

A Bayesian approach to capture-recapture models with misidentification

Davide Di Cecco, Andrea Tancredi

SISBAYES 2025

Capture-recapture and species sampling models

Capture-recapture and species sampling models, *inter alia*, represent the main approaches to estimate the size of a target population

- Capture-recapture models estimate the number of individuals
- Species sampling models estimate the number of species
- Differences in data structure:
 - ▶ capture histories or matrices vs frequency counts of species
 - ▶ but...if the capture occasions in the capture-recapture experiment are not fixed, the data structure is essentially the same (repeated count data)

- Both approaches assume perfect identification of units.
- Biased estimates if errors occur. Common in:
 - ▶ Camera traps: photo quality, observer skill, mark distinctiveness.
 - ▶ Genotyping: sample contamination, sequencing errors.

Number of Species in Microbial Diversity Studies

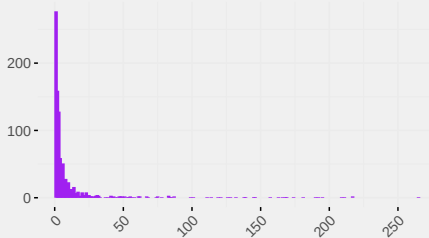
We focus on the problem of estimating the number of microbial species in specific environments

- Environmental samples (e.g., soil, water) are analyzed through sequencing of the 16S rRNA gene, which is found in the genome of prokaryotes.
- Sequences classified using clustering algorithms:
 - ▶ Based on similarity scores with threshold defined by researchers

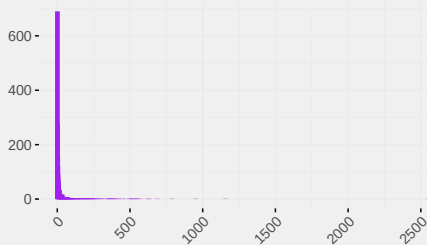
Diversity estimation in microbial communities is complex due to:

- High species richness, skewed abundance distribution
- Unlike macroecology, rare species dominate microbial samples
- **Singletons** (species observed once) are the modal class

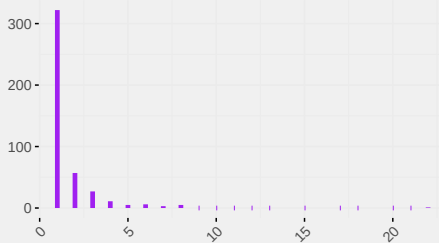
Apples s=11740



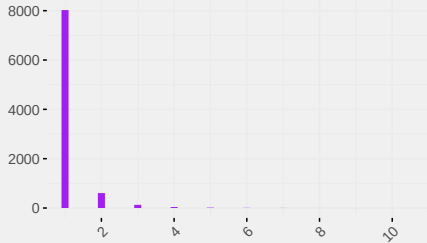
Hawaii s=57610



Epstein s=705



Swine s=9988



Artefactual Inflation:

- PCR-based sequencing introduces errors → ghosts species
- These errors inflate species counts, especially singletons
- Confirmed by simulations with cultivated in vivo communities

Key Effects on Diversity Estimation:

- Overcounting of singletons affects:
 - ▶ Species richness — total number of distinct (including unsampled) species
 - ▶ Community heterogeneity — measured via Shannon's diversity index

Standard approach without misidentification

- N distinct species. Each species observed X_i^* times.
- Observed data $n_1^*, n_2^*, \dots$ where

$$n_j^* = \sum_{i=1}^N \mathbb{I}_{\{X_i^*=j\}} \quad \text{and} \quad n^* = \sum_{j \geq 1} n_j^*$$

- **Likelihood.** Let $f_j^*(\theta) = P(X_i^* = j; \theta)$, e.g.¹

$$f_j^*(\theta) = \frac{\Gamma(j+r)}{\Gamma(r)j!} p^r (1-p)^j$$

$$L(N, \theta) = \binom{N}{n_0^*, n_1^*, n_2^*, \dots} \prod_{j \geq 0} (f_j^*(\theta))^{n_j^*}$$

where $n_0^* = N - n^*$ is the number of unobserved species.

