UNIVERSITÀ DEGLI STUDI DI PADOVA

Dipartimento di Fisica e Astronomia "Galileo Galilei"

Corso di Laurea in Fisica

Tesi di Laurea

Proprietà di scala della biodiversità in comunità microbiche

Relatore

Prof. Samir Suweis

Correlatrice

Dott.ssa Anna Tovo

Laureanda

Eleonora Manoli

Anno Accademico 2017/2018

Abstract

Uno dei problemi aperti in Ecologia è quello di caratterizzare la biodiversità di un ecosistema stimandola attraverso censimenti locali, che tipicamente coprono solamente una minima percentuale dell'area su cui si estende il sistema in esame. In questa tesi, verranno presentati alcuni dei metodi proposti in letteratura per superare tale problema, con particolare attenzione ad un modello di upscaling presentato nell'articolo "Upscaling species richness and abundances in tropical forests", recentemente pubblicato su Science Advances [1].
Dopo aver testato l'affidabilità dei modelli proposti, applicheremo per la prima volta il metodo di upscaling per lo studio della biodiversità in comunità microbiche, utilizzando dati di metagenomica relativi ai batteri dell'intestino umano [2].

Indice

Introduzione

Una sfida cruciale per la ricerca in ecologia è quella di comprendere come le varie grandezze, tra cui la biodiversità e l'abbondanza delle specie, cambiano attraverso le diverse scale spaziali [3]. Infatti, in un ecosistema, le condizioni fisiche variano in maniera complicata nello spazio e nel tempo e le diverse popolazioni possono interagire in modi che dipendono dalla scala.

Le diverse cause dei cambiamenti ecologici, siano esse naturali o antropogeniche, tendono a manifestarsi a scale diverse, e come risultato i cambiamenti nella biodiversità possono essere diversi a seconda della scala a cui li si analizza. Anche i nostri interessi riguardo ai sistemi naturali variano a seconda della scala; gli obiettivi di conservazione possono riguardare scale globali, nazionali o regionali, mentre questioni di servizi ecosistemici, cioè i benefici forniti dagli ecosistemi, riguardano le proprietà delle comunità a scala locale.

Introduciamo ora alcune definizioni che utilizzeremo nel corso della tesi. Con il termine *biodiversità* indicheremo sia la ricchezza di specie, sia la loro abbondanza relativa nello spazio e nel tempo; la *ricchezza di specie* è semplicemente il numero di specie in uno spazio definito ad un dato istante e l'*abbondanza relativa delle specie* si riferisce alla loro rarità o predominanza in termini di individui che compongono la specie stessa relativamente alla popolazione dell'intera comunità ecologica. Definiamo infine una *comunità ecologica* come un gruppo di specie simili a livello trofico che competono o potrebbero competere in un'area per le stesse o per simili risorse.

Solitamente, per motivi pratici, gli ecosistemi vengono monitorati a scale ridotte, ma spesso l'interesse ricade sulle proprietà di una comunità a scale diverse. Poiché non è possibile monitorare un intero paesaggio o un intero continente, tipicamente si studiano dei campioni, ma questi danno informazioni solo sulla biodiversità alla scala considerata che non si possono utilizzare direttamente per inferire la biodiversità alla scala globale, non godendo quest'ultima di proprietà additiva. Estrapolare la ricchezza delle specie da scala locale a globale non è dunque una cosa semplice e per riuscire in questo intento sono stati sviluppati molti metodi. Uno strumento statistico comunemente usato per descrivere la predominanza o rarità della presenza delle specie in una comunità ecologica è la **species abundance distribution** (**SAD**) cioè la descrizione delle abbondanze, ovvero il numero di individui, per ogni specie osservata all'interno di una comunità [4]. Precisamente definiamo la SAD come un vettore di abbondanze di tutte le specie presenti in una comunità ecologica. Spesso è rappresentata in un istogramma in cui l'asse x rappresenta il logaritmo in base 2 delle abbondanze e l'asse y la frequenza di specie con tale abbondanza [5]. Un'altra quantità studiata in ecologia è la **species area relationship** (**SAR**), una curva che descrive come cresce il numero di specie (diverse) al crescere dell'area campionata su cui l'ecosistema si estende.

Diamo ora una breve descrizione della **teoria neutrale**, che è il quadro teorico per la modellazione della dinamica delle popolazioni nel quale si svolge questo lavoro [6]. Con *neutrale* si intende che la teoria tratta gli organismi di una comunità come identici nella loro probabilità di nascita, morte, riproduzione, migrazione e speciazione. Si utilizza il termine neutrale quindi per descrivere l'ipotesi dell'equivalenza di tutti gli individui appartenenti alle specie di una data comunità ecologica. Notiamo che questa definizione di neutralità non esclude il fatto che gli individui possano interagire anche con processi ecologici complessi. Dunque la caratteristica che definisce una teoria neutrale in ecologia non è la semplicità delle regole che governano le interazioni tra gli individui ma piuttosto la completa identità delle interazioni stesse. Nonostante le sue drastiche ipotesi alla base, si è mostrato che questa teoria descrive molto bene le proprietà emergenti di comunità ecologiche di un dato livello trofico e che sono molto biodiverse [7].

Nei prossimi capitoli ci occuperemo di descrivere alcuni tra i metodi che sono stati sviluppati per dedurre la ricchezza delle specie a scala globale a partire da un campione ridotto di SAD di un dato ecosistema.

1 Modelli di campionamento e analisi

La ricchezza delle specie è la misura più intuitiva e più frequentemente usata per caratterizzare la diversità di un dato ecosistema.
La ricchezza di specie dipende però fortemente dal metodo di campionamento e dalla completezza del campione. Il modo tipico di censimento di una comunità ecologica porta ad avere principalmente due tipi di dati: dati di abbondanza e dati di incidenza [8] [9].
Per fissare la notazione: consideriamo una comunità costituita da N individui appartenenti a S specie distinte. Sia N_i il numero di individui della i-esima specie con i=1,2,...,S, $N_i > 0$ e $N = \sum_{n=1}^{S} N_i$. L'abbondanza relativa della specie i-esima è $p_i = N_i/N$, dunque $\sum_{n=1}^{S} p_i = 1$. Qui N, S, N_i e p_i rappresentano i valori veri, ma sconosciuti, dei parametri fondamentali dell'insieme in esame.

Dati di abbondanza

In molti studi biologici o ecologici solitamente gli individui vengono osservati in un dato momento e vengono classificati in base alla specie di appartenenza. Si prenda ad esempio un campione di n individui dall'insieme in esame e si ipotizzi di osservare un totale di S^* specie: questo è il *campione di riferimento*. Questo tipo di data-set può essere ottenuto usando due schemi di campionamento differenti:

1. *campionamento di tipo discreto* in cui l'unità campionaria è un individuo. Ad esempio, si campiona un numero fissato N^* di individui in una certa area. Qui la grandezza del campione n è fissata e ogni specie può essere rappresentata al massimo da N^* individui;

2. *campionamento di tipo continuo* nel quale il campione viene quantificato misurando su scale continue come tempo, area o volume d'acqua. Si fissa, per esempio, una certa area da studiare o un certo periodo di tempo nel quale analizzare il sito in esame. Con questo protocollo di campionamento il numero di individui osservati è una variabile casuale e ogni specie può essere rappresentata da un numero qualsiasi di individui.

Dati di incidenza

In alcune indagini le unità di campionamento non sono gli individui, ma porzioni di area o periodi di tempo: queste vengono campionate casualmente e indipendentemente. Ad esempio un'area di interesse può essere divisa in un certo numero di celle approssimativamente della stessa sotto-area e, una volta selezionate casualmente alcune di queste, l'indagine viene condotta solo su di esse.
A volte risulta impossibile contare esattamente il numero di individui per ogni specie che appaiono in un campione (ad esempio per microrganismi, invertebrati o piante) e quindi viene registrata solo la loro incidenza (presenza o assenza) nel campione. Dunque le stime si basano su degli insiemi di unità di campionamento in cui è registrata solo la presenza o assenza di una certa specie in un dato campione invece che la sua abbondanza.

Avendo a disposizione questo tipo di dati si possono seguire due approcci per stimare la diversità del campione: quello parametrico e quello non parametrico. In questo lavoro useremo dati di abbondanza ottenuti con campionamento continuo, infatti consideriamo una frazione a di un'area A nella quale sono stati registrati il numero di individui pre-

senti con relativa specie di appartenenza. D'ora in poi ci occuperemo solo di questo caso particolare.

Modelli parametrici e non parametrici

Negli approcci parametrici che analizzeremo si assume che la distribuzione dell'abbondanza delle specie abbia una certa forma funzionale, governata da dei parametri. Facendo il fit della curva dell'abbondanza relativa delle specie dei dati osservati si ottengono i valori dei parametri che, secondo le caratteristiche e le proprietà della distribuzione ipotizzata, permettono di dedurre le informazioni sulla diversità del sistema osservato.
Negli approcci non parametrici, invece, non si fanno assunzioni sulla distribuzione sottostante alla curva dell'abbondanza delle specie. L'intuizione e concetto base su cui si fondano i metodi di stima non parametrici è che le specie dominanti, cioè quelle a cui appartengono un elevato numero di individui, non danno alcuna informazione sulla ricchezza delle specie inosservate, mentre le specie rare, contengono quasi tutte le informazioni sulla biodiversità. Dunque, la maggior parte degli stimatori non parametrici si basa sulle frequenze di apparizione di basso ordine, specialmente sul numero di *singletons* e *doubletons*, cioè sul numero specie che vengono registrate contenere uno o due individui.

