

# Supplementary Material

## Methods, Monte Carlo tests and Appendices

This document contains complementary information about the how the data analysis was actually performed, supporting figures, and the two appendices cited in the main article. The C programs used in this paper are available on request.

## Methods

### General fitting method: maximum likelihood

In order to assess the quality of the agreement to the fitted model, a typical  $\chi^2$  is calculated as:

$$\chi^2 = \sum_{i=1}^a \frac{(O_i - E_i)^2}{E_i}, \quad (1)$$

where  $a$  is total number of abundance classes in which our data are divided,  $E_i$  is the average number of species at the abundance class  $i$ , given by a theoretical model, and  $O_i$  is the number of species actually observed. Under the assumption of statistical independence of the constituent terms, and a constant variance matching expected values through the different abundance classes, the sum in Eq. (1) is distributed following a  $\chi^2$  distribution with  $a - 1 - p$  degrees of freedom, where  $p$  is the number of fitted parameter of the theoretical model. These assumptions are rarely met in the analysis of abundance data by applying  $\chi^2$  tests (see Fig S2). To avoid this problem, here Monte Carlo tests are used instead (see below). Furthermore, in order to preserve statistical independence (and unlike Preston histograms) abundance bins are always chosen to be doubling non-overlapping intervals, i.e. abundance class  $i$  is defined to contain all species with abundance  $n$  such that  $2^{i-1} \leq n < 2^i$ ,  $i = 1, \dots$

In all cases  $\chi^2$  values were computed by pooling together the last two abundance intervals. Notice that the probability values for the log-series must be computed with 10 degrees of freedom as well, since this distribution is also one-parametric. The estimated  $\theta$  for the log-series corresponding to Fisher's  $\alpha$  takes the same value found by Williams using Fisher's (1943) method. It can be shown that Fisher's method furnishes the maximum likelihood estimate for the parameter  $\alpha$  (and consequently also for  $x$ ). The estimated parameters and the confidence intervals (at the 0.9 level) were calculated by using maximum likelihood (Hilborn & Mangel, 1997). Since the ZSM (Eq. (5)) is a two-parametric distribution, the corresponding confidence for the maximum likelihood estimates should be properly given as a confidence surface. Here, in an approximate way, we have associated confidence intervals to each parameter by sequentially keeping constant first  $\theta = 44$ , and letting  $m$  define its likelihood profile, and then fixing  $m = 0.15$ , and letting  $\theta$  define its likelihood profile.

## The likelihood ratio test

For clarity, we explain how to perform this test for the Lepidoptera community. We calculated the negative log-likelihood for the maximum likelihood estimates for the metaZSM and the localZSM:

$$\mathbf{L}\{S_1, \dots, S_m | \theta, m\} = 1086.06 \quad (2)$$

$$\mathbf{L}\{S_1, \dots, S_m | \theta\} = 1088.04. \quad (3)$$

The observed ratio of likelihoods is 1.98. The statistic defined by  $\mathcal{R} = 2(\mathbf{L}\{S_1, \dots, S_m | \theta\} - \mathbf{L}\{S_1, \dots, S_m | \theta, m\})$  has a  $\chi^2$  distribution with 1 degree of freedom. Since  $Pr(\chi^2 > 3.96|1) < 0.05$  ( $Pr(\chi^2 < 3.84|1) = 0.95$ ), we can conclude that the ZSM is only slightly significantly better than the mZSM at the confidence level of 0.05%.

## Monte Carlo Tests

### Sampling the metacommunity: Ewens' algorithm

As can be seen in Table 1, when performing the fitting of the mZSM to the Lepidoptera data,  $Pr(\chi^2 > 16|df = 10)$  is 0.1, by using the standard  $\chi^2$  test. Therefore, if we accept a number of assumptions (see above) and assume that these data are a random sample from the metacommunity, we should observe that random samples of the same size deviate from the theory by more than 16 — the  $\chi^2$  deviation observed in the data — 10% of the time. However, since we have a way to generate random samples of a given size from the metacommunity (see Hubbell, 2001, Ewens' algorithm, pg. 291), we know exactly how randomness enters the data under this null hypothesis. Thus, we can build a tailored Monte Carlo test in the hope of improving our confidence in rejecting the null hypothesis.

