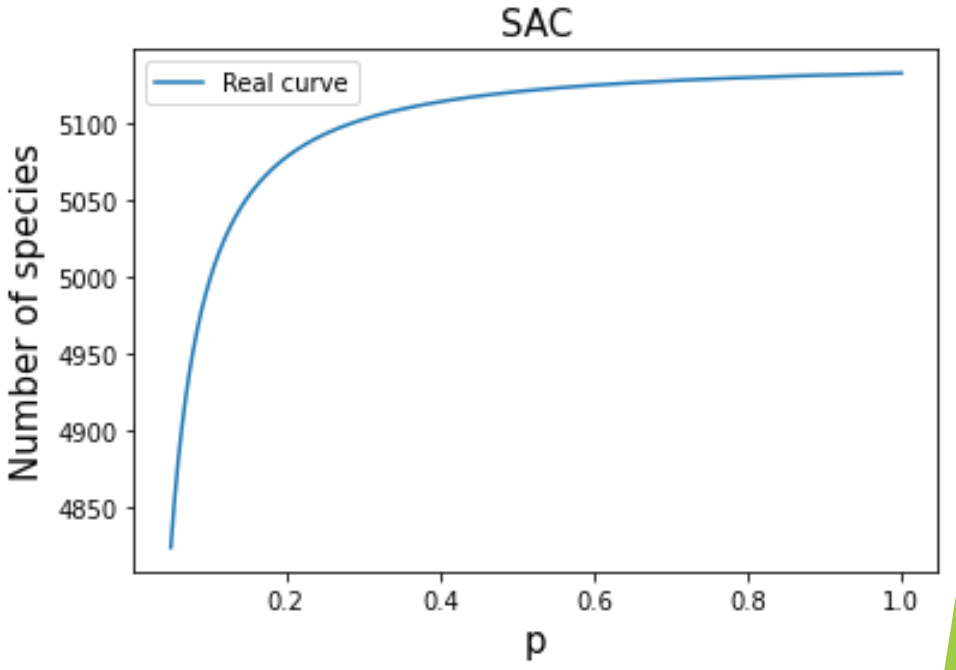
- Generate abundances from a LN RSA with parameters μ = 5 and σ = 1.
- Set the number of species S = 5000.

Results relative to 10 simulations:

- S = 5144.6 +- 0.8
- Relative Error = 2.89 +- 0.06

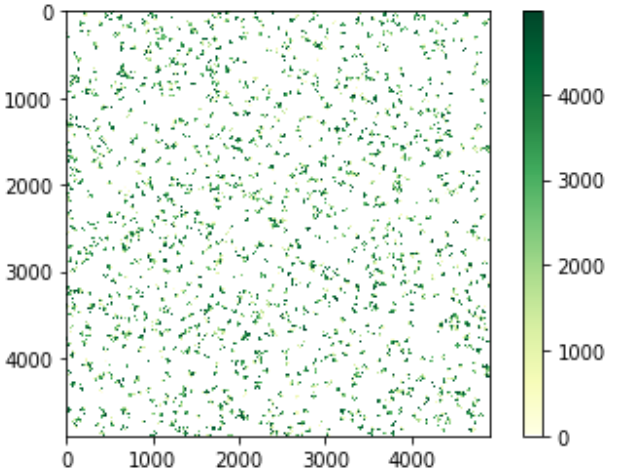






RESULTS – IN-SILICO FORESTS

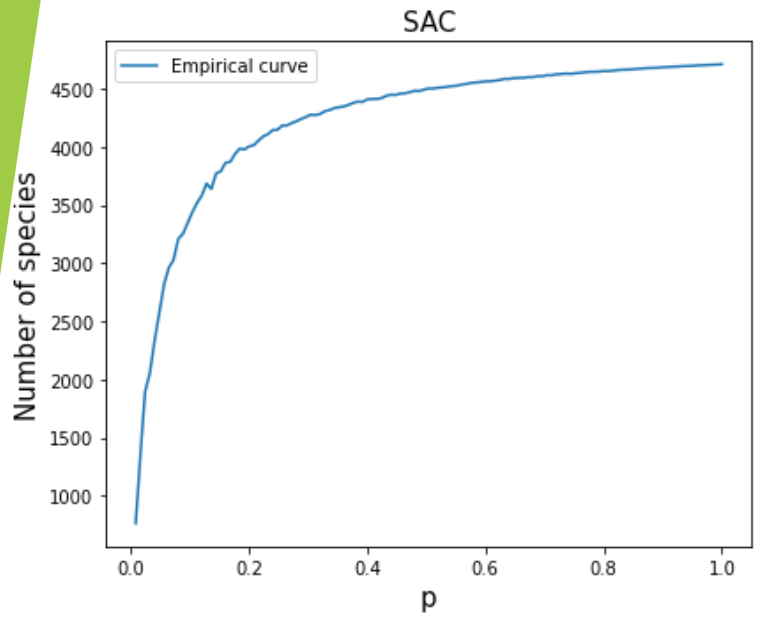
NEGATIVE BINOMIAL - CLUSTERING DISTRIBUTION