2 Derivazione delle distribuzioni binomiale negativa e logaritmica

Nel modello parametrico analizzato in questo lavoro, assumeremo che la SAD sia ben descritta da una binomiale negativa. Prima di entrare nel merito dei modelli per la stima del numero di specie, vediamo come si può ricavare tale distribuzione (come anche la distribuzione logaritmica, molto utilizzata in ecologia), modellando la dinamica stocastica dell'abbondanza delle specie attraverso la così detta *birth-death master equation* [7].
Entrambe le distribuzioni (logaritmica e binomiale negativa) possono essere derivate da processi ecologici fondamentali quali appunto nascita, morte e migrazione.

Sia $P_{n,s}(t)$ la probabilità che ad un certo tempo t, la specie s abbia esattamente n individui, con $s \in \{1, ..., S\}$. Assumiamo che la dinamica della popolazione di ogni specie $s = 1, 2, ..., S$ con popolazione n sia governata dai rates di nascita e di morte, rispettivamente, $b_{n,s}$ e $d_{n,s}$. L'equazione che regola l'evoluzione di $P_{n,s}(t)$ per $n \geq 0$ sarà quindi:

$$\frac{\partial P_{n,s}(t)}{\partial t} = P_{n\text{-}1,s}(t)\, b_{n\text{-}1,s} + P_{n+1,s}(t)\, d_{n+1,s} - P_{n,s}(t)\, b_{n,s} - P_{n,s}(t)\, d_{n,s}. \tag{2.1}$$

In questo lavoro, per evitare che n sia negativo, imponiamo delle condizioni al contorno riflettenti: $b_{\text{-}1,s} = d_{0,s} = 0$. Si trova che per $n > 0$ la soluzione stazionaria è:

$$P_{n,s} = P_{0,s} \prod_{i=0}^{n-1} \frac{b_{i,s}}{d_{i+1,s}} \tag{2.2}$$

dove il termine $P_{0,s}$ è il fattore di normalizzazione che può essere trovato imponendo la condizione $\sum_{n=1}^{\infty} P_{n,s} = 1$. Notiamo che la somma inizia da n=1 in quanto non si considerano specie con abbondanza nulla.

2.1 Distribuzione binomiale negativa

Ipotizziamo che tutti gli individui, siano essi appartenenti a specie rare o comuni, abbiano la stessa probabilità di morire, sopravvivere e riprodursi. In questo caso i tassi di nascita e morte pro capite non dipendono dal numero di individui n appartenenti alla specie. Definiamo quindi $b_{n,s}$ come:

$$b_{n,s} = b_s(n + r_s), \tag{2.3}$$

dove r_s è un parametro che tiene conto di eventi di immigrazione o di interazioni intraspecifiche.
Analogamente definiamo $d_{n,s}$:

$$d_{n,s} = d_s n. \tag{2.4}$$

Sostituendo questi ultimi termini nella (2.2) e denotando con $\xi_s = b_s/d_s$, si ottiene:

$$P_{n,s} = P_{0,s} \binom{n + r_s - 1}{n} \xi_s^n.$$

La costante di normalizzazione può essere calcolata imponendo:

$$1 = \sum_{n=1}^{\infty} P_{n,s} = P_{0,s} \sum_{n=1}^{\infty} \binom{n + r_s - 1}{n} \xi_s^n = P_{0,s}[1 - (1 - \xi_s)^{r_s}](1 - \xi_s)^{-r_s}.$$

Dunque la probabilità che una specie s abbia n individui all'equilibrio è data da una binomiale negativa di parametri (r_s, ξ_s) e normalizzata per abbondanze non nulle ($n \geq 1$):

$$P_{n,s}^{NB} = \frac{1}{1-(1-\xi_s)^{r_s}} \binom{n+r_s-1}{n} \xi_s^n (1-\xi_s)^{r_s}. \tag{2.5}$$

Sotto l'ipotesi della teoria neutrale secondo la quale le specie sono considerate demograficamente equivalenti, possiamo rimuovere l'indice s di specie dall'equazione sopra, ottenendo così una SAD per l'ecosistema in esame. Notiamo che quindi in questo framework, la dinamica della popolazione della specie è determinata puramente da processi demografici random (nascita, morte e migrazione) e ogni specie rappresenta una diversa realizzazione dello stesso processo stocastico.

2.2 Distribuzione logaritmica di Fisher

Notiamo che, scegliendo in modo diverso il termine $b_{n,s}$, si può ottenere sempre attraverso la *birth-death master equation* (2.1), un'altra importante distribuzione: la distribuzione logaritmica di Fisher. Consideriamo ora nella dinamica della comunità ecologica anche il fenomeno di speciazione (cioè l'entrata di nuove specie nel sistema dovute a mutazioni, invece che a migrazione da comunità esterne). La speciazione avverrà con un rate molto piccolo ν, e si avrà:

$$b_{n,s} = b_s n + \delta_{n,0}\nu. \tag{2.6}$$

Aggiungendo la condizione al contorno riflettente $b_{0,s} = \nu$ si ha quindi che il tasso di nascita tiene conto della riproduzione e della speciazione. In particolare, il parametro ν assicura che, se le specie si estinguono, la comunità rimane sempre popolata da un individuo. Dunque sostituendo la (2.4) e la (2.6) nella (2.2) e definendo $x_s = b_s/d_s$, si trova la seguente soluzione stazionaria:

$$P_{n,s} = P_{0,s} \frac{\nu}{b_s} \frac{x_s^n}{n}. \tag{2.7}$$

La costante di normalizzazione $P_{0,s}$ si determina imponendo:

$$1 = \sum_{n=1}^{\infty} P_{n,s} = P_{0,s} \frac{\nu}{b_s} \sum_{n=0}^{\infty} \frac{x_s^n}{n} = P_{0,s} \frac{\nu}{b_s} [-\log(1-x_s)].$$

Dunque abbiamo:

$$P_{n,s}^{LS} = -\frac{1}{\log(1-x_s)} \frac{x_s^n}{n}. \tag{2.8}$$

Anche in questo caso assumiamo che le specie siano equivalenti e possiamo dunque omettere l'indice s.

2.2.1 La distribuzione di Fisher come caso particolare della binomiale negativa

Osserviamo che la distribuzione binomiale negativa converge ad una distribuzione logaritmica nel limite di r che tende a zero:

$$\lim_{r\to 0} P_n^{NB} = \lim_{r\to 0} \frac{(1-\xi)^r}{1-(1-\xi)^r} \binom{n+r-1}{n} \xi^n = \frac{\xi^n}{-n\ln(1-\xi)}, \tag{2.9}$$

dove si è usato il fatto che:

$$\binom{n+r-1}{n} = \frac{\Gamma(n+r)}{\Gamma(n+1)\Gamma(r)} \approx \frac{r}{n},$$

per $r \approx 0$.
Notiamo dunque che la (2.9) coincide con la (2.8) ponendo $x = \xi$.

3 Metodi di upscaling

Con il termine *upscaling* si intende l'operazione matematica di inferire informazioni "a scala più grande" data la conoscenza del sistema ad una scala "più piccola". In questo contesto ecologico significa quindi predire il numero di specie presenti a scale più grandi di quelle che, per motivi pratici, possono essere indagate sperimentalmente solo a scala locale e di cui si hanno dunque dati empirici. Prendiamo ad esempio la foresta Amazzonica: la sua estensione è talmente vasta che è possibile campionare solamente una piccolissima percentuale di essa ($p^* = 0.00016\%$) [1].
In questa sezione vediamo in dettaglio come è possibile ricostruire la biodiversità di un intero ecosistema a partire da un campione ridotto di SAD, occupandoci del caso in cui ad essere studiata è una frazione dell'area totale del sistema in esame.
Analizzeremo prima due metodi parametrici, quello della binomiale negativa e della distribuzione logaritmica di Fisher [1], e poi un metodo non parametrico, quello dello stimatore di $Chao_{wor}$ [10].

Ipotesi sul metodo di upscaling

Nella nostra analisi assumiamo che la probabilità p che un individuo si trovi all'interno di una zona a contenuta in una regione A sia proporzionale all'area della zona stessa: $p{=}a/A$. Ci riferiamo a quest'ipotesi con il nome di *ipotesi di campo medio*. Una conseguenza di ciò è che campionare una frazione a di un'area A in cui ogni individuo viene catalogato in una lista in base alla specie di appartenenza è equivalente a campionare la stessa frazione p degli individui della lista. Questa è l'unica procedura imparziale che può essere adottata quando non si hanno informazioni né sulle posizioni degli individui né sulle correlazioni spaziali. Per ritenere quest'ipotesi soddisfatta bisogna verificare che la regione in esame non presenti forti disomogeneità e anisotropie, altrimenti alcune specie tenderebbero ad abitare habitat specifici all'interno della regione e dunque l'assunzione di avere una distribuzione spazialmente omogenea degli individui non sarebbe più valida.