By using the Ewens' algorithm, 40,000 samples of the same number of 15,609 individuals were generated at random from a model metacommunity with  $\theta = 39.8$ . For each pseudo-sample, we used the same maximum likelihood procedure to estimate the  $\theta$  parameter and computed the  $\chi^2$  (Eq. (1)) deviation from the theoretical prediction (mZSM) in the same way as we analyzed the real data. In 95% of cases the Monte Carlo  $\chi^2$  statistic was lower than the observed value in the data (Fig. S2). In this figure we present the behavior of the parametric  $\chi^2$  (Fig. S2 A) and the Monte Carlo  $\chi^2$  (Fig. S2 B) statistic for comparison purposes.

### Sampling the local community: Etienne's algorithm

Recently, Etienne & Olf (2004) developed a genealogy-based approach to derive the multivariate abundance distribution of the local community as an extension of Ewens' multivariate distribution. The same approach allows an extension of Ewens' algorithm to be readily carried out (Etienne, 2004). This procedure enables the rapid generation of pseudo-samples from a dispersal-limited ( $m < 1$ ) local community embedded in a metacommunity characterized by biodiversity number,  $\theta$ . By using this algorithm, 10000 samples of 15,609 individuals were generated from a model local community characterized

by  $\theta = 41$  and  $m = 0.77$ . The same maximum likelihood procedure was again used to estimate the parameters and to compute the  $\chi^2$  statistic, just as was done for the real data. In 95% of cases the Monte Carlo  $\chi^2$  statistic was lower than that observed in the data.

## A Relationship between different approaches

In this appendix we show how Eq. (10), the SAD under the assumption of infinite metacommunities:

$$S(n) = \theta \int_0^1 P_s(n; J, m, x) \frac{(1-x)^{\theta-1}}{x} dx, \quad (\text{A1})$$

is related to the solution given in Volkov *et al.* (2003). In order to arrive at the solution reported there, the authors first derive the stationary abundance distribution in a metacommunity of size  $J_M$  undergoing neutral zero-sum dynamics. By assuming linear transition rates, (Eqs. (5) in the main text), this distribution turns out to be a Fisher logseries. They then show that the average number of species in the metacommunity in the stationary state can be written as:

$$S_M(n) = S_M \int_0^{J_M} d\mu \hat{\rho}(\mu) P_s(n; J, m, \mu/J_M), \quad (\text{A2})$$

where  $F(\mu) \equiv P_s(n; J, m, \mu/J_M)$ , and  $\hat{\rho}(\mu)$  is the distribution of abundances in the metacommunity. Finally, they take  $\hat{\rho}(\mu)$  to be a singularity-free description of a logseries:

$$\hat{\rho}(\mu) = \frac{1}{\Gamma(\epsilon)\delta^\epsilon} \exp(-\mu/\delta)\mu^{\epsilon-1}. \quad (\text{A3})$$

Notice that this is a gamma density distribution function with parameters  $\epsilon$  and  $\delta$ . Fisher showed that the Poissonian sampling of this distribution gives rise to the negative binomial as a sample distribution. He then considered the probability of having a certain number of species in a sample of size  $J$  with a given abundance  $n$  (in relation to the probability of any species being actually observed in the sample). This sample distribution, in the limit of vanishingly small  $\epsilon$ , takes the form of his famous logseries distribution. By introducing Eq. (A3) into Eq. (A2) and following similar steps, including the consideration of an infinite metacommunity, Volkov *et al.* (2003) arrive at the expression

$$S(n) = \theta \int_0^1 P_s(n; J, m, x) \exp(-\theta x) \gamma dx. \quad (\text{A4})$$

This is the central expression, Eq. (7), in Volkov *et al.* (2003), but here we have made simple change of variable and the notation introduced in this paper has been used. This is also a sample distribution for infinite metacommunities, giving the average number of species represented by  $n$  individuals in a local community of size  $J$  characterized by a degree of isolation (immigration)  $m$ , and a biodiversity number  $\theta$ . For completeness we note that the conditional probability  $P_s(n; J, m, x)$  can be written as (Volkov *et al.*, 2003; McKane *et al.*, 2004):

$$P_s(n; J, m, x) = \binom{J}{n} \frac{\Gamma(n + \gamma x)}{\Gamma(\gamma x)} \frac{\Gamma(\nu - n)}{\Gamma(\nu - J)} \frac{\Gamma(\lambda + \nu - J)}{\Gamma(\lambda + \nu)}, \quad (\text{A5})$$

where

$$\begin{aligned}\lambda &= \gamma x \\ \nu &= J + \gamma(1 - x) \\ \gamma &= \frac{m(J - 1)}{1 - m}\end{aligned}\tag{A6}$$