¹Negative binomial, R.A.Fisher 1943, British ecological society

- Negative binomial for f^* may be adequate in real applications. Each species follows a $\text{Poisson}(\lambda_i)$ with $\lambda_i \sim \text{Gamma}(r, p/(1 - p))$
- **Boundary problem** $r \approx 0$ i.e. $\lambda_i \approx 0$ and $\hat{N} = n/(1 - e^{-\lambda}) \rightarrow \infty$ may inflate the N estimation ... *if heterogeneity is very great r becomes small and approaches zero... Now, the abundance in nature of different species of the same group generally varies very greatly, so that, as I first found in studying Corbet's series of Malayan butterflies, the negative binomial, which often fits such data well, has a value of r so small as to be.... indistinguishable from zero.* *That it is not really zero for collections of wild species follows from the fact that the total number of species, and therefore the total number not included in the collection, is really finite.* (Fisher 1943)
- Alternatives: finite Poisson mixture models, NPMLE and penalized NPMLE (Lindsay and coauthors) **BNP** approaches (starting from Lijoi et al. BKA 2007)

$M_{h,\alpha}$ model for species sampling

- For each species, we observe $X_i = X_i^* - M_i$ occurrences where

$$M_i | X_i^* = x_i^* \sim \text{Binomial}(\alpha, x_i^*) \quad \text{thinning}$$

- Marginal distribution of X_i

$$P(X_i = j | f^*, \alpha) = \tilde{f}_j(\theta, \alpha) = \sum_{m_i=0}^{\infty} \binom{j + m_i}{m_i} \alpha^{m_i} (1 - \alpha)^j f_{j+m_i}^* \quad j \geq 0$$

If $X_i^* \sim \text{NB}(r, p)$ then $X_i \sim \text{NB}(r, p/(1 - \alpha(1 - p)))$

- Thinned counts $\tilde{n}_j = \sum_{i=1}^N \mathbb{I}_{\{x_i=j\}}$ have distribution

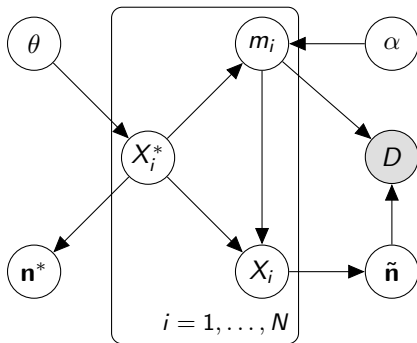
$$p(\tilde{\mathbf{n}} | N, \theta, \alpha) = \binom{N}{\tilde{n}_0, \tilde{n}_1, \tilde{n}_2, \dots} \prod_{j \geq 0} (\tilde{f}_j(\theta, \alpha))^{\tilde{n}_j}$$

where $\tilde{\mathbf{n}} = \{\tilde{n}_j\}_{j \geq 0}$ and $\tilde{n}_0 = n_0^* + \sum_{i=1}^N \mathbb{I}_{\{m_i=x_i^*\}}$

- Observed data are $D = (n_1, n_2, \dots)$ where

$$n_1 = \tilde{n}_1 + m \quad m = \sum_{i=1}^N m_i \quad n_j = \tilde{n}_j \quad j \geq 2$$

- A DAG



Likelihood and Bayesian inference

- Likelihood

$$\begin{aligned} p(D|N, \theta, \alpha) &= \sum_{\tilde{n}_0, \tilde{n}_1} p(D, \tilde{n}_0, \tilde{n}_1|N, \theta, \alpha) \\ &= \sum_{\tilde{n}_1} p(M = n_1 - \tilde{n}_1|\tilde{\mathbf{n}}, \theta, \alpha) P(\tilde{\mathbf{n}}|\theta, \alpha, N) \end{aligned}$$

where $M|\tilde{\mathbf{n}}, \theta, \alpha$ is the sum of independent r.v. $\sum_{i=1}^N [M_i|X_i, \theta, \alpha]$

- $X_i^* \sim Poi(\lambda)$: $M_i|X_i, \lambda, \alpha \sim Poi(\lambda\alpha)$ and $M|\tilde{\mathbf{n}}, \lambda, \alpha \sim Poi(N\lambda\alpha)$
- $X_i^* \sim NB(r, p)$: $M_i|X_i, \theta, \alpha \sim NB(r + x_i, 1 - \alpha(1 - p))$ and

$$M|\tilde{\mathbf{n}}, \lambda, \alpha \sim NB(rN + \sum_j j\tilde{n}_j, 1 - \alpha(1 - p))$$

Prior and MCMC

Assume $p(\theta, \alpha, N) = p(\theta)p(\alpha)/N$ with a $\text{Beta}(a_\alpha, b_\alpha)$ prior for α . In the NB model we take an inverse gamma prior for r and a $\text{Beta}(a_p, b_p)$ prior for p .