3.1 Metodo della binomiale negativa

Quello che vorremmo ottenere attraverso l'*upscaling* date le informazioni sull'abbondanza delle specie alla scala a, è conoscere la forma della SAD ed il numero totale di specie presenti a scala globale, cioè in tutta l'area A dell'ecosistema in esame. Denotiamo con $P(n|1)$ la probabilità che una specie abbia esattamente n individui a scala globale (con il numero 1 si intende l'intero ecosistema ($p = 1$)). Questa probabilità coincide esattamente con l' *abbondanza relativa delle specie* (RSA), che è semplicemente la SAD normalizzata rispetto al numero totale di specie. Notiamo che P($n|1$) deve essere definita solamente per $n \geq 1$ poiché per poter osservare una data specie nell'intero ecosistema, questa deve essere popolata da almeno un individuo.
In questo metodo di *upscaling* si ipotizza che la SAD segua una distribuzione binomiale negativa $\mathcal{P}(n|r,\xi)$ di parametri (r,ξ):

$$P(n|1) = c(r,\xi)\mathcal{P}(n|r,\xi) \tag{3.1}$$

con

$$\mathcal{P}(n|r,\xi) = \binom{n+r-1}{n}\xi^n(1-\xi)^r \tag{3.2}$$

e

$$c(r,\xi) = \frac{1}{1-(1-\xi)^r} \tag{3.3}$$

dove c è la costante di normalizzazione. Quest'ultima è stata calcolata imponendo $\sum_{n=1}^{\infty} P(n|1) = 1$, dove appunto la normalizzazione è calcolata per popolazioni $n \geq 1$. Notiamo che invece $\mathcal{P}(n|r,\xi)$ è normalizzata per $n \geq 0$: questo perché all'interno dei campioni esiste una probabilità non nulla di che una specie presente nell'intero ecosistema abbia n=0 individui. Dunque in questo modo si tiene conto del numero di specie mancanti nei campioni.
Consideriamo ora un campione di area a e definiamo p=a/A la scala del campione, cioè la frazione di ecosistema osservato. Come primo passaggio calcoliamo la SAD del campione assumendo che quest'ultima non sia influenzata da correlazioni spaziali. Quest'ipotesi è ben soddisfatta ed è stata verificata usando dati di foreste generati *in silico* a vari gradi di correlazione spaziale [1].
Sotto queste ipotesi la probabilità che una specie presenti k individui in un'area a=pA, condizionata dal fatto che presenta n individui nella regione totale A è data dalla distribuzione binomiale:

$$\mathcal{P}_{binom}(k|n,p) = \begin{cases} \binom{n}{k} p^k (1-p)^{n\text{-}k} & \text{se } k = 0, ..., n \\ 0 & \text{se } k{>}n \end{cases} \tag{3.4}$$

Infatti, in assenza di correlazioni spaziali, la probabilità che uno degli individui di una specie si trovi in una regione di area a è esattamente p.

Mostriamo ora un risultato chiave per il metodo di upscaling.

3.1.1 Proprietà di invarianza per forma della distribuzione binomiale negativa

Sia $P(n|1) = c(r,\xi)\mathcal{P}(n|r,\xi)$ la SAD dell'ecosistema a scala globale e denotiamo con $\mathcal{P}(k|n,p)$ la probabilità che una specie abbia abbondanza k alla scala $p\in(0,1)$ condizionata dal fatto che alla scala globale A sono presenti n individui di quella specie. Se $\mathcal{P}(k|n,p) = \mathcal{P}_{binom}(k|n,p)$ segue una distribuzione binomiale, allora la SAD $\mathcal{P}_{sub}(k|p)$ alla scala di campionamento p è ancora una binomiale negativa per $k \geq 1$ con il parametro ξ riscalato e lo stesso r:

$$\mathcal{P}_{sub}(k|p) = \begin{cases} c(r,\xi)\mathcal{P}(k|r,\hat{\xi}_p), & k \geq 1 \\ 1 - c(r,\xi)/c(r,\hat{\xi}_p), & k{=}0 \end{cases} \tag{3.5}$$

con

$$\hat{\xi}_p = \frac{p\xi}{1-\xi(1-p)}. \tag{3.6}$$

Infatti la probabilità $\mathcal{P}_{sub}(k|p)$ di trovare una specie popolata da $k \geq 0$ individui nel campione di area $a{=}pA$ è:

$$\begin{aligned} k \geq 1 : \mathcal{P}_{sub}(k|p) &= \sum_{n \geq k} \mathcal{P}_{binom}(k|n,p) P(n|1) \\ &= \sum_{n \geq k} \binom{n}{k} p^k (1-p)^{n-k} \cdot c(r,\xi) \binom{n+r-1}{n} \xi^n (1-\xi)^r \\ &= c(r,\xi) \binom{k+r-1}{k} \left(\frac{p\xi}{1-\xi(1-p)} \right)^k \left(\frac{1-\xi}{1-\xi(1-p)} \right)^r \\ &= c(r,\xi) \binom{k+r-1}{k} \hat{\xi}_p^k (1-\hat{\xi}_p)^r = c(r,\xi) \cdot P(k|r,\hat{\xi}_p) \end{aligned} \tag{3.7}$$

dove abbiamo usato la (3.6) per $\hat{\xi}_p$ per la penultima uguaglianza, e

$$\begin{aligned} k = 0 : \mathcal{P}_{sub}(0|p) &= 1 - \sum_{n \geq 1} \mathcal{P}_{sub}(k|p) \\ &= 1 - \sum_{n=1}^{\infty} c(r,\xi) \cdot \binom{k+r-1}{k} \hat{\xi}_p^k (1-\hat{\xi}_p)^r \\ &= 1 - c(r,\xi) \cdot \sum_{k=1}^{\infty} \mathcal{P}(k|r,\hat{\xi}_p) = 1 - \frac{c(r,\xi)}{c(r,\hat{\xi}_p)}. \end{aligned} \tag{3.8}$$

3.1.2 Il numero di specie a scala globale

Vogliamo ora inferire la biodiversità a scala globale, conoscendo solamente le informazioni sulle specie ottenute da un campione ad una certa scala p^*, ovvero le abbondanze delle $S^* \leq S$ specie presenti nel campione esaminato. Denotando il numero di specie di abbondanza k alla scala p^* con $S^*(k)$, otteniamo, per $k \geq 1$:

$$\frac{S^*(k)}{S^*} \equiv P(k|p^*) = \frac{\mathcal{P}_{sub}(k|p^*)}{\sum_{k' \geq 1} \mathcal{P}_{sub}(k'|p^*)} = \frac{\mathcal{P}(k|r,\hat{\xi}_{p^*})}{\sum_{k' \geq 1} \mathcal{P}(k'|r,\hat{\xi}_{p^*})} = c(r,\hat{\xi}_{p^*}) \mathcal{P}(k|r,\hat{\xi}_{p^*}) \tag{3.9}$$

che, dalla (3.1), è una binomiale negativa normalizzata per $k \geq 1$, mentre $\mathcal{P}(k|r,\hat{\xi}_{p^*})$ è normalizzata per $k \geq 0$. Per quanto detto sopra otteniamo dunque il seguente risultato: partendo da una distribuzione binomiale negativa per la SAD a scala globale, anche la SAD a scala ridotta risulta distribuita secondo una binomiale negativa di parametri lo stesso r e $\hat{\xi}_p^*$ riscalato. Una SAD avente la proprietà di avere la stessa forma funzionale a scale differenti è detta essere *invariante per forma*.

Fittando la SAD dei dati alla scala p^* possiamo dunque trovare i parametri r e $\hat{\xi}_p^*$ e, invertendo l'equazione (3.6), troviamo:

$$\xi = \frac{\hat{\xi}_{p^*}}{p^* + \hat{\xi}_{p^*}(1-p^*)}. \tag{3.10}$$

Usando ancora la (3.6) per eliminare ξ dall'ultima equazione, otteniamo la seguente relazione per il parametro ξ alle due scale p e p^*:

$$\hat{\xi}_p = \frac{p\hat{\xi}_{p^*}}{p^* + \hat{\xi}_{p^*}(p-p^*)} \equiv U(p,p^*|\hat{\xi}_{p^*}) \tag{3.11}$$

dalla quale, ovviamente, è possibile riottenere sia la (3.6) che la (3.10) ponendo $\xi \equiv \hat{\xi}_{p=1}$.