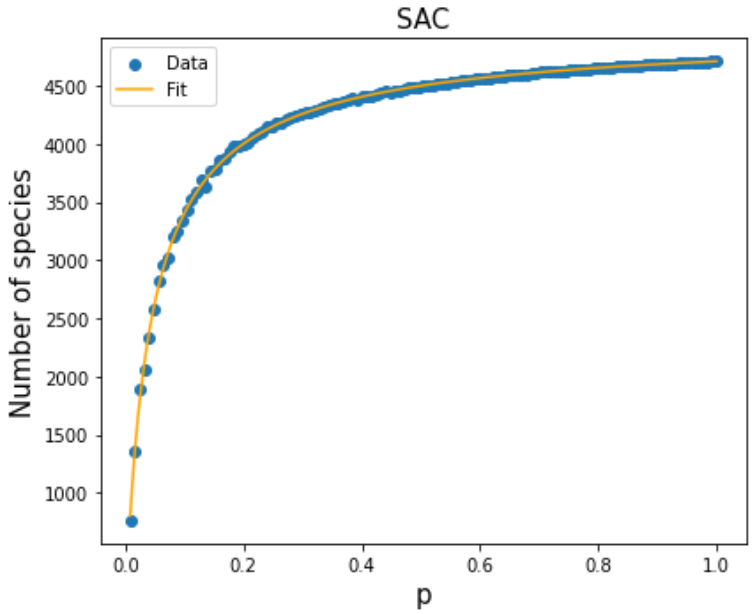


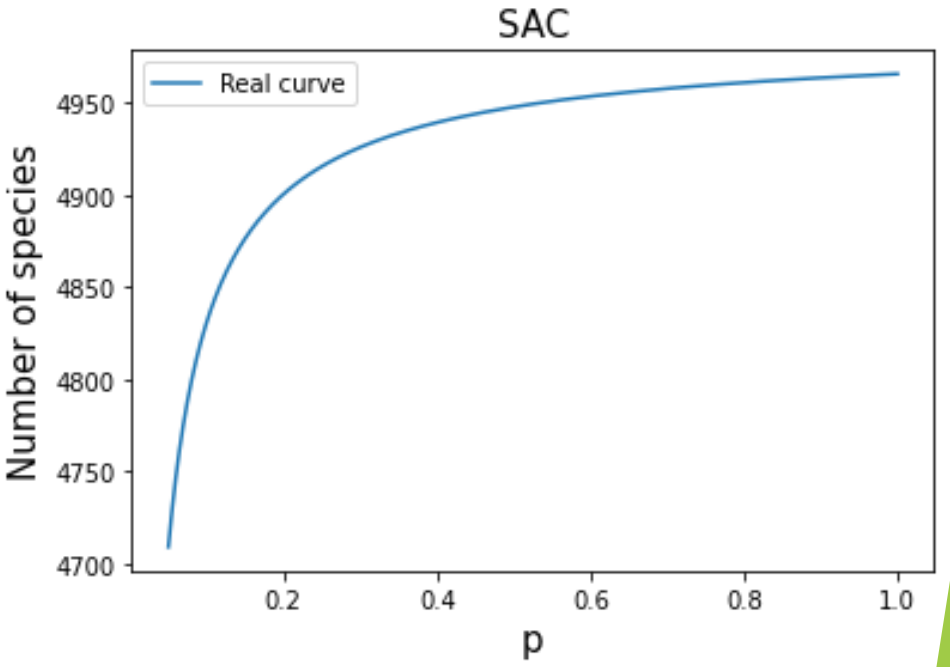
- Generate abundances from a NB RSA with parameters r = 0.8 and ξ = 0.999.
- Set the number of species S = 4974.

Results relative to 10 simulations:

- S = 4955 +- 2
- Relative Error = -0.4 +- 0.1

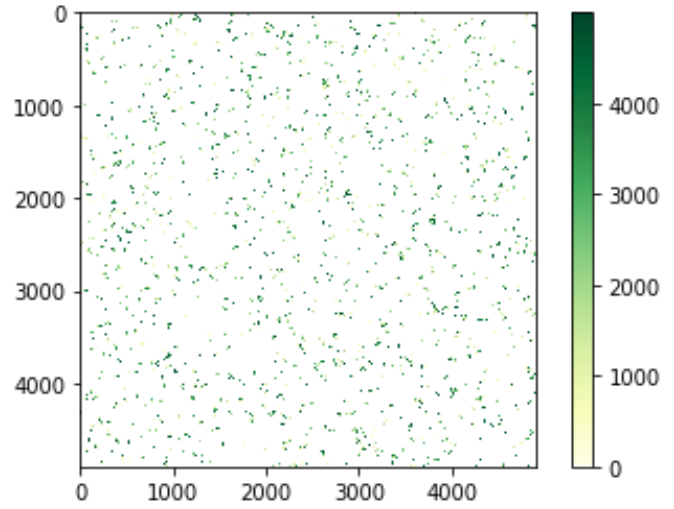






RESULTS – IN-SILICO FORESTS

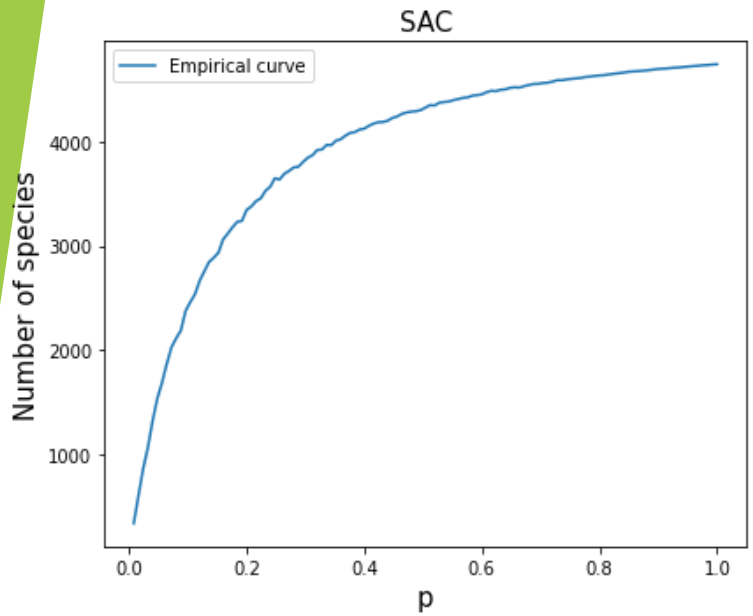
LOG NORMAL- CLUSTERING DISTRIBUTION

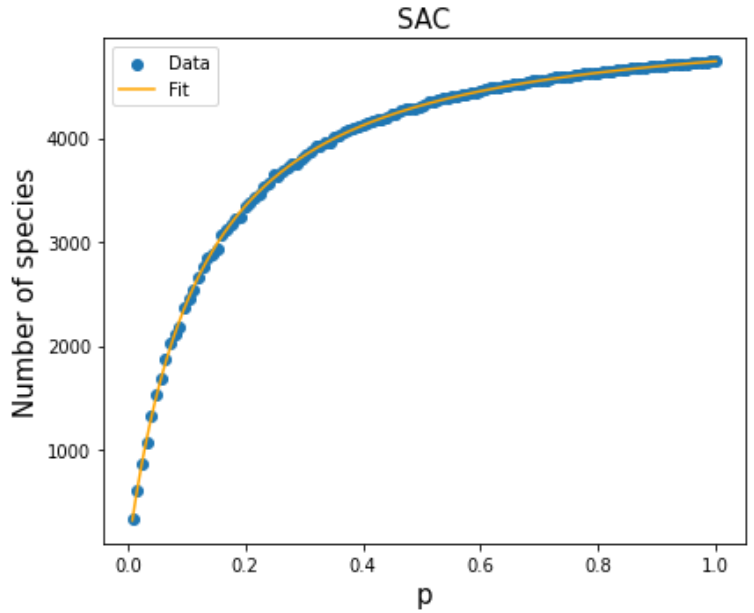


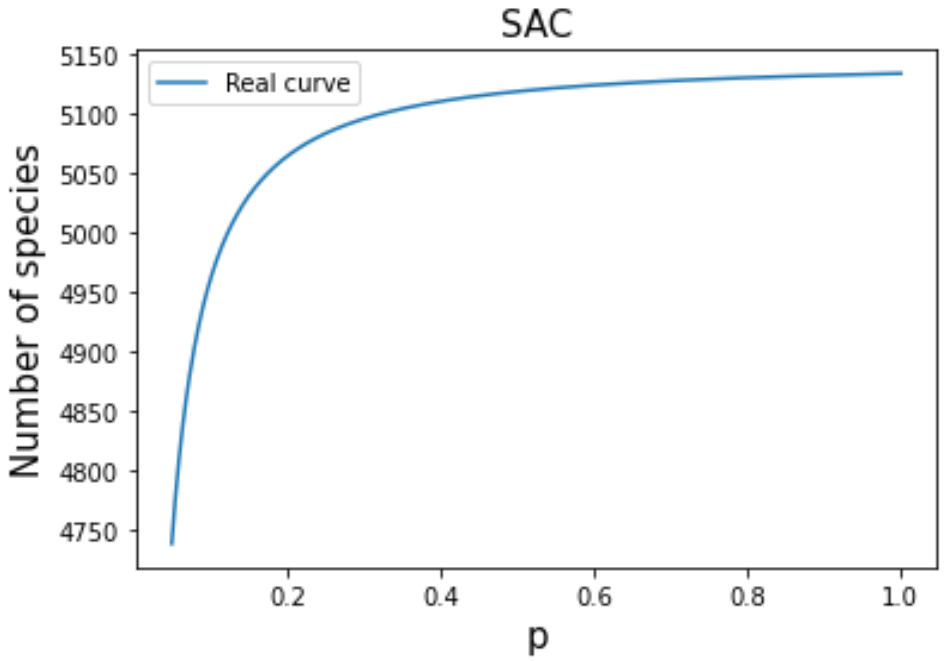
- Generate abundances from a LN RSA with parameters μ = 5 and σ = 1.
- Set the number of species S = 5000.