MCMC algorithm for the $NB(r, p)$ model with misidentification

- 1) Metropolis-Hastings step for $p(\tilde{n}_0, \tilde{n}_1 | D, \theta, \alpha)$ which is a negative Multinomial for $\tilde{n}_0, \tilde{n}_1$ multiplied by the factor $p(M = n_1 - \tilde{n}_1 | \tilde{\mathbf{n}}, \theta, \alpha)$
- 2) Draw $\alpha \sim \text{Beta}(a_\alpha + m, b_\alpha + s - m)$ where $s = \sum_{j \geq 1} j n_j$
- 3) Draw $p \sim \text{Beta}(a_p + rN, b_p + s)$
- 4) Metropolis for $\frac{p^{rN} \Gamma(s + rN)}{\Gamma(s - m + rN)} \prod_{j \geq 0} \left(\frac{\Gamma(j + r)}{\Gamma(r) j!} \right)^{\tilde{n}_j} p(r)$

True abundance distribution and entropy

- Possibility to recover the true X_i^*
- Let $M_i \sim NB(r_i, p)$, $i = 1, \dots, N$, then

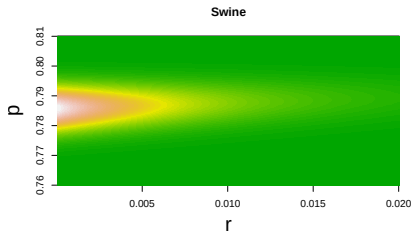
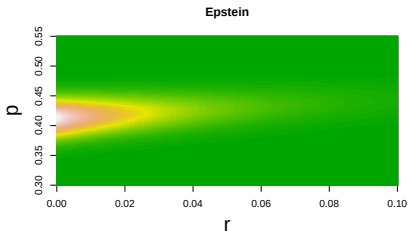
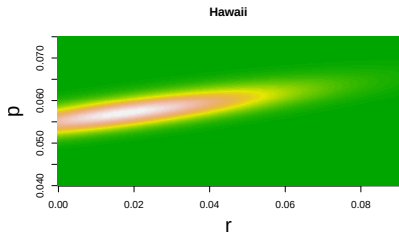
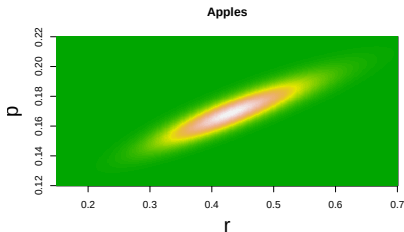
$$P(m_1, \dots, m_N | m) = \frac{\Gamma(\sum_i r_i) \Gamma(m+1)}{\Gamma(m + \sum_i r_i)} \prod_{i=1}^{N^*} \frac{\Gamma(m_i + r_i)}{\Gamma(r_i) \Gamma(m_i + 1)}$$

- At each MCMC step draw $X_i^* = X_i + M_i$ for $i = 1, \dots, N$ where $M_1 \dots M_N$ are generated from a Multinomial Dirichlet distribution with size m and parameters $x_1 + r, \dots, x_N + r$
- We can draw the posterior predictive distribution for the entropy

$$E = \exp \left(- \sum_{j \geq 1} n_j^* \frac{j}{s} \ln \frac{j}{s} \right)$$

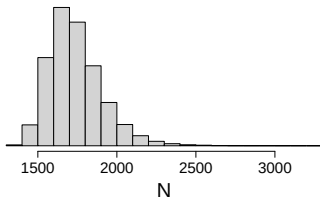
Results

- Negative Binomial model: conditional likelihood for (p, r)