Vogliamo ora determinare la relazione tra il numero totale di specie S alla scala globale p=1 e il numero totale di specie osservate localmente S_p alla scala p. D'ora in avanti per denotare il numero di specie alla scala locale useremo la notazione $S^* \equiv S_{p^*}$. Notiamo che:

$$\mathcal{P}_{sub}(k{=}0|p^*) = \frac{S - S^*}{S} \tag{3.12}$$

$$\mathcal{P}_{sub}(k|p^*) = \frac{S^*(k)}{S}. \tag{3.13}$$

Usando la seconda delle (3.5), il numero di specie presenti nell'intera comunità ecologica è dato, in termini dei dati del campione osservato, da:

$$S = \frac{S^*}{1 - \mathcal{P}_{sub}(k=0|p^*)} = S^* \frac{1-(1-\xi)^r}{1-(1-\hat{\xi}_p^*)^r}. \tag{3.14}$$

Notiamo che, se si assume che la SAD segua una distribuzione binomiale negativa a scala globale, il valor medio dell'abbondanza totale riscala linearmente con l'area, infatti:

$$\begin{aligned} \mathbb{E}(N^*) &= \sum_{k=1}^{\infty} kSc(r,\xi)\binom{k+r-1}{k}\hat{\xi}_{p^*}^k(1-\hat{\xi}_{p^*})^r = Sc(r,\xi)r\frac{\hat{\xi}_{p^*}}{1-\hat{\xi}_{p^*}} \\ &= Sc(r,\xi)r\frac{p\xi}{1-\xi} = p\mathbb{E}(N) \end{aligned} \tag{3.15}$$

3.2 Metodo della distribuzione logaritmica

Confronteremo i risultati ottenuti con il metodo di upscaling della binomiale negativa, con quelli derivati attraverso il modello più classico della distribuzione logaritmica, che ricordiamo essere un caso particolare di quello appena presentato. Il modello della distribuzione logaritmica per la SAD nacque nei primi anni '40 quando il chimico e naturalista britannico Alexander Steven Corbet, dopo aver trascorso due anni in Malesia a studiare e catalogare le specie di farfalle, tornò in Inghilterra e mostrò i suoi dati al collega Ronald Aylmer Fisher. Corbet si chiese quante nuove specie avrebbe trovato se fosse tornato in Malesia per altri due anni. Per rispondere a questa domanda, Fisher, da molti considerato come il padre della statistica, affrontò il problema della stima del numero di specie, che da quel momento ha trovato larghe applicazioni in vari campi scientifici. Dunque in questo contesto Fisher introdusse la distribuzione che da lui prende nome e che viene molto utilizzata in teoria dell'ecologia per descrivere la SAD di un ecosistema, ovvero la distribuzione logaritmica di parametro x [11]:

$$P(n|1) = P^{LS}(n|x) = \alpha(x)\frac{x^n}{n},\ \alpha(x) = -(\log(1-x))^{-1}, \tag{3.16}$$

dove $\alpha(x)$ è la costante di normalizzazione. Assumendo anche questa volta che la SAD del campione non sia influenzata da correlazioni spaziali, anche la distribuzione logaritmica, essendo un caso particolare della distribuzione binomiale negativa, soddisfa la proprietà di invarianza.

3.2.1 Proprietà di invarianza per forma della distribuzione logaritmica

Sia $P(n|1) = \alpha(x)\mathcal{P}^{LS}(n|x)$ la SAD alla scala globale e denotiamo con $\mathcal{P}(k|n,p)$ la probabilità che una specie abbia abbondanza k nel campione alla scala $p \in (0,1)$ condizionata dal fatto alla scala globale A la specie possiede n individui.

Se $\mathcal{P}(k|n,p) = \mathcal{P}_{binom}(k|n,p)$ è distribuita secondo una binomiale, allora la SAD alla scala del campione, $\mathcal{P}^{LS}_{sub}(k|p)$, è ancora una distribuzione logaritmica per $k \geq 1$ con il parametro x riscalato:

$$\mathcal{P}^{LS}_{sub}(k|p) = \begin{cases} \alpha(x)\mathcal{P}^{LS}(k|\hat{x}_p) & k \ \geq 1 \\ 1 - \alpha(x)/\alpha(\hat{x}_p) & k{=}0 \end{cases} \tag{3.17}$$

con

$$\hat{x}_p = \frac{px}{1 - x(1-p)}. \tag{3.18}$$

Infatti la probabilità $\mathcal{P}^{LS}_{sub}(k|p)$ di trovare una specie con popolazione $k \geq 0$ nel sotto campione di area $a{=}pA$ è:

$$\begin{aligned} k \geq 1 : \mathcal{P}^{LS}_{sub}(k|p) &= \sum_{n \geq k} P_{binom}(k|n,p) P(n|1) \\ &= \sum_{k \geq n} \binom{n}{k} p^k (1-p)^{n\text{-}k} \cdot \alpha(x) \frac{x^n}{n} \\ &= \alpha(x) \left(\frac{px}{1 - x(1-p)} \right)^k \frac{1}{k} \\ &= \alpha(x) \frac{\hat{x}_p^k}{k} = \alpha(x) \cdot \mathcal{P}^{LS}(k|\hat{x}_p) \end{aligned} \tag{3.19}$$

dove abbiamo usato la relazione (3.18) nella penultima uguaglianza,e

$$\begin{aligned} k = 0 : \mathcal{P}^{LS}_{sub}(0|p) &= 1 - \sum_{k \geq 1} P^{LS}_{sub}(k|p) \\ &= 1 - \sum_{k=1}^{\infty} \alpha(x) \cdot \frac{\hat{x}_p^k}{k} \\ &= 1 - \alpha(x) \cdot \sum_{k=1}^{\infty} \mathcal{P}^{LS}(k|\hat{x}_p) = 1 - \frac{\alpha(x)}{\alpha(\hat{x}_p)}. \end{aligned} \tag{3.20}$$

Notiamo che (3.18) è analoga a (3.6). Dunque l'analogo di (3.10) è

$$x = \frac{\hat{x}^*_p}{p^* + \hat{x}^*_p(1-p^*)} \tag{3.21}$$

e l'equazione (3.11) vale anche in questo caso.
La SAD può essere ottenuta come nell'equazione (3.9) ed è data da:

$$P(k|p^*) = \frac{P^{LS}_{sub}(k|p^*)}{\sum_{k' \geq 1} P^{LS}_{sub}(k'|p^*)} = \alpha(\hat{x}^*_p) \frac{\hat{x}^k_{p^*}}{k} = \alpha(\hat{x}^*_p) P^{LS}(n|\hat{x}^*_p) \tag{3.22}$$

.

3.2.2 Il numero di specie a scala globale

Il numero di specie con popolazione $k \geq 1$ presenti nel campione di area $a{=}pA$ è dato da:

$$S_p(k) \equiv S P^{LS}_{sub}(k|p) = S\alpha(x) \frac{\hat{x}^k_p}{k} = \hat{\alpha} \frac{\hat{x}^k_p}{k} \tag{3.23}$$

dove abbiamo unito le costanti S e $\alpha(x)$ in un unico termine $\hat{\alpha}$ che non dipende dalla scala p del campione. Quando ci riferiremo alla scala p^* useremo, per brevità di notazione, $S^*(k) \equiv S_{p^*}(k)$.
Allora il numero totale di specie S^* e l'abbondanza totale N^* alla scala p^* sono date rispettivamente da:

$$S^* = \sum_{k=1}^{\infty} S^*(k) = -\hat{\alpha}\log(1-\hat{x}_{p^*}) \tag{3.24}$$

$$N^* = \sum_{k=1}^{\infty} kS^*(k) = \hat{\alpha}\frac{\hat{x}_{p^*}}{1-\hat{x}_{p^*}}. \tag{3.25}$$

Poiché S^* e N^* sono note dal campione, possiamo trovare $\hat{\alpha}$ risolvendo la seguente equazione:

$$N^* - \hat{\alpha}(\exp(\frac{S^*}{\hat{\alpha}}) - 1) = 0, \tag{3.26}$$

che si ottiene inserendo l'espressione di $\hat{x}_{p^*}$ da (3.24) nella (3.25).

Vogliamo ora dedurre le informazioni a scala globale p=1 dai dati disponibili alla scala p=p^*. Dalle considerazioni precedenti sappiamo che $\hat{\alpha}$ è un parametro indipendente dalla scala, dunque abbiamo le seguenti relazioni per S e N:

$$S = -\hat{\alpha}\log(1-x) \tag{3.27}$$

$$N = \hat{\alpha}\frac{x}{1-x}. \tag{3.28}$$

dalle quali otteniamo:

$$S = \hat{\alpha}\log\left(1+\frac{N}{\hat{\alpha}}\right),\ \hat{\alpha} = S\alpha(x). \tag{3.29}$$

Dunque per dedurre la biodiversità a scala globale, S, è necessaria una stima dell'abbondanza totale N. Prendiamo $N = N^*/p^*$. Notiamo che questo è consistente con il nostro quadro teorico nel quale assumiamo che la SAD sia *invariante per forma*: infatti si può dimostrare che, se si assume che la SAD segua una distribuzione di Fisher a scala globale, il valor medio dell'abbondanza totale riscala linearmente con l'area:

$$\mathbb{E}(N^*) = \sum_{k=1}^{\infty} kS^*(k) = \sum_{k=1}^{\infty} k\hat{\alpha}\frac{\hat{x}_{p^*}^k}{k} = \alpha\frac{\hat{x}_{p^*}}{1-\hat{x}_{p^*}} = \hat{\alpha}\frac{px}{1-x} = p^*\mathbb{E}(N), \tag{3.30}$$

dove abbiamo usato la (3.18).

Per dedurre in modo alternativo la biodiversità a scala globale, analogamente a quanto fatto per il metodo della binomiale negativa, si potrebbe usare anche la relazione seguente:

$$S = \frac{S^*}{1-P_{sub}^{LS}(k{=}0|p^*)} = S^*\frac{\log(1-x)}{\log(1-\hat{x}_{p^*})}. \tag{3.31}$$

In questo caso non bisogna avere una stima del numero totale di individui N nell'area A. È stato verificato su dati riguardanti la biodiversità nelle foreste che i due metodi restituiscono previsioni equivalenti [1].

