STANDARD OPERATING PROTOCOL · EnGeL-SAMPL-EFFORT

Measuring sampling effort and completeness: rarefaction, extrapolation and richness estimators (iNEXT, SpadeR)

Protocol ID	EnGeL-SAMPL-EFFORT
Version	1.1.0
Date	2026-11-20
Status	draft
Area	Sampling and experimental design (SAMPL)
Authors	Luís Cunha (lead; analysis script; review and approval; ORCID 0000-0002-5870-2537)
Reviewed and approved by	Luís Cunha
Laboratory	Environmental Genomics Laboratory (EnGeL-UC), Environmental Ecology & Ecotoxicology Group (E3G), Centre for Functional Ecology, Department of Life Sciences, University of Coimbra · Laboratório Associado TERRA
Contact	luis.cunha@uc.pt · ewormuc@gmail.com
Licence	Text: CC BY 4.0 · Code: MIT
Main hazards	None: computational protocol

1 Purpose and scope

This protocol tells you whether a sampling effort was enough: how many species (OTUs) were detected compared with how many are expected, how complete the sample is (sample coverage), and how richness and diversity would change with more or less sampling. It applies to incidence data (presence/absence per sampling unit, for example per soil core or per DNA sample) and to abundance data (counts of individuals per species). Use it to compare sites or treatments at equal sample size or equal coverage, and, together with EnGeL-SAMPL-POWER (sample size before fieldwork), to plan the sampling effort of new studies.

2 Principle

Rarefaction (interpolation) and extrapolation curves of Hill numbers (q = 0: species richness; q = 1: exponential of Shannon entropy; q = 2: inverse Simpson) are drawn against sample size and against sample coverage (Chao et al., 2014). iNEXT computes the curves with bootstrap confidence intervals (Hsieh et al., 2016). SpadeR gives asymptotic richness estimators (for example Chao1 and Chao2, ICE, jackknife) and the number of species shared between two sites (Chao et al., 2016).

3 Before you start

Item	Details
Hands-on time	[TO CONFIRM]
Input	Incidence table (species × sampling units, 0/1) or abundance counts; one table per site or treatment to compare
Prepare in advance	R packages installed; sampling units and their grouping (sites, treatments) defined
Key safety points	Not applicable

4 Materials

4.1 Computer

- Computer with R and RStudio (lab workstation or personal computer)

4.2 Consumables

- None

4.3 Software and input data

- R [TO CONFIRM: version] with the packages iNEXT, SpadeR and ggplot2 [TO CONFIRM: package versions]
- Community data: incidence table (OTUs or species × sampling units, 0/1), e.g. OTU_incidence.xlsx from EnGeL-BIOINF-COMMUNITY; or abundance counts per species (e.g. morphospecies counts from EnGeL-FAUNA-MAR-MORPH)

5 Reagent preparation

Data format. Start from the incidence table (or the abundance counts). For incidence data, prepare one matrix per site or treatment with species (or OTUs) as rows and sampling units as columns (1 = detected, 0 = not detected): iNEXT and SpadeR call this `incidence_raw`. For abundance data, prepare a vector of counts per species for each site (`abundance`). Combine the matrices or vectors of all sites in a named list.

6 Procedure

6.1 1. Set up

1. Open the project's R project and load the packages:

```
library(iNEXT)
library(SpadeR)
library(ggplot2)
```

> **CRITICAL** The source script loads BiodiversityR, but the functions `ChaoSpecies`, `Diversity`, `ChaoShared` and `SimilarityMult` used below come from SpadeR: load SpadeR.

2. Optional, to learn the functions: run the steps below on the example data supplied with the packages (`data(spider)` in iNEXT, abundance data; `data(ChaoSpeciesData)` and `data(ChaoSharedData)` in SpadeR, incidence data).

6.2 2. Rarefaction and extrapolation (iNEXT)

3. Read your data into a named list of incidence matrices (one per site or treatment) and run iNEXT for q = 0, 1 and 2:

```
out <- iNEXT(data_list, q = c(0, 1, 2), datatype = "incidence_raw",
             endpoint = NULL, nboot = 500)   # endpoint NULL: twice the reference sample size
```

For abundance data use `datatype = "abundance"`. To compute the curves at chosen sample sizes, give them in `size =` (the tutorial uses 1, 5, 20, 50, 100, 200 and 400 individuals for the spider example) [TO CONFIRM: sizes or end-point for EnGeL data].

4. Plot the curves: sample-size-based curves (`type = 1`), the sample completeness curve (`type = 2`) and coverage-based curves (`type = 3`), with one panel per site:

```
ggiNEXT(out, type = 1, facet.var = "Assemblage")
ggiNEXT(out, type = 2)
ggiNEXT(out, type = 3, facet.var = "Assemblage")
```

5. Save the numerical results (`out$DataInfo`, which includes the sample coverage, and `out$AsyEst`, the asymptotic estimates) as tables in `results/`.