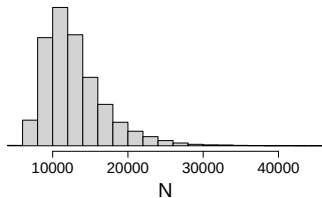


- Negative Binomial model: posterior for N

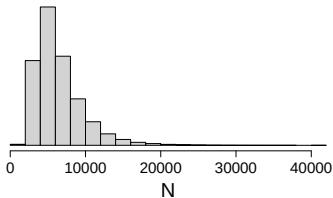
Apples



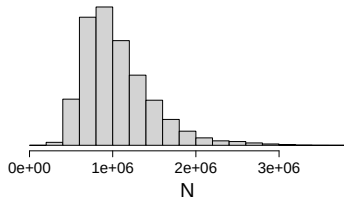
Hawaii



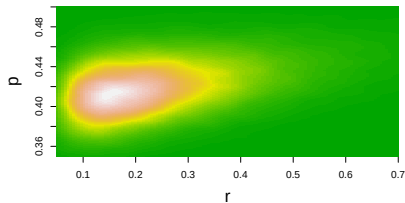
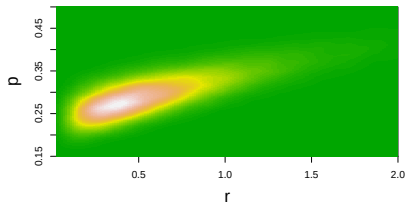
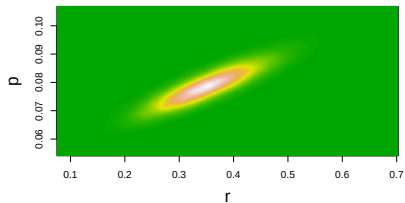
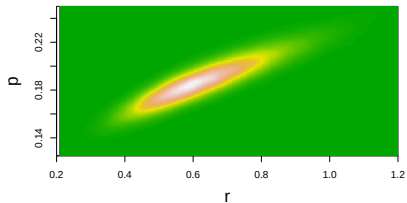
Epstein



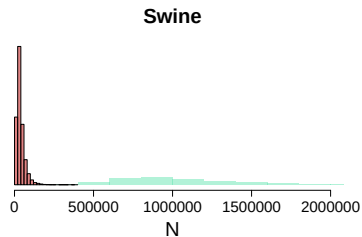
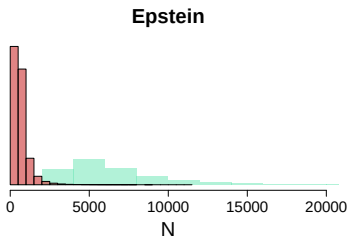
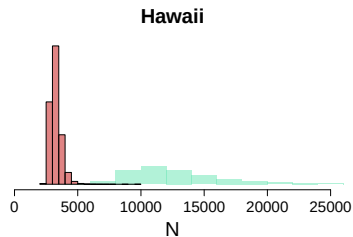
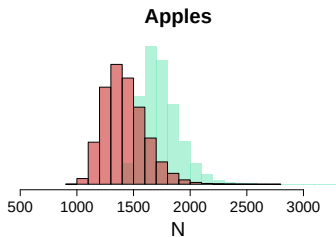
Swine