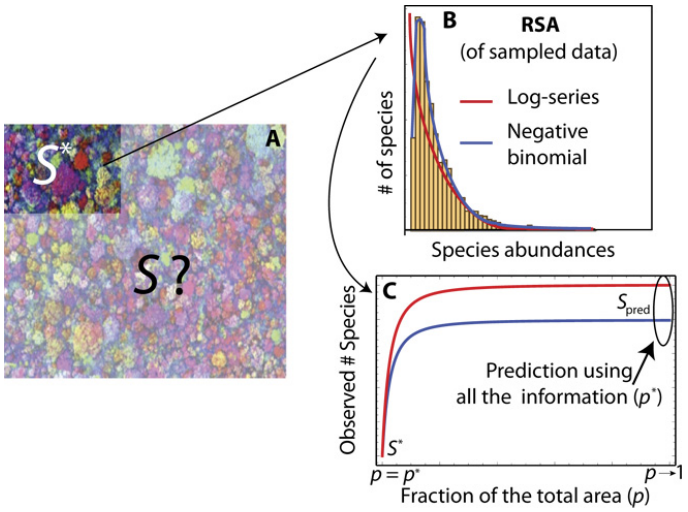

Figura 3.1: **Rappresentazione schematica dei modelli parametrici di upscaling**. Questi consistono in tre passaggi. (**A**)Campionamento: conosciamo l'abbondanza di S^* specie alla scala di campionamento p^*. (**B**)Fit: facciamo un fit della SAD con una binomiale negativa o una distribuzione logaritmica. (**C**)Upscaling: usando i parametri del miglior fit ottenuti in (B) e usando le equazioni (3.29) e (3.14) deduciamo la biodiversità dell'intero ecosistema. [1]

3.3 Metodo $Chao_{wor}$

Introduciamo ora il metodo non parametrico sviluppato da Chao, nato nell'ambito di uno schema di campionamento senza reinserimento. Il pedice *wor* sta infatti per "*without repetition*". Questo è il sistema di indagine più usato quando si devono campionare individui che non si desidera osservare ripetutamente (ad esempio nello studio delle foreste, nel quale gli alberi vengono censiti in piccole aree che sono selezionate senza ripetizione). In questo schema di campionamento ogni individuo o ogni unità di campionamento possono essere indagati una sola volta.
Assumiamo che in un ecosistema ci siano S specie indicizzate da 1 a S. Sia N_i (abbondanza assoluta) il numero di individui appartenenti alla i-esima specie, i=1,...,S, e $N_i > 0$. La popolazione totale dunque è data da $N = \sum_{i=1}^{S} N_i$.
Supponiamo di prendere dall'intero ecosistema un campione di N^* individui, campionandoli senza reinserimento. Sia X_i la frequenza della specie campionata cioè il numero di individui della i-esima specie osservati nel campione. Solo le specie con $X_i > 0$ sono osservabili nel campione. Sia S_k^* il numero di specie nel campione che sono rappresentate esattamente da k individui, dunque S_0^* denota il numero di specie che non sono state osservate nel campione. Dunque abbiamo che $N^* = \sum_{i=1}^{S^*} X_i = \sum_{k>1} kS_k^*$. Definiamo $p^* = N^*/N$ la frazione di campionamento e S^* il numero di specie osservate nel campione, $S^* = \sum_{k>1} S_k^*$.
Generalmente la probabilità che una specie venga rilevata, o rate di rilevamento, dipende sia dall'abbondanza della specie nel campione sia da caratteristiche specifiche degli individui come ad esempio il modo di spostarsi e muoversi all'interno dell'ambiente, colore, forma e dimensione.
Consideriamo dunque il caso generale in cui la probabilità di trovare un individuo possa variare a seconda della specie di appartenenza e indichiamola con $\theta_i > 0$ per la i-esima specie. Sotto queste ipotesi, definendo $p_i = N_i/N$ come l'abbondanza relativa, il rate di rilevamento per la i-esima specie diventa $\psi_i = \frac{N_i\theta_i}{\sum_{k=1}^{S} N_k\theta_k} = \frac{p_i\theta_i}{\sum_{k=1}^{S} p_k\theta_k}$ con $i = 1, ..., S$. Intuitivamente, il numero di individui che hanno la stessa possibilità di essere osservati è dato

da $N_i\psi_i$, ma poiché questo potrebbe essere un numero non intero, definiamo una variabile a valori interi Z_i, che rappresenta il numero di individui che hanno la stessa possibilità di essere osservati per la i-esima specie. Siccome $Z \geq 1$ e la frazione di individui campionata è N^*/N, si può modellare il vettore $\mathbf{Z}$=$(Z_1, Z_2, ..., Z_S)$ attraverso una distribuzione multinomiale troncata di parametri N e $(\psi_1^*, \psi_2^*, ..., \psi_S^*)$, dove $\psi_i^* = \psi_i / P\{z : z_i \geq 1, i = 1, ..., S\}$, $\mathbf{z}$=$(z_1, z_2, ..., z_S)$ e $\sum_{i=1}^{S} z_i = N$. Per ogni dato valore di $\mathbf{z}$=$(z_1, z_2, ..., z_S)$, le frequenze con cui appaiono gli individui della specie i-esima nel campione, $(X_1, X_2, ..., X_S)$, seguono una distribuzione ipergeometrica generalizzata:

$$P(X_i = x_i, i = 1, 2, ..., S) = \binom{z_1}{x_1}\binom{z_2}{x_2}...\binom{z_S}{x_S}/\binom{N}{N^*} \tag{3.32}$$

$$z_i \geq 1, \sum_{i=1}^{S} z_i = N.$$

Sulla base di questo modello generale, la distribuzione marginale per ognuna delle frequenze con le quali vengono individuate le specie è una distribuzione ipergeometrica:

$$P(X_i = x_i) = \binom{z_i}{k}\binom{N - z_i}{N^* - k}/\binom{N}{N^*}. \tag{3.33}$$

3.3.1 Il numero di specie a scala globale

Vediamo dunque com'è possibile dedurre, sotto queste ipotesi, il numero di specie a scala globale a partire da un vettore di abbondanze ottenuto esaminando una frazione dell'intero ecosistema. Per semplicità poniamo $n = N^*$.
Il valore di aspettazione per i contatori di frequenze S_k^* usando la (3.33) è:

$$\mathbb{E}(S_k^*) = \sum_{i}^{S} P(X_i = k) = \sum_{i=1}^{S} \binom{z_i}{k}\binom{N - z_i}{n - k}/\binom{N}{n} \tag{3.34}$$

In particolare:

$$\mathbb{E}(S_0^*) = \sum_{i=1}^{S} \binom{N - z_i}{n}/\binom{N}{n}$$

$$\mathbb{E}(S_1^*) = \sum_{i=1}^{S} \binom{z_i}{1}\binom{N - z_i}{n - 1}/\binom{N}{n} = \sum_{i=1}^{S} \frac{nz_i}{N - z_i - n + 1}\binom{N - z_i}{n}/\binom{N}{n}$$

$$\mathbb{E}(S_2^*) = \sum_{i=1}^{S} \binom{z_i}{2}\binom{N - z_i}{n - 2}/\binom{N}{n} = \sum_{i=1}^{S} \frac{n(n-1)z_i(z_i-1)}{2(N - z_i - n + 1)(N - z_i - n + 2)}\binom{N - z_i}{n}/\binom{N}{n}$$

Per la disuguaglianza di Cauchy-Schwarz si ha:

$$\left\{\sum_{i=1}^{S} \frac{nz_i}{N - z_i - n + 1}\binom{N - z_i}{n}/\binom{N}{n}\right\}^2 \leq$$

$$\left\{\sum_{i=1}^{S} \binom{N - z_i}{n}/\binom{N}{n}\right\} \times \left\{\sum_{i=1}^{S} \left(\frac{nz_i}{N - z_i - n + 1}\right)^2 \binom{N - z_i}{n}/\binom{N}{n}\right\},$$

dove vale il segno di uguaglianza quando tutte le z_i sono uguali.