Results relative to 10 simulations:

- S = 5142+- 2
- Relative Error = 2.8 +- 0.2







RESULTS – BCI

Filter the dataset and extract data correspoding to a fraction $p^* = 0.032$ of whole area:

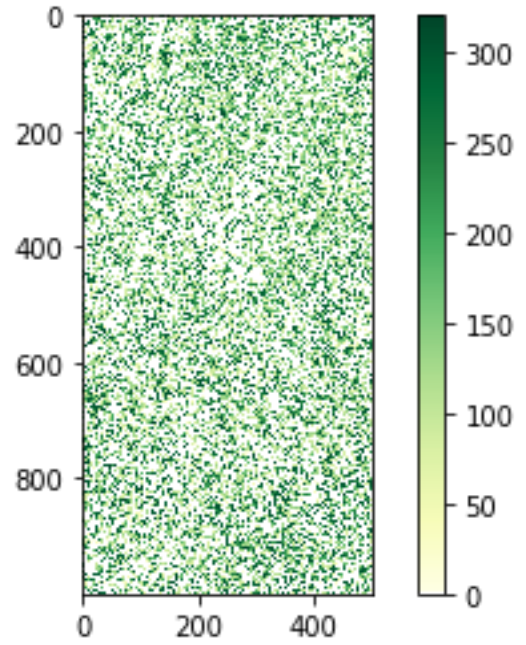
- Distribute individuals in a grid in the location where they are found

- Divide the forest into M = 800 units cells
- Compute the M×S* presence/absence matrix, for S* = 320
- Sub-sample the 10% of the cells (corresponding to a fraction p_ = 0.1 of the sample forest area)
- Apply the algorithm for $0 < p < p\_$
- Compute ecological patterns: SAC, RSA, RSO

- Eventually, consider the whole dataset as the sample scale
- Apply the framework for $0 < p < p^*$
- Extract parameters of RSA and number of species at global scale

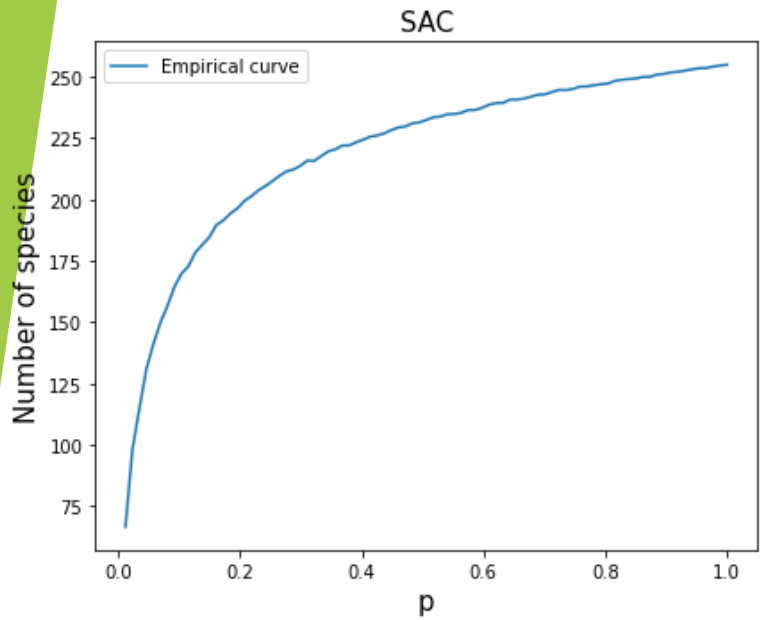
RESULTS – BCI

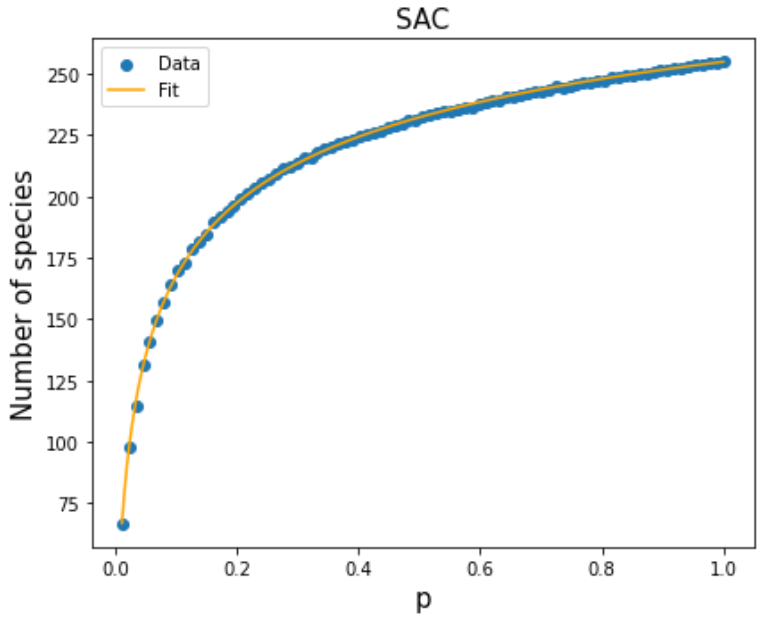
- Distribute individuals of the sample p* with S* = 320 species
- Sub-sample p_ = 0.1p*
- Apply the algorithm

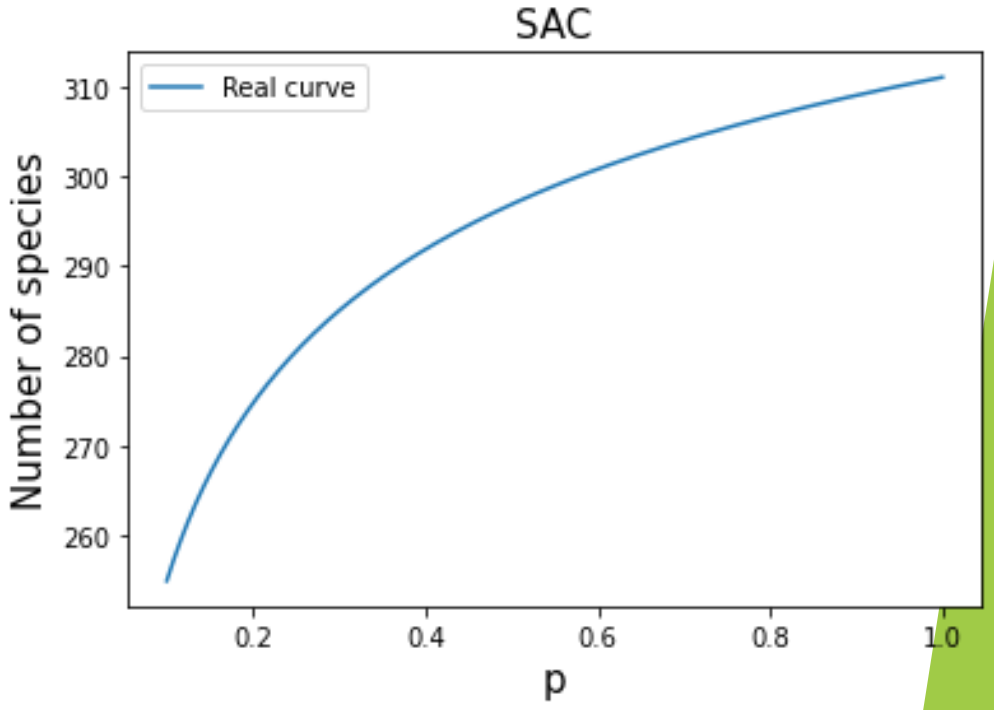


Results relative to 100 simulations:

- S = 310.2 +- 0.2
- Relative Error = -3.0 +- 0.4





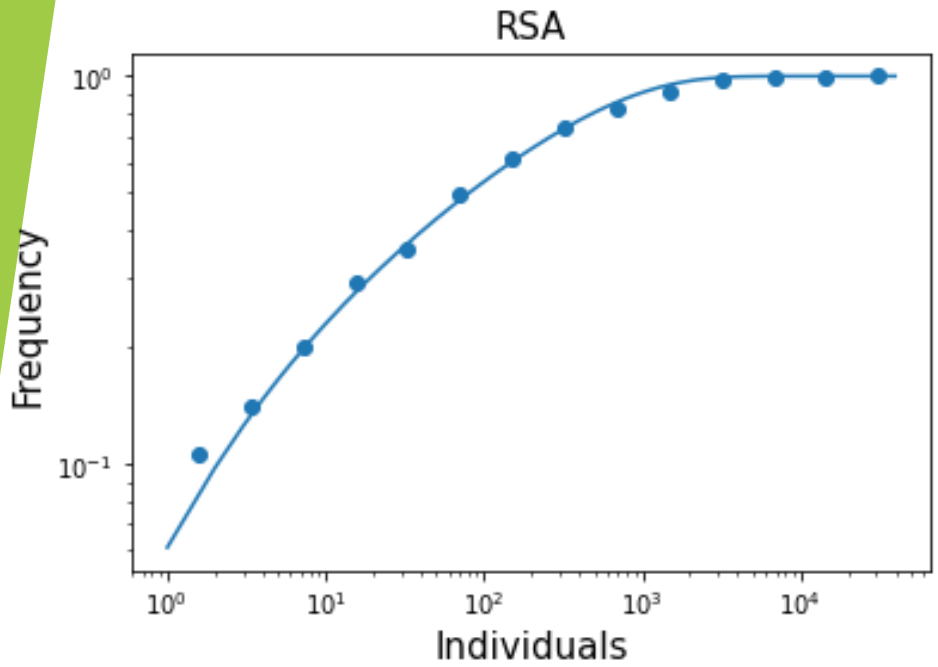


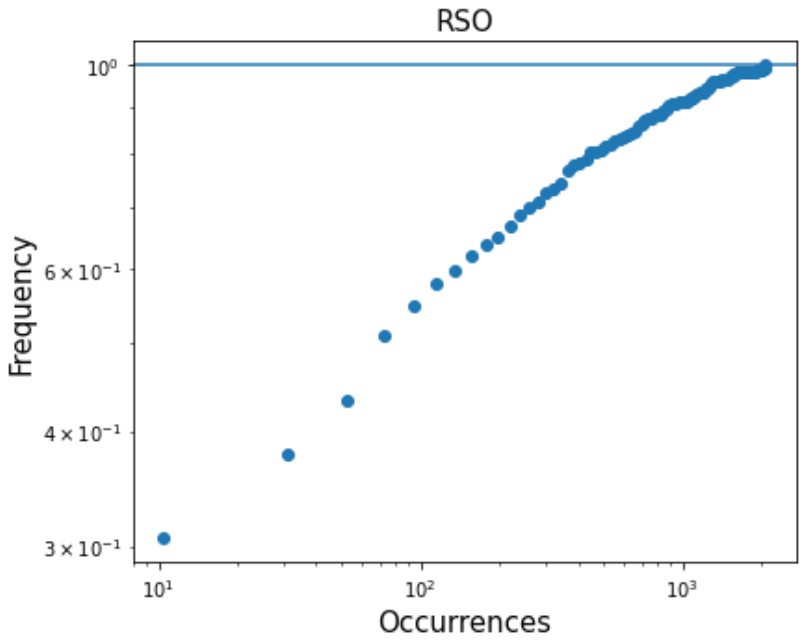
RESULTS – BCI

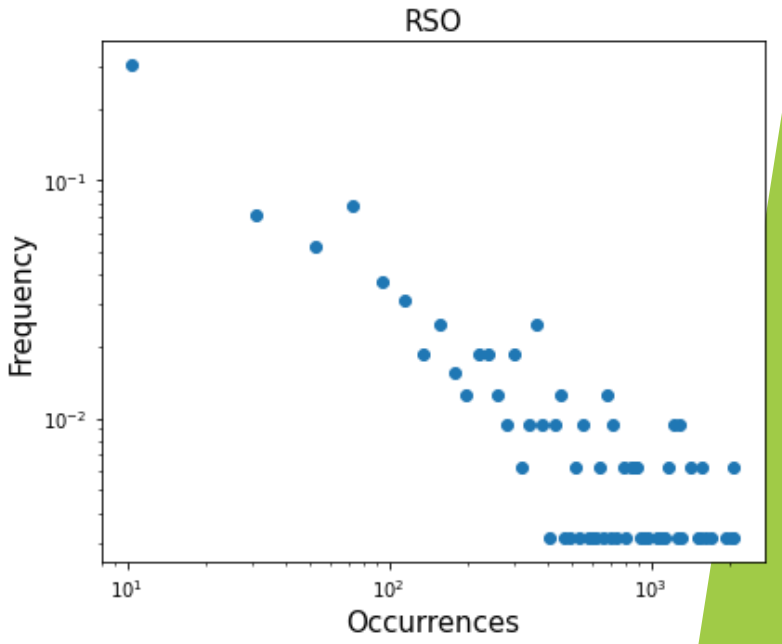
- Evaluate ecological patterns RSA, RSO at scale p*

- Use abundance data for RSA available at sample scale p*
- Superimpose the RSA curve with the estimated parameters

- Compute the RSO from the sample p*:
  - dependence on number of cells, 2000 cells results reported
  - power law behaviour







RESULTS – BCI

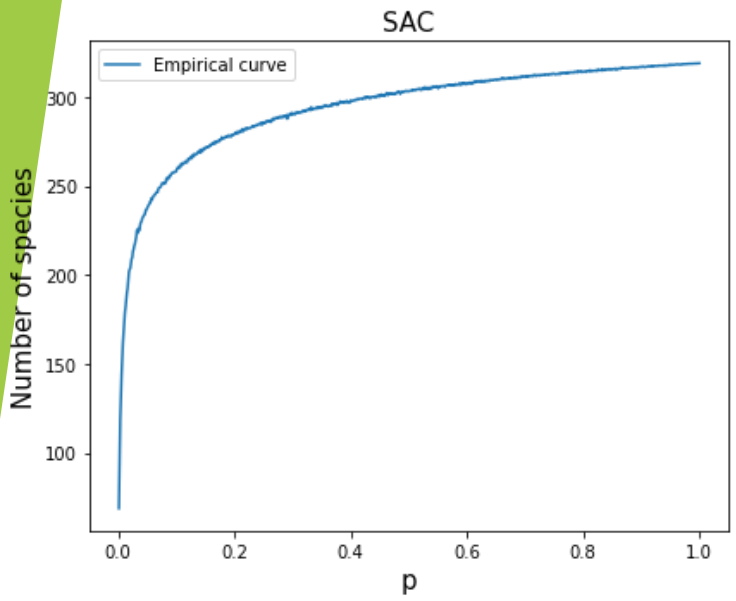
- Scale from sample scale p* to global scale p = 1
- Give an estimate for total number of species S and RSA parameters