- $M_{h,\alpha}$ Negative Binomial model: posterior for p, r

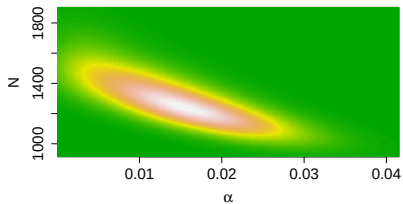


- $M_{h,\alpha}$ Negative Binomial model: posterior for N

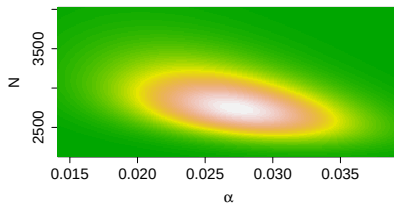


- $M_{h,\alpha}$ Negative Binomial model: posterior for α, N

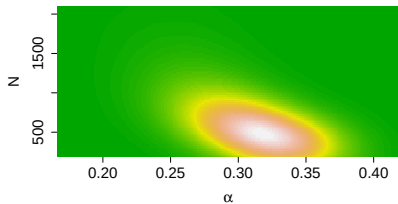
Apples



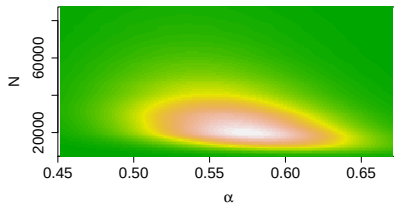
Hawaii



Epstein



Swine



- Model comparison: DIC

Method	Apples	Hawaii	Epstein	Swine
$M_{h,\alpha}$ NB	213.6	299.3	68.2	89.1
NB	160.4	332.3	94.6	291.9
$M_{h,\alpha}$ Poisson	1630.9	26934.6	112.3	168.1
Poisson	2194.5	31536.7	279.4	754.9

- Shannon's diversity E assuming no misidentification, and posterior modes and 95% credibility intervals of E^* under the $M_{h,\alpha}$ NB

	Apples	Hawaii	Epstein	Swine
E	291.0	477.3	334.4	8288.5
E^*	288.2	469.1	201.7	4826.5
	(285.1;291.0)	(466.8;471.4)	(154.3;254.6)	(4186.7; 5397.8)

Method	Without considering errors			
	Apples ($n = 1000$)	Hawaii ($n = 2319$)	Epstein ($n = 437$)	Swine ($n = 8833$)
Chao	1241 (1177;1328)	3182 (3024;3374)	1347 (1083;1718)	62057 (57415;67142)
Chao-Bunge	1629 (1434;1894)	8826 (5947;13798)	—	—
Lanumteang-Bohning	1482 (1258;1903)	4044 (3398;5077)	2256 (818;9119)	115280 (65918;207328)
breakaway	1617 (1121;4153)	8826 (5987;13862)	2162 (780;9102)	39792 (37751;41975)
wlrm_transformed	1219 (1153;1311)	2819 (2690;2992)	1849 (1277;2809)	120430 (86204;169796)
Poisson	1046 (1030;1060)	2423 (2400;2445)	672 (618;731)	39767 (37705;41923)
NB	1685 (1486;2166)	11318 (7096;21694)	4778 (2625;14789)	872584 (497416;2·10 ⁶)
Considering errors in singletons				
breakaway_nof1	1397 (1015;11501)	1631 (1630;1640)	604 (455;1951)	846113 (52882;1.6·10 ⁷)
Chiu-Chao	839 (697;1023)	2491 (2061;3023)	275 (77;676)	10252 (5220;16540)
$M_{h,\alpha}$ Poisson	786 (766;805)	1710 (1688;1731)	167 (142;194)	2796 (2414;3221)
$M_{h,\alpha}$ NB	1365 (1097;1778)	3090 (2562;4134)	409 (190;1527)	23319 (7547;90353)

Discussion

- $M_{h,\alpha}$ for misidentification in the repeated count data framework
- Assumes spurious singletons due to identification errors.
- Possible source of **one-inflation**
- Extensions to other mixed Poisson distributions
- **Discounting**: Estimate $\tilde{n}_0, \tilde{n}_1$ from $\{n_j\}_{j \geq 2}$, ignoring n_1 .
 - ▶ Valid estimation for $N = \tilde{n}_0 + \tilde{n}_1$ for models closed under thinning

+

References

- Link W, Yoshizaki J, Bailey L, Pollock K. (2010) Uncovering a latent multinomial: analysis of mark–recapture data with misidentification. - *Biometrics*
- Tuoto, T., Di Cecco, D. and Tancredi, A. (2022) Bayesian analysis of one–inflated models for elusive population size estimation. *Biometrical Journal*
- Di Cecco D, Tancredi, A Tuoto T (2024) Semi-parametric Bayesian approach for population size estimation modelling the excess of singletons. *JRSSC*
- Di Cecco, D., Tancredi, A. (2024) Estimating the number of sequencing errors in microbial diversity studies. *Environmental and Ecological Statistics*
- Di Cecco, D., Tancredi, A. A Bayesian parametric approach to capture-recapture with misidentification. *Submitted*

Thank you