La parte sinistra della disuguaglianza è $\{\mathbb{E}(S_1^*)\}^2$ e la prima sommatoria della parte destra è $\{\mathbb{E}(S_0^*)\}$. Per quanto riguarda la seconda sommatoria di destra riscrivendo:

$$\left(\frac{nz_i}{N-z_i-n+1}\right)^2 = \frac{n}{n-1}\left(\frac{n(n-1)z_i(z_i-1)}{(N-z_i-n+1)^2}\right) + \frac{n^2 z_i}{(N-z_i-n+1)^2}$$

essa diventa:

$$\left\{\sum_{i=1}^{S}\left(\frac{nz_i}{N-z_i-n+1}\right)^2 \binom{N-z_i}{n}/\binom{N}{n}\right\} \approx \frac{2n}{n-1}\mathbb{E}(S_2^*)$$
$$+\sum_{i=1}^{S}\left[\frac{n}{N-z_i-n+1}\right]\frac{nz_i}{N-z_i-n+1}\binom{N-z_i}{n}/\binom{N}{n}.$$

Il contributo delle specie con z_i grande all'ultimo termine dell'equazione sopra è trascurabile, mentre per le specie con z_i molto più piccolo di N, abbiamo:

$$\frac{n}{N-z_i-n+1} = \frac{n/N}{(N-z_i-n+1)/N} \approx \frac{n/N}{1-n/N} = \frac{p^*}{1-p^*}.$$

Quindi otteniamo la seguente disuguaglianza:

$$\{\mathbb{E}(S_1^*)\}^2 \leq \{\mathbb{E}(S_0^*)\}\left\{\frac{n}{n-1}2\mathbb{E}(S_2^*) + \frac{p^*}{1-p^*}\mathbb{E}(S_1^*)\right\},$$

che è equivalente a:

$$\mathbb{E}(S_0^*) \geq \frac{\mathbb{E}(S_1^{*2})}{\frac{n}{n-1}2\mathbb{E}(S_2^*) + \frac{p^*}{1-p^*}\mathbb{E}(S_1^*)}. \tag{3.35}$$

Sostituendo il valore di aspettazione con le frequenze osservate otteniamo come limite inferiore per la ricchezza delle specie:

$$S_{p=1} = S^* + \frac{S_1^{*2}}{\frac{n}{n-1}2S_2^* + \frac{p^*}{1-p^*}S_1^*}. \tag{3.36}$$

4 Applicazione della teoria ecologica alle comunità microbiche

Gli ecosistemi di comunità microbiche presenti nel corpo umano giocano un ruolo molto importante per la nostra salute [12]. Ogni individuo può essere visto come un insieme di habitat occupati da comunità microbiche formatesi attraverso i processi fondamentali dell'ecologia: diffusione, diversificazione locale, selezione ambientale e migrazione. I tanti e svariati membri delle comunità hanno un ruolo cruciale nel mantenimento della salute umana liberando essi nutrienti ed energia altrimenti inaccessibili, favorendo la differenziazione dei tessuti, stimolando il sistema immunitario e proteggendo l'ospite dall'invasione da parte di agenti patogeni. Un certo numero di disturbi clinici, come l'obesità, la malnutrizione e malattie infiammatorie, sono stati associati all'alterazione della composizione delle comunità microbiche presenti nell'ospite.
Il corpo umano, dunque, può essere visto come un ecosistema e la salute di un individuo può essere associata ai servizi forniti all'organismo dalle comunità microbiche.
Recenti scoperte di variazioni inaspettate nella composizione del microbioma di individui sani hanno evidenziato l'importanza di identificare i processi che possano dare origine ad un tale cambiamento: la teoria dell'ecologia cerca di spiegare e predire questi fenomeni. Inoltre, il modello ecologico trasportato nel mondo delle pratiche cliniche può portare ad un miglioramento delle cure fornite ai pazienti: infatti, una visione completa della comunità che si va ad alterare con una certa terapia e non focalizzata solamente sul batterio a cui è dovuto il disturbo, può portare ad un nuovo approccio clinico che nella cura di una malattia tiene conto dell'intero microbioma dell'individuo.

4.1 Sequenziamento del DNA degli individui in comunità microbiche

Ottenere dati di biodiversità per una comunità microbica non è una cosa semplice: dopo averne prelevato una campione, per riconoscere le specie presenti al suo interno è necessario sequenziare il DNA in esso contenuto.
Lo sviluppo di tecniche di sequenziamento di nuova generazione (NGS, *next-generation sequencing*) ha portato ad un incremento delle risorse impegnate in questo tipo di ricerca, aumentando rapidamente le nostre conoscenze sulla composizione e sulle funzioni delle popolazioni batteriche in diversi ambienti [2]. Nel contesto clinico, il microbioma dell'intestino umano è stato soggetto ad indagini sofisticate che hanno rivelato una forte interazione tra i microrganismi, il sistema immunitario e il metabolismo. Una ridotta biodiversità o uno squilibrio tra le popolazioni di specie batteriche all'interno comunità microbiche dell'intestino umano sono state associate ad una serie di fenotipi come l'obesità, malattie infiammatorie dell'intestino, diabete di tipo II e numerosi altri disturbi.
La maggior parte degli studi riguardanti la comprensione delle dinamiche che governano le popolazioni batteriche sono stati condotti attraverso i cosiddetti approcci metagenomici, che studiano cioè l'insieme dei diversi materiali genetici, in modo semplice ed efficace in termini di costi. I principali metodi di sequenziamento del DNA sono il metodo *shotgun* e il metodo $16S$. Il metodo *shotgun* è una tecnica sperimentale di sequenziamento dell'intero genoma di un organismo [13]. Poiché a causa dell'elevata lunghezza della sequenza genetica è impossibile sequenziare il genoma in un unico passaggio, esso consiste nella creazione di numerosi piccoli frammenti di DNA che vengono clonati e sequenziati separatamente da entrambi i versi. Questi poi vengono riassemblati *in silico* attraverso criteri di compatibilità e sovrapposizione, in modo da ottenere una lunga sequenza continua. Con il metodo

$16S$ invece si va a sequenziare il gene ribosomale $16S$ che da una parte, essendo contenuto in una regione molto conservata, aiuta l'amplificazione, e dall'altra parte, differendo da una specie all'altra, ne permette la classificazione.
Le sequenze così ottenute vengono poi confrontate con quelle presenti nei database che contengono le informazioni sui metagenomi dei batteri finora sequenziati. In particolare il progetto microbioma umano (HMP, *human microbiome project*) contiene una vasta raccolta di sequenze di microorganismi associati al corpo umano, inclusi eucarioti, batteri, archei e virus, ottenute sia con il metodo *shotgun* che con il sequenziamento $16S$. Attingendo a queste informazioni è dunque possibile ottenere informazioni su quali specie batteriche siano presenti all'interno di campioni di interesse e con quali abbondanze.

4.2 Applicazione dei metodi di *upscaling*

Dato un campione di N individui (cioè, nel caso di dati metagenomici, di N sequenze di DNA, ovvero i cosiddetti *reads*), grazie ai metodi di classificazione tassonomica sopra citati, è possibile assegnare la specie solamente ad una frazione di questi; le restanti sequenze vengono scartate perché nei database di riferimento non risultano esserci batteri con tali stringhe di DNA. Analogamente a quanto viene fatto in ecologia, utilizzando i metodi *upscaling* si potrebbe stimare il numero di specie e la loro abbondanza tra questi batteri non identificati, attraverso i dati tassonomici delle specie riconosciute con successo (in una frazione p^*). La frazione p^* può essere quindi stimata come il rapporto tra il numero di sequenze che hanno trovato riscontro nel database, N^*, e il numero di sequenze inizialmente presenti nel campione, N, i.e., $p^* = N^*/N$. Le specie assenti nel campione di riferimento in ecologia corrispondono quindi, in questo contesto, alle specie a cui appartengono le sequenze che non riescono ad essere classificate.

Per questo lavoro sono stati ottenuti, utilizzando il software Kaiju [14], due vettori di abbondanze batteriche a partire da campioni sequenziati con metodo *shotgun*. In particolare un campione riguarda il microbioma un individuo sano mentre l'altro riguarda quello di un individuo affetto dal morbo di Crohn, una malattia dell'intestino. I dati utilizzati sono stati presi dallo studio svolto nell'articolo "Characterization of the gut microbiome using 16S or shotgun metagenomics" [2].

4.2.1 Test

Per sondare l'efficienza nell'ambito delle comunità microbiche dei metodi di *upscaling* precedentemente descritti sono stati condotti dei test su ognuno dei due campioni, procedendo in questo modo:

- per ogni campione sono stati selezionati 100 sottocampioni in modo casuale, ognuno contenete l'1% della popolazione totale;
- ad ogni sottocampione sono stati applicati i metodi di *upscaling*, i due parametrici della binomiale negativa e della distribuzione logaritmica e il metodo non parametrico $Chao_{wor}$, con con $p^* = 0.01$;
- sono stati predetti i numeri di specie alla scala globale (quella del campione) e confrontati con quelli reali che conosciamo a tale scala.

Analizzando i risultati dei test abbiamo notato i seguenti fatti:

- il metodo $Chao_{wor}$ è inapplicabile. Questo infatti si basa sul conteggio del numero di specie rare, popolate cioè da uno o due individui. Nel caso di comunità microbiche

non vengono identificate specie con questo tipo di caratteristiche alla sotto-scala analizzata e dunque il metodo non fornisce alcun risultato;

- il metodo parametrico della distribuzione logaritmica di Fisher sovrastima il numero di specie, in particolare predice circa il doppio del numero di specie realmente presenti nel campione iniziale;

- il metodo parametrico della binomiale negativa fornisce in tutti i casi analizzati una stima corretta del numero di specie presenti nel campione.

Sono stati calcolati gli errori percentuali sulle specie predette alla scala globale da ognuno dei 100 sottocampioni. Per il metodo della binomiale negativa questi sono dell'ordine dello 0.01% e del 2.5% rispettivamente per l'individuo sano e per l'individuo malato. Il metodo della distribuzione logaritmica produce invece degli errori significativamente grandi: poiché il numero di individui e il numero di specie sono uguali per ogni sottocampione, per ognuna delle 100 prove questo metodo predice, attraverso le equazioni (3.26) e (3.29), lo stesso numero di specie e dunque gli errori percentuali sono uguali. Questi risultano dell' 85% per l'individuo sano e dell' 83% per l'individuo malato.