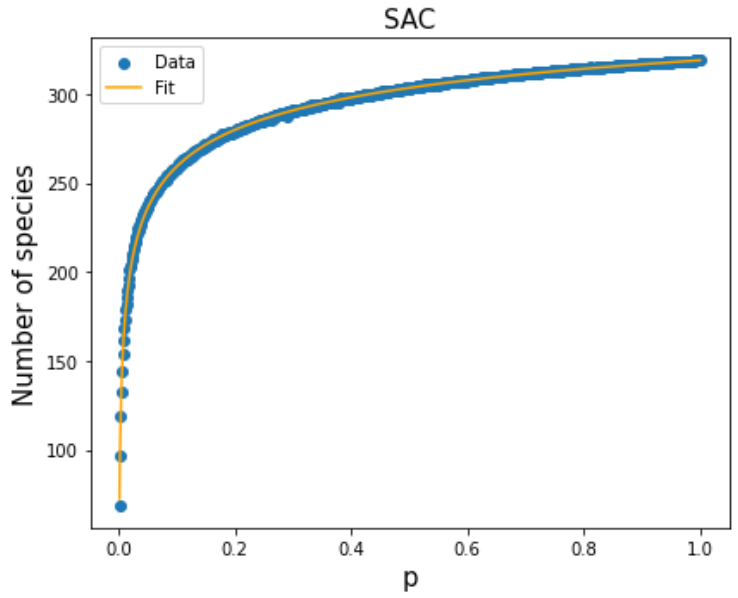
Results relative to 100 simulations:

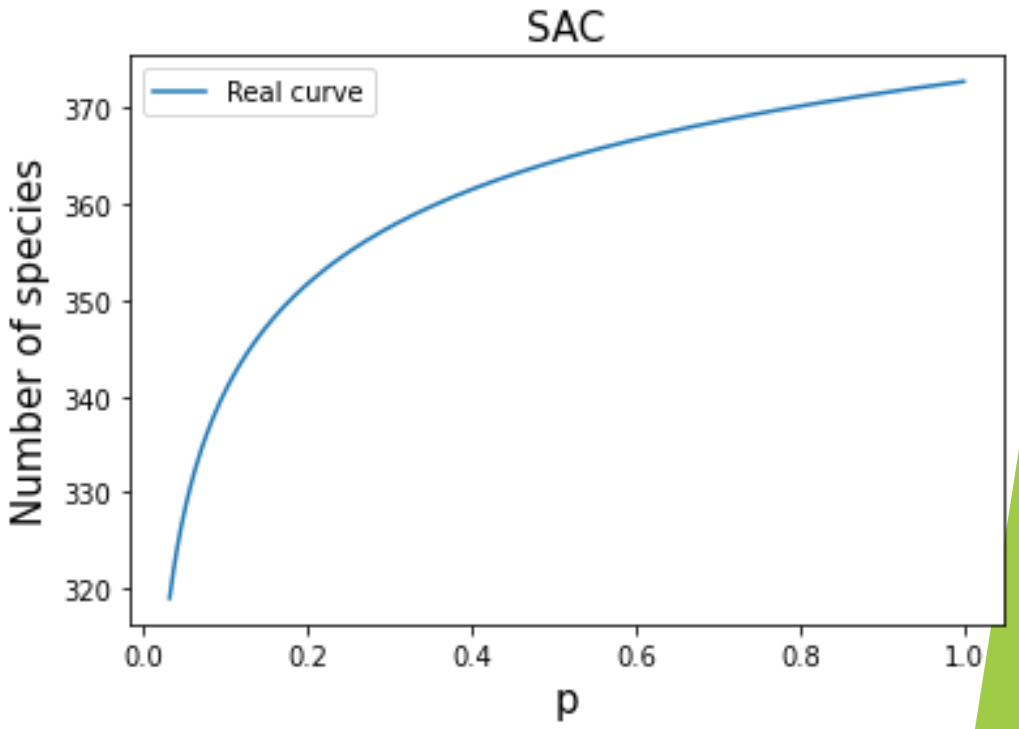
- S = 372.9 +- 0.5

Parameters RSA at global scale:

- r = 0.179 +- 0.008
- ξ = 0.9998 +- 0.0008







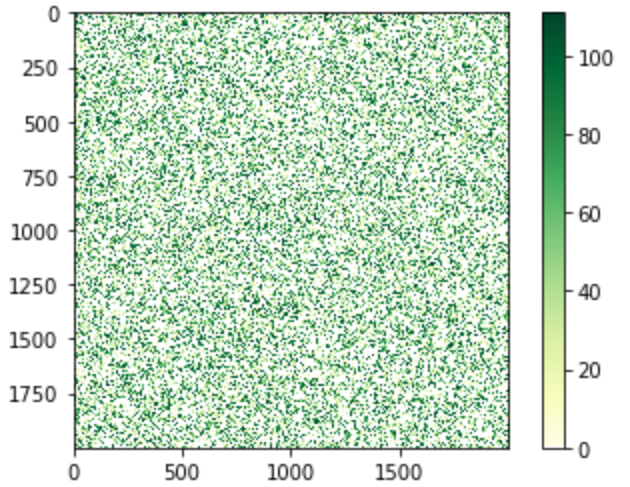
RESULTS – BIRDS DATA

Filter the dataset and extract data corresponding to birds identification over a 2Km x 2Km area consisting of 10-points-counts separated by at least 300m.

- Distribute individuals in a grid with 2 approaches
  - At random
  - Clustering at site positions, with radius 300m

- Divide the area into M = 1600 units cells, of side 50m
- Compute the M×S* presence/absence matrix, for S* = 111
- Sub-sample a fraction of the cells (corresponding to a fraction p_ of the total area)
- Apply the algorithm for 0<p<p_
- Compute ecological patterns: SAC, RSA, RSO

RESULTS – BIRDS DATA



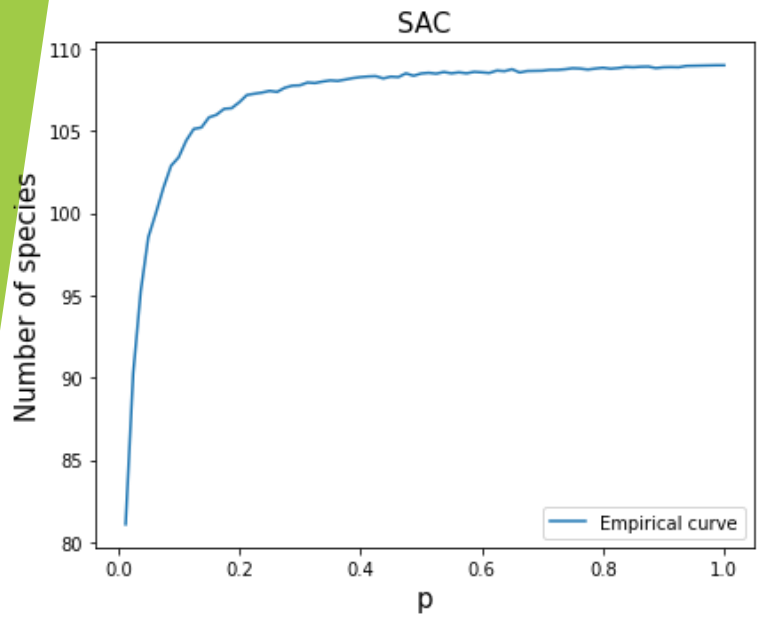
- Distribute individuals of the sample with S* = 111 species at **random**
- Sub-sample p_ = 0.05 p*
- Apply the algorithm