6.3 3. Richness estimators and shared species (SpadeR)

6. Asymptotic richness estimators and diversity profiles for one site, saving the full output as text:

```
rich <- ChaoSpecies(site_matrix, datatype = "incidence_raw", k = 10, conf = 0.95)
capture.output(rich, file = "results/ChaoSpecies_site.txt")
div  <- Diversity(site_matrix, datatype = "incidence_raw")
capture.output(div, file = "results/Diversity_site.txt")
```

7. Species shared between two sites (the matrix holds the sampling units of both sites, side by side; `units` gives the number of units of each site):

```
ChaoShared(two_site_matrix, datatype = "incidence_raw", units = c(n1, n2),
           se = TRUE, nboot = 500, conf = 0.95)
```

The lab uses 500 bootstrap replicates (steadier confidence intervals than 200, at the cost of run time). For more than two sites, `SimilarityMult()` compares several sites at once (the tutorial uses `q = 2` and relative similarity; use `nboot = 500`).

7 Calculations

Not applicable: the estimators are computed by the packages. Report the estimator used for each value (e.g. Chao2 for incidence data).

8 Quality control

Check	Acceptance criterion
Data format	Matrices contain only 0 and 1 (incidence) or non-negative integers (abundance); no empty sampling units
Sample coverage of each site	≥ 90%; below this, the sampling is reported as incomplete and comparisons are made at equal coverage
Comparisons between sites	Made at equal sample size or equal coverage, within the extrapolation range recommended by the package authors (up to about twice the reference sample size)

9 Data and metadata

- **Outputs:** iNEXT tables (`DataInfo`, `AsyEst`, curve values), plots (sample-size-based, completeness, coverage-based), SpadeR text outputs.
- **Where:** project folder in the lab's shared folder; the script in the project's `scripts` folder.

Minimum to record
- Input data (file, date), type (incidence or abundance), sampling unit and number of units per site
- R and package versions (`sessionInfo()`)
- Parameters: q values, endpoint or sizes, `k`, `nboot`, confidence level
- Sample coverage and estimated richness per site

10 Safety and waste

Not applicable (computational protocol).

11 Troubleshooting

Problem	Likely cause	Solution
`could not find function "ChaoSpecies"`	SpadeR not loaded	`library(SpadeR)`.
Very wide confidence intervals in the extrapolated part	Extrapolation too far beyond the reference sample	Extrapolate to at most about twice the reference sample size.
Error about the data type	Matrix not in `incidence_raw` format, or counts given as `incidence_raw`	Check that the data are 0/1 matrices (incidence) or count vectors (abundance) and set `datatype` accordingly.

12 Remarks

The source is a tutorial: it first runs the functions on the packages' example data, then on lab data (a transposed taxon table, `R_taxo_all_transp_nonames.csv`, with 20 sampling units in one call) [TO CONFIRM: study and data set behind the lab example]. Metabarcoding data are best analysed as incidence data (presence/absence per sampling unit), because read numbers are not abundances.

13 References

Chao, A., Gotelli, N.J., Hsieh, T.C., et al. (2014). Rarefaction and extrapolation with Hill numbers: a framework for sampling and estimation in species diversity studies. Ecological Monographs, 84(1), 45–67. https://doi.org/10.1890/13-0133.1

Chao, A., Ma, K.H., Hsieh, T.C., Chiu, C.-H. (2016). SpadeR: Species-richness prediction and diversity estimation with R. R package [TO CONFIRM: version].

Hsieh, T.C., Ma, K.H., Chao, A. (2016). iNEXT: an R package for rarefaction and extrapolation of species diversity (Hill numbers). Methods in Ecology and Evolution, 7(12), 1451–1456. https://doi.org/10.1111/2041-210X.12613

14 Version history and citation

Version	Date	Change	By
1.0.0	2026-11-17	First EnGeL-UC version, from the lab's R tutorial script for iNEXT and SpadeR (iNEXT_tutorial.R).	Luís Cunha

Version	Date	Change	By
1.0.1	2026-11-17	Author of the source script confirmed (Luís Cunha).	Luís Cunha
1.0.2	2026-11-20	Cross-reference to EnGeL-SAMPL-POWER for planning sample size.	Luís Cunha
1.1.0	2026-11-20	Values confirmed by Luís Cunha: sample coverage of at least 90% regarded as sufficient; 500 bootstrap replicates.	Luís Cunha

How to cite

Cunha L (2026). Measuring sampling effort and completeness: rarefaction, extrapolation and richness estimators (iNEXT, SpadeR). EnGeL-UC protocol EnGeL-SAMPL-EFFORT, version 1.1.0. Zenodo. DOI [TO CONFIRM].

Licence

Text: CC BY 4.0. Code: MIT. Logos are the property of their institutions and are not covered by the licence.

15 Open questions

1. R and package versions (iNEXT, SpadeR); hands-on time.
2. Sample sizes or endpoint to use for EnGeL data.
3. Study and data set behind the lab example in the source script.
4. DOI.