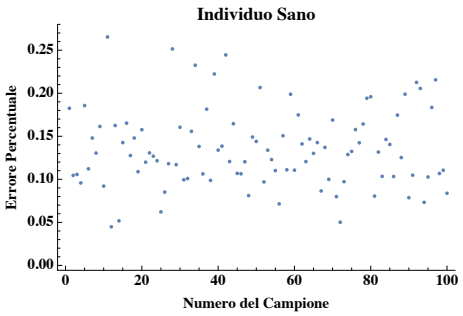

Figura 4.1: Errore percentuale sulle specie predette da campioni dell'1% con il metodo della binomiale negativa per un individuo sano.

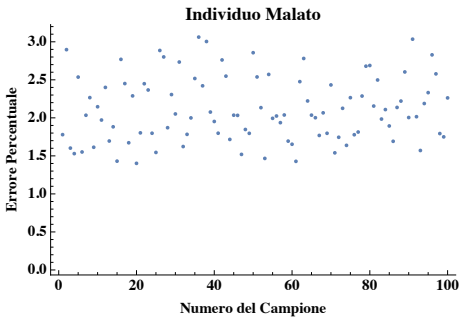

Figura 4.2: Errore percentuale sulle specie predette da campioni dell'1% con il metodo della binomiale negativa per un individuo malato.

Inoltre sono state calcolate le RSA predette a scala globale a partire dai parametri stimati alla scala dei sottocampioni. Si nota che i punti ottenuti per la binomiale negativa riproducono l'andamento della RSA a scala globale, mentre quelli ottenuti per la distribuzione logaritmica non ne predicono il picco.

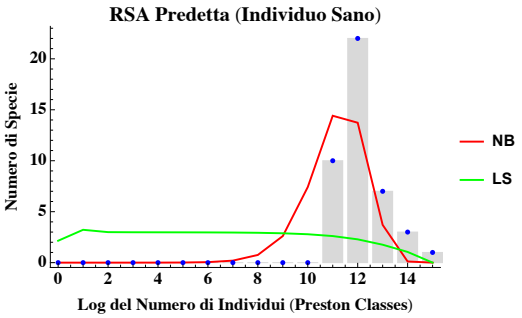

Figura 4.3: RSA predetta per individuo sano. Le previsioni della binomiale negativa riproducono l'andamento reale mentre quelle della distribuzione logaritmica falliscono nel riprodurre il picco.

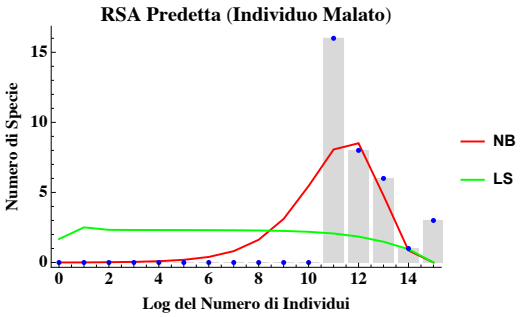

Figura 4.4: RSA predetta per individuo malato. Le previsioni della binomiale negativa riproducono l'andamento reale mentre quelle della distribuzione logaritmica falliscono nel riprodurre il picco.

Abbiamo sotto-campionato ognuno dei due campioni, selezionando casualmente per 100 volte il $10\%, 20\%, ..., 90\%$ degli individui. Per ogni sottocampione è stato poi calcolato il numero di specie predette alla scala globale. In figura 4.5 e 4.4 abbiamo inserito i grafici del numero medio predetto ad ogni sotto-scala per l'individuo sano e malato. Vediamo che il metodo della binomiale negativa predice correttamente il numero di specie anche partendo da campioni a scale ridotte, mentre quello della distribuzione logaritmica si avvicina al valore vero solo per scale molto grandi.

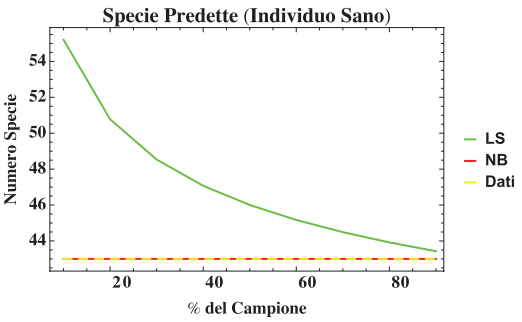

Figura 4.5: Individuo sano. Specie predette a scala globale.

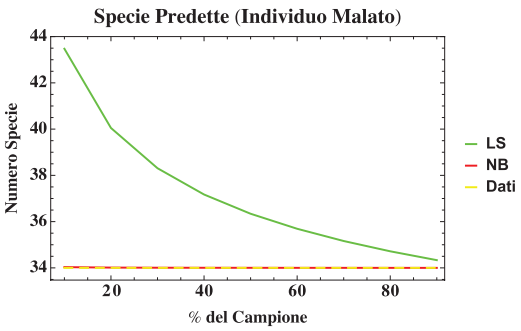

Figura 4.6: Individuo malato. Specie predette a scala globale.

Alla luce dei risultati dei test applichiamo ai nostri campioni solo i due metodi parametrici.

4.2.2 Risultati di *upscaling*

Nelle seguenti figure sono rappresentate le RSA dei due campioni. Questi istogrammi, detti *Preston Plot*, rappresentano la distribuzione relativa delle specie raggruppando nell'asse delle ascisse il logaritmo in base due del numero di individui e mostrando nell'asse delle ordinate il numero di specie per ogni classe, cioè per ogni *Preston Class*. Fare un istogramma di Preston significa costruire un sistema di suddivisione in categorie di abbondanze che raddoppiano $(1, 2, 4, 8, 16...)$ e contare quante specie appartengono alle varie classi. Le specie che hanno esattamente $1, 2, 8, 16..$ individui vanno divise equamente tra le due categorie adiacenti. Questo tipo di classificazione delle specie trasforma effettivamente i dati di abbondanza relativa delle specie nel loro logaritmo in base 2.

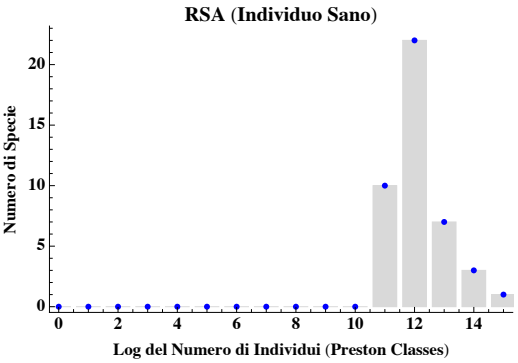

Figura 4.7: RSA individuo sano. Preston Plot.

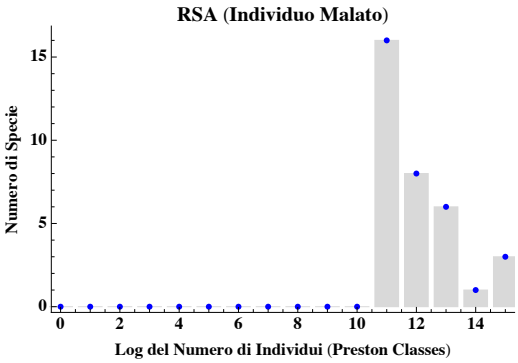

Figura 4.8: RSA individuo malato. Preston Plot.

	$\mathbf{S}^*$	$\mathbf{N}^*$	**N**	$\mathbf{p}^*$
Sano	43	176531	733551	0.240653
Crohn	34	159376	268205	0.594232

Tabella 4.1: Dati Iniziali.

Nella tabella 4.1 sono indicati, per ognuno dei due campioni, il numero di specie S^* e di individui N^* riconosciuti, il numero di sequenze N ricostruite prima che venissero scartate quelle che non hanno trovato riscontro nel database e la p^* corrispondente, stimata come $p^* = N^*/N$.

Assumendo per la RSA una forma binomiale negativa e fittando il corrispondente pattern empirico otteniamo i seguenti parametri:

	r	$\hat{\xi}_{p^*}$	ξ
Sano	2.4 ± 0.4	0.9994 ± 0.0001	0.99986 ± 0.00003
Crohn	1.2 ± 0.3	0.99975 ± 0.00007	0.99985 ± 0.00004

Tabella 4.2: Parametri Binomiale Negativa.

Assumendo invece una distribuzione logaritmica otteniamo:

	$\hat{x}_{p^*}$	x
Sano	0.999 ± 0.002	0.999995 ± 0.000004
Crohn	0.99998 ± 0.00002	0.99999 ± 0.00001

Tabella 4.3: Parametri Distribuzione Logaritmica.

dove ξ e x sono i parametri delle distribuzione a scala globale calcolati con le equazioni (3.10) e (3.21) rispettivamente.

In figura 4.9 e 4.10 possiamo vedere le RSA ottenute fittando i parametri alla scala del campione (tabelle 4.2 e 4.3).