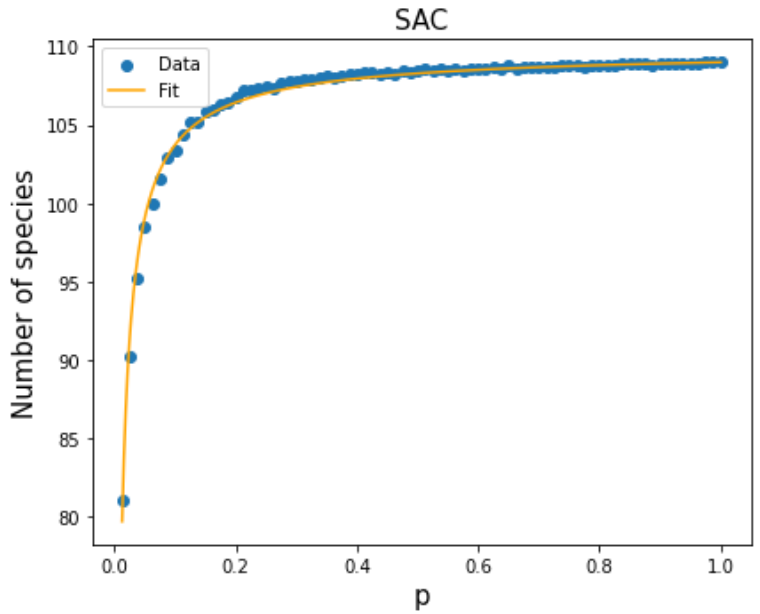
Results relative to 100 simulations:

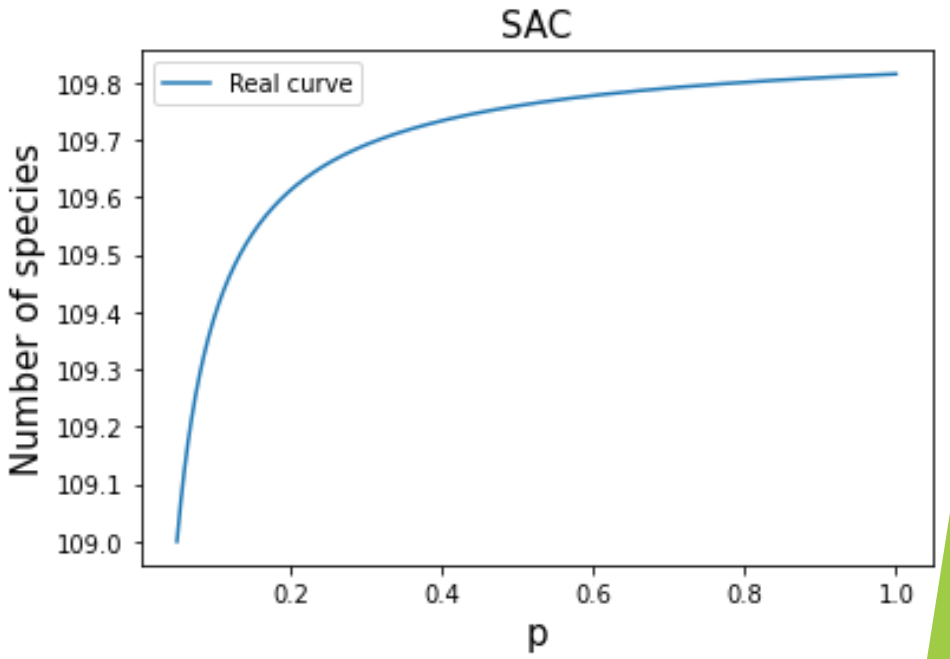
- S = 109.156 +- 0.005
- Relative Error = -1.66 +- 0.07

Parameters RSA at sample scale:

- r = 0.91 +- 0.01
- r = 0.9998 +- 0.0016





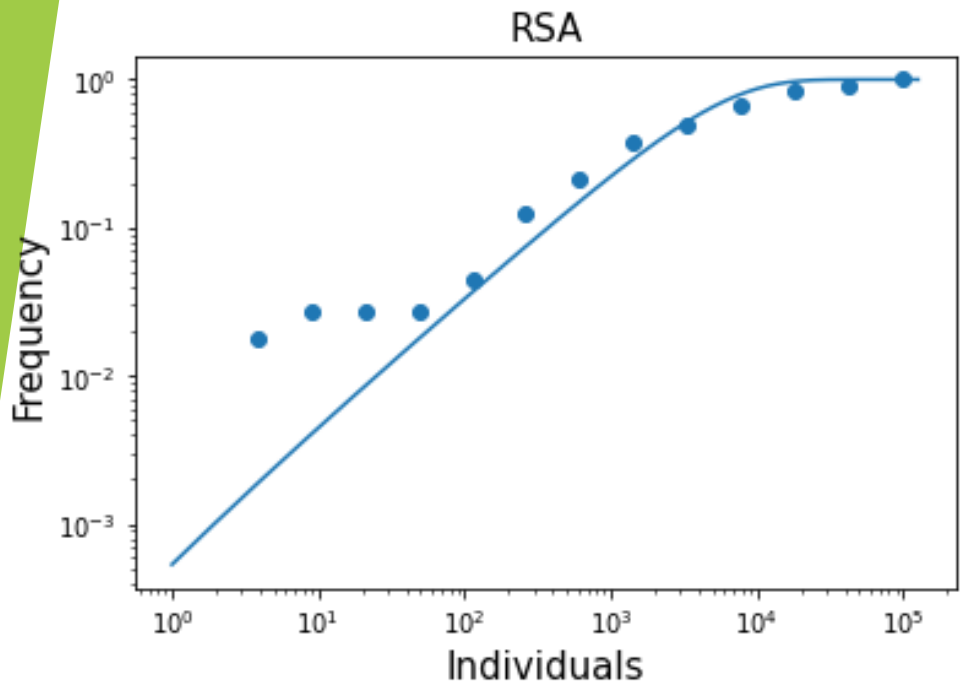


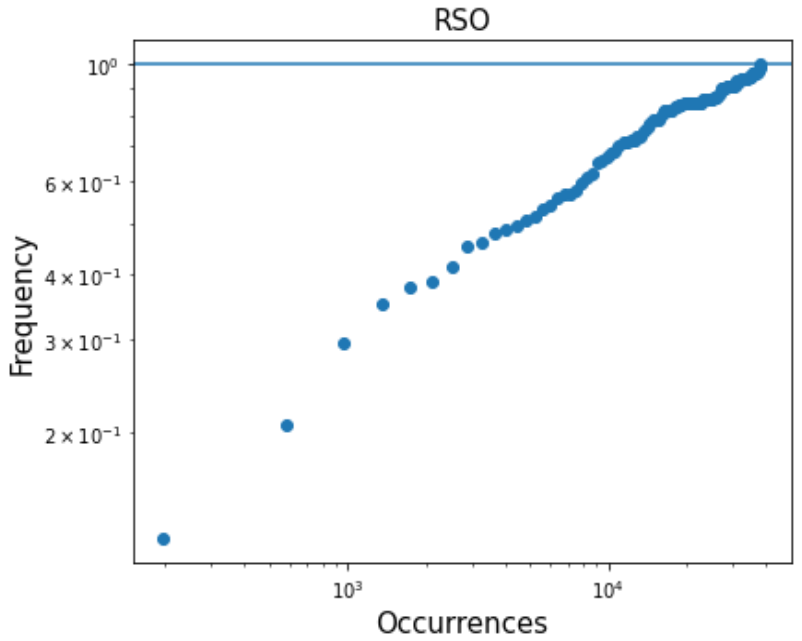
RESULTS – BIRDS DATA

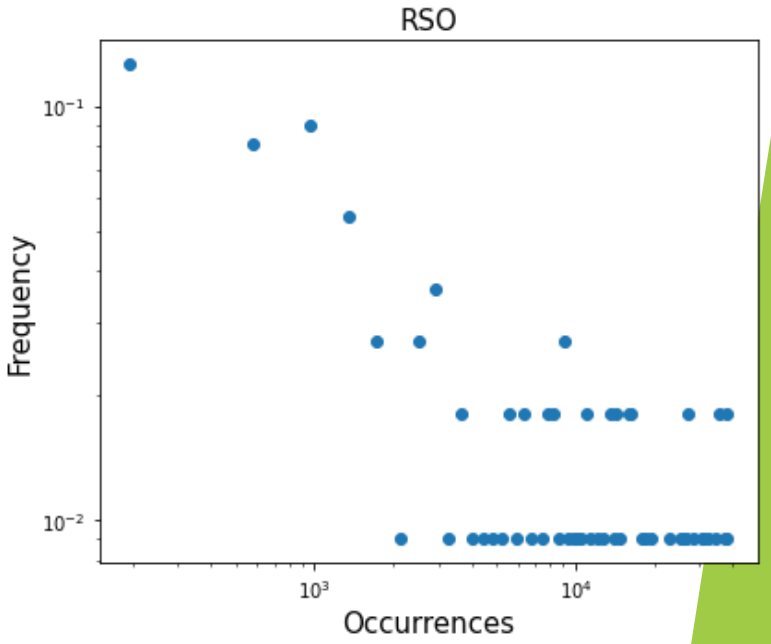
- Evaluate ecological patterns RSA, RSO at sample scale

- Use abundance data for RSA available at sample scale
- Superimpose the RSA curve with the estimated parameters

- Compute the RSO from the sample:
  - dependence on number of cells, 40000 cells (side 10m) results reported
  - power law behaviour

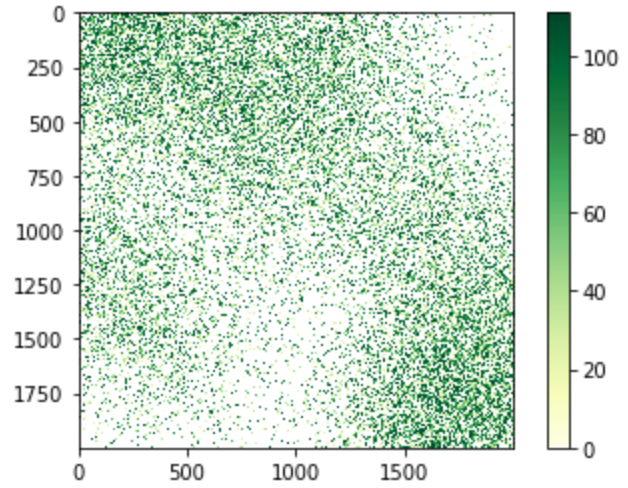






RESULTS – BIRDS DATA

- Distribute individuals of the sample with S* = 111 species with **clustering**
- Sub-sample p_ = 0.025 p*
- Apply the algorithm

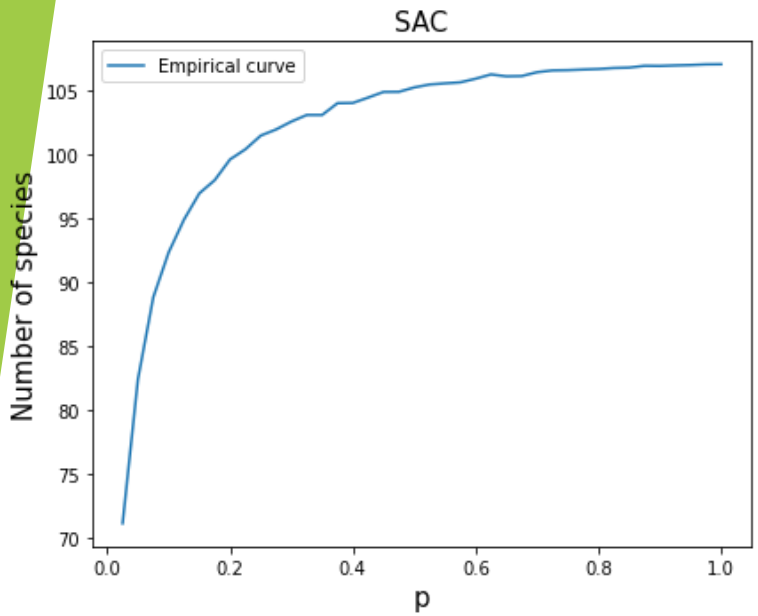


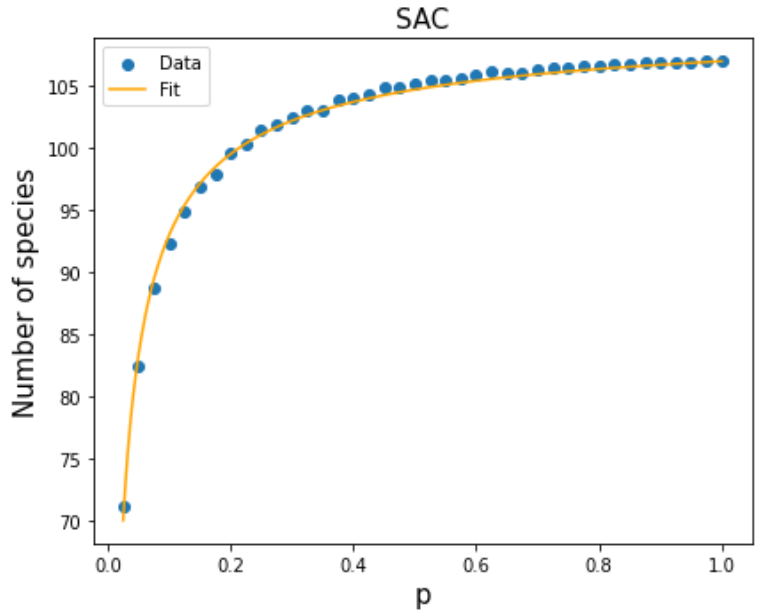
Results relative to 100 simulations:

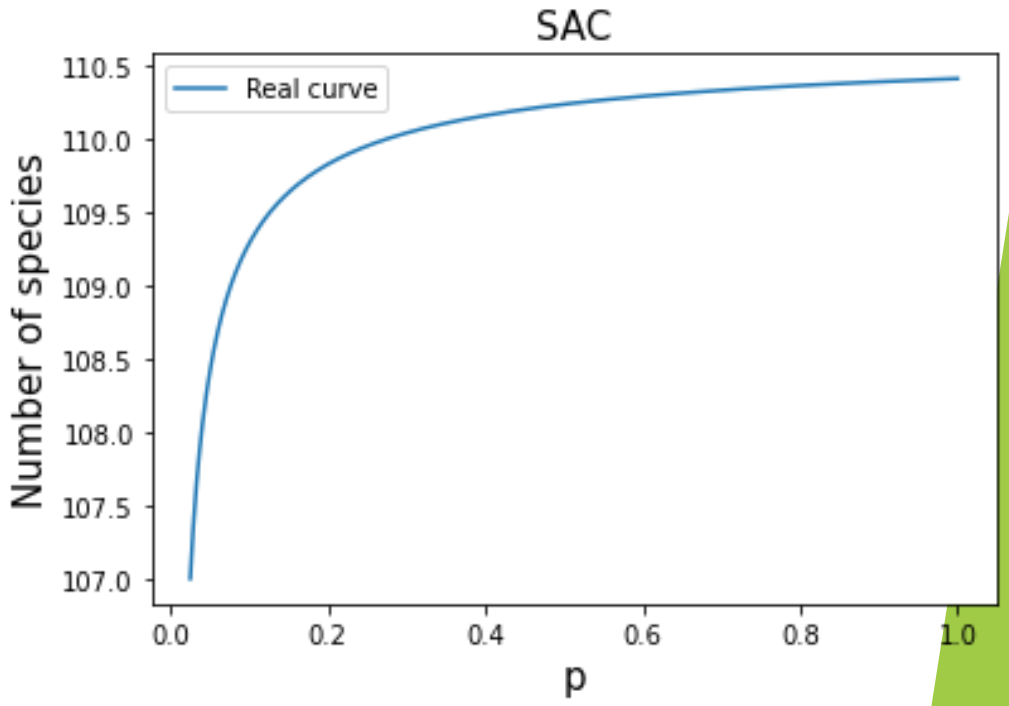
- S = 110.21 +- 0.03
- Relative Error = -0.7 +- 0.2

Parameters RSA at sample scale:

- r = 0.82 +- 0.02
- ξ = 0.9997 +- 0.0024





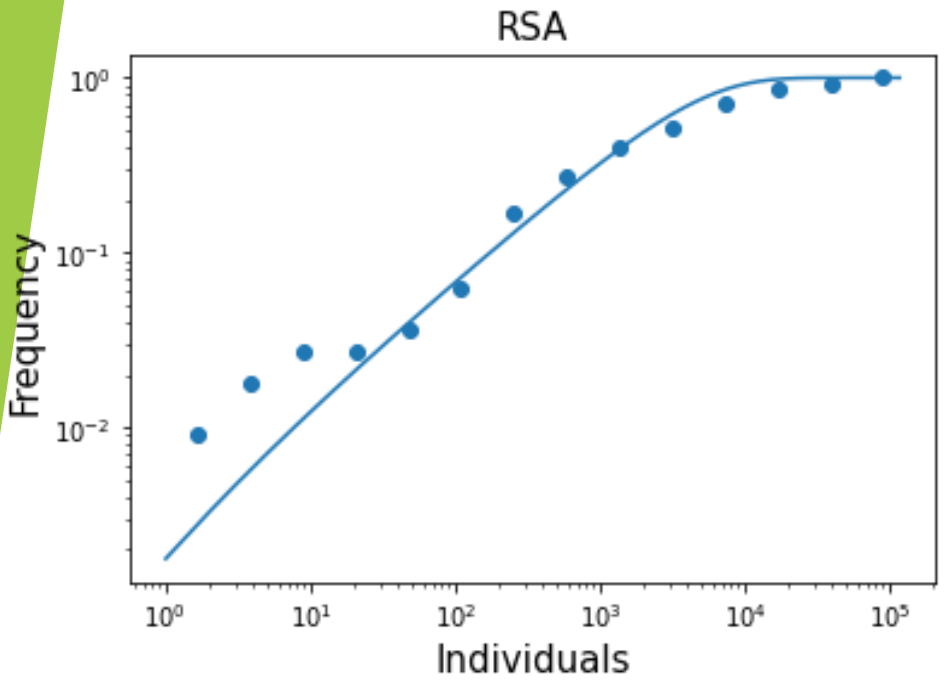


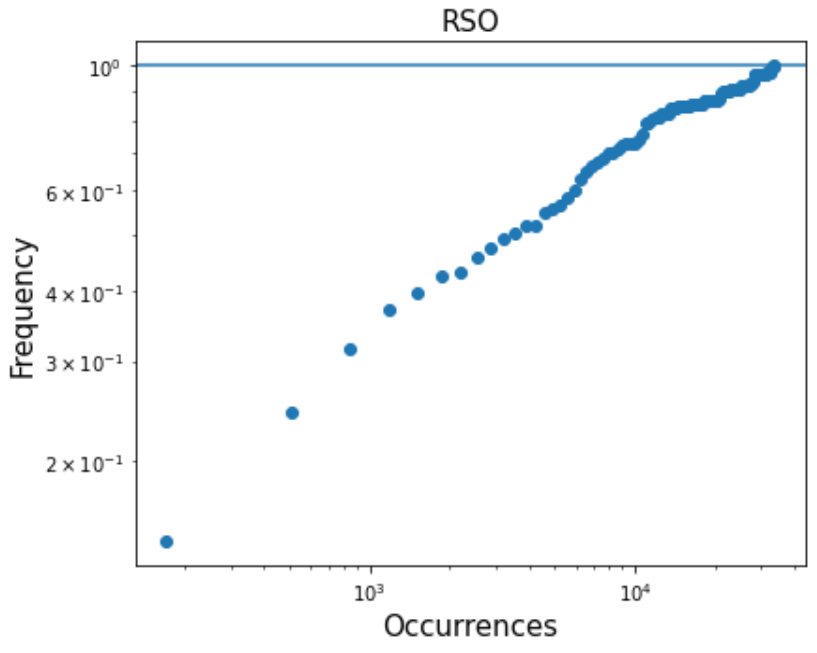
RESULTS – BIRDS DATA

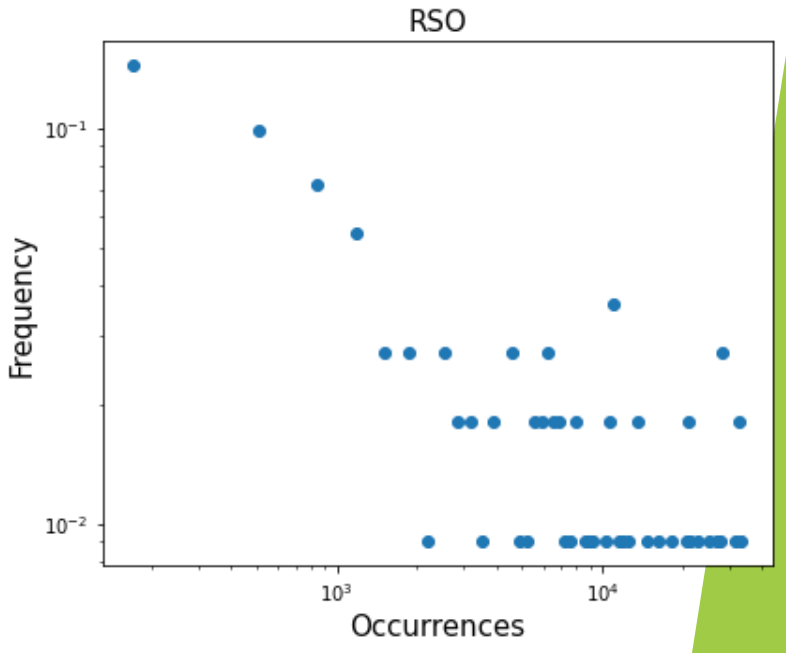
- Evaluate ecological patterns RSA, RSO at sample scale

- Use abundance data for RSA available at sample scale
- Superimpose the RSA curve with the estimated parameters

- Compute the RSO from the sample:
  - dependence on number of cells, 40000 cells (side 10m) results reported
  - power law behaviour







CONCLUSIONS

- The algorithm for extracting macro-ecological patterns at global scale starting from local presence/absence data works.
- Issues arise when data present spatial correlations.

THANKS