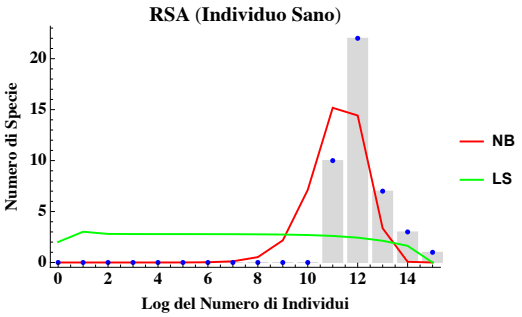

Figura 4.9: RSA individuo sano: Preston Plot e curve di fit.

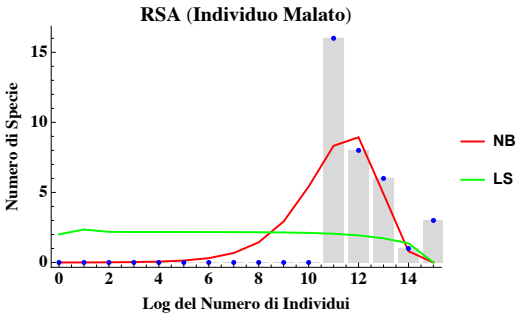

Figura 4.10: RSA individuo malato: Preston Plot e curve di fit.

Per calcolare il numero di specie con una data abbondanza k abbiamo calcolato le (2.5) e (2.8) di parametri ottenuti dai fit per $n = k$ e moltiplicato il risultato per il numero di specie presenti nel campione di riferimento.

Anche in questo caso, a differenza della distribuzione logaritmica, la binomiale negativa riproduce la reale distribuzione delle specie.

Abbiamo poi calcolato il numero di specie predetto alla scala globale, $p = 1$, ottenendo i risultati mostrati nella seguente tabella:

	S^*	S_{NB}	S_{LS}
Sano	43	43	48
Crohn	34	34	35

Tabella 4.4: Risultati di upscaling.

Notiamo che, sia per l'individuo sano sia per quello affetto da morbo di Crohn, con il metodo della binomiale negativa il numero di specie predette alla scala globale coincide con il numero di specie realmente presenti nel campione, mentre il numero predetto dal metodo della distribuzione di Fisher si discosta poco dal valore di S^*.
Questi risultati sono ragionevoli se si guarda la SAR dei campioni. Per entrambi i campioni abbiamo selezionato casualmente, per ognuna delle scale del $10\%, 20\%, ..., 90\%$, 100 sottocampioni. In media, ad ogni scala, si trova che il numero di specie presenti nei sottocampioni coincide con quello delle specie presenti alla scala del campione di riferimento. I risultati di *upscaling* sono dunque compatibili con questa evidenza.

5 Conclusioni

In questo lavoro abbiamo analizzato alcuni dei modelli, detti metodi di *upscalig*, nati in ambito ecologico, che permettono di fare previsioni sulla biodiversità di un intero ecosistema pur avendo informazioni solamente della composizione di una minima parte di questo. Conoscendo il numero di individui, N^*, e il numero di specie, S^*, presenti alla scala p^* in esame, abbiamo visto che, a seconda del metodo a cui si fa riferimento, il numero di specie, S, alla scala globale può essere stimato con:

- L'Eq. (3.14) se si assume che la SAD segua una distribuzione binomiale negativa

$$S = S^* \frac{1-(1-\xi)^r}{1-(1-\hat{\xi}_p^*)^r};$$

- Le Eqs. (3.29) o (3.31) se si assume che la SAD segua una distribuzione logaritmica di Fisher

$$S = \hat{\alpha} \log\left(1 + \frac{N}{\hat{\alpha}}\right),$$

$$S = S^* \frac{\log(1-x)}{\log(1-\hat{x}_{p^*})};$$

- L'Eq. (3.36) se utilizziamo il metodo non parametrico di $Chao_{wor}$

$$S = S^* + \frac{S_1^{*2}}{\frac{n}{n-1} 2S_2^* + \frac{p^*}{1-p^*} S_1^*}.$$

In seguito abbiamo messo in evidenza l'analogia che esiste tra le specie assenti nel campione di riferimento in ecologia, e le specie le cui reads non vengono classificate nell'ambito delle comunità microbiche. Alla luce di questa analogia abbiamo applicato i metodi descritti a dati metagenomici riguardanti la popolazione batterica dell'intestino di un individuo sano e di un altro affetto dal morbo di Crohn.
I test e le analisi condotte su questi campioni hanno rivelato che:

- Il metodo $Chao_{wor}$ non può essere applicato a dati di questo tipo;
- Il metodo della distribuzione logaritmica di Fisher fa buone previsioni se la scala del campione di riferimento non si allontana troppo da quella globale (Figure 4.5 e 4.6), ma in nessun caso riesce a riprodurre l'andamento della RSA del sistema in esame (Figure 4.9 e 4.10);
- Il metodo della binomiale negativa fa previsioni ottime (Figure 4.5 e 4.6) e riproduce parzialmente l'andamento della RSA del sistema in esame (Figure 4.9 e 4.10).

Alcune possibili sviluppi futuri di quanto analizzato in questo lavoro potrebbero essere:

- Utilizzare il metodo della binomiale negativa per ricostruire l'abbondanza relativa delle specie in comunità microbiche e predire il numero di specie batteriche nell'intero intestino, una volta stimata correttamente la percentuale di sampling, p^*. Avere informazioni complete sulla comunità che ospita il batterio causa del disturbo può cambiare e migliorare l'approccio clinico adottato;

- Utilizzare le informazioni che si possono ottenere dalle RSA previste per monitorare i cambiamenti nella composizione del microbioma degli individui, ad esempio in casi di infezione e malattia o durante la somministrazione di antibiotici;
- Studiare le differenze nelle popolazioni di comunità batteriche di individui con diverse storie cliniche.

Per concludere, con questo lavoro abbiamo voluto mostrare come i metodi di *upscaling*, sviluppati e molto utilizzati in ambito ecologico, possano dare interessanti contributi anche nell'ambito della micorbiologia.

Elenco delle figure

Elenco delle tabelle

Bibliografia

[1] Anna Tovo, Samir Suweis, Marco Formentin, Marco Favretti, Igor Volkov, Jayanth R. Banavar, Sandro Azaele, and Amos Maritan. Upscaling species richness and abundances in tropical forests. *Science Advances*, 3(10), 2017.

[2] Juan Jovel, Jordan Patterson, Weiwei Wang, Naomi Hotte, Sandra O'Keefe, Troy Mitchel, Troy Perry, Dina Kao, Andrew L. Mason, Karen L. Madsen, and Gane K.-S. Wong. Characterization of the gut microbiome using 16s or shotgun metagenomics. *Frontiers in Microbiology*, 7:459, 2016.

[3] Sandro Azaele, Amos Maritan, Stephen J. Cornell, Samir Suweis, Jayanth R. Banavar, Doreen Gabriel, and William E. Kunin. Towards a unified descriptive theory for spatial ecology: predicting biodiversity patterns across spatial scales. *Methods in Ecology and Evolution*, 6(3):324–332.

[4] Brian J. McGill, Rampal S. Etienne, John S. Gray, David Alonso, Marti J. Anderson, Habtamu Kassa Benecha, Maria Dornelas, Brian J. Enquist, Jessica L. Green, Fangliang He, Allen H. Hurlbert, Anne E. Magurran, Pablo A. Marquet, Brian A. Maurer, Annette Ostling, Candan U. Soykan, Karl I. Ugland, and Ethan P. White. Species abundance distributions: moving beyond single prediction theories to integration within an ecological framework. *Ecology Letters*, 10(10):995–1015.

[5] F. W. Preston. The commonness, and rarity, of species. *Ecology*, 29(3):254–283, 1948.

[6] STEPHEN P. HUBBELL. *The Unified Neutral Theory of Biodiversity and Biogeography (MPB-32)*. Princeton University Press, 2001.

[7] Sandro Azaele, Samir Suweis, Jacopo Grilli, Igor Volkov, Jayanth R Banavar, and Amos Maritan. Statistical mechanics of ecological systems: Neutral theory and beyond. *Reviews of Modern Physics*, 88(3):035003, 2016.

[8] Anne Chao and Chun-Huo Chiu. *Species Richness: Estimation and Comparison*, pages 1–26. American Cancer Society, 2016.

[9] Alan Gelfand and Shinichiro Shirota. Preferential sampling for presence/absence data and for fusion of presence/absence data with presence-only data. *bioRxiv*, 2018.

[10] Anne Chao and Chih-Wei Lin. Nonparametric lower bounds for species richness and shared species richness under sampling without replacement. *Biometrics*, 68(3):912–921.

[11] R. A. Fisher, A. Steven Corbet, and C. B. Williams. The relation between the number of species and the number of individuals in a random sample of an animal population. *Journal of Animal Ecology*, 12(1):42–58, 1943.

[12] Elizabeth K. Costello, Keaton Stagaman, Les Dethlefsen, Brendan J. M. Bohannan, and David A. Relman. The application of ecological theory toward an understanding of the human microbiome. *Science*, 336(6086):1255–1262, 2012.

[13] James M. Heather and Benjamin Chain. The sequence of sequencers: The history of sequencing dna. *Genomics*, 107(1):1 – 8, 2016.

[14] Peter Menzel, Kim Lee Ng, and Anders Krogh. Fast and sensitive taxonomic classification for metagenomics with kaiju. 7:11257, 2016.