Both Eqs. (A4) and ((10), in the main text) are asymptotic sample distribution assuming infinite metacommunities. The difference between them can be summarized as follows. The former assumes a singularity-free continuous description of the log-series at the meta-community level — a gamma density function of a vanishing small  $\epsilon$  parameter (Fisher *et al.*, 1943). The latter equation takes the asymptotic representation of the exact distribution (see Eq. (6) in the main text) as the continuous description for the stationary species abundance distribution in the metacommunity. In a forthcoming work, we will discuss other aspects of the relationship between the solutions obtained by Volkov *et al.* (2003), Vallade & Houchmandzadeh (2003), McKane *et al.* (2004), Etienne & Olf (2004) along with a simple way to derive the exact stationary distribution in the metacommunity.

## B A simple formula for the Poisson mZSM

In this appendix we provide the derivation of Eq. (13) from Eq. (12) in the main text.

The expression for the expected number of species with  $n$  individuals in a random sample from the metacommunity, given by (12), after the change of variables  $y = Jx$  may be written as

$$S(n) = \frac{\theta}{n} \int_0^J f_{n,1}(y) \left(1 - \frac{y}{J}\right)^{\theta-1} dy, \quad (\text{B1})$$

where

$$f_{n,\delta}(y) = \frac{1}{\Gamma(n)\delta^n} \exp(-y/\delta) y^{n-1} \quad (\text{B2})$$

is the probability density for a gamma distribution with parameters  $n$  and  $\delta$ . Therefore, since  $J$  is very large, Eq. (B1) may be written

$$S(n) \approx \frac{\theta}{n} \int_0^\infty f_{n,1}(y) \left(1 - \frac{y}{J}\right)^{\theta-1} dy = \frac{\theta}{n} E \left[ \left(1 - \frac{Y}{J}\right)^{\theta-1} \right] \quad (\text{B3})$$

where  $Y$  is a gamma variate with parameters  $n$  and  $\delta = 1$ . By expanding the function  $f(Y) = \left(1 - \frac{Y}{J}\right)^{\theta-1}$  in Taylor's series around  $Y = n$ , and given that  $E[Y] = n$  and  $\text{Var}(Y) = n$  for a gamma variate of probability density (B2), we find:

$$S(n) = \frac{\theta}{n} \left(1 - \frac{n}{J}\right)^{\theta-1} + \frac{1}{2} \frac{\theta(\theta-1)(\theta-2)}{J^2} \left(1 - \frac{n}{J}\right)^{\theta-3} + \mathcal{O}\left(\frac{1}{J^3}\right) \quad (\text{B4})$$

Since the sample size is usually large ( $J > 100$ ), we can use the formula given in (13) in the main text as a very good approximation.

## References

- Etienne, R. (2004). Bayesian analysis of species abundance data. Manuscript.
- Etienne, R. S. & Olf, H. (2004). A novel genealogical approach to neutral biodiversity theory. *Ecol. Lett.* **7**:170–175.
- Fisher, R., Corbet, A. & Williams, C. (1943). The relation between the number of species and the number of individuals in a random sample of an animal population. *J. Anim. Ecol.* **12**:42–58.
- Hilborn, R. & Mangel, M. (1997). *The Ecological Detective. Confronting Models with Data*. Princeton University Press, Princeton.
- Hubbell, S. P. (2001). *The Unified Neutral Theory of Biodiversity and Biogeography*. Princeton University Press, Princeton.
- McKane, A., Alonso, D. & Solé, R. V. (2004). Analytical solution to Hubbell’s neutral model of local community dynamics. *Theor. Popul. Biol.* **65**:67–73.
- Vallade, M. & Houchmandzadeh, B. (2003). Analytic solution of a neutral model of biodiversity. *Phys. Rev. E* **68**:061902.
- Volkov, I., Banavar, J. R., Hubbell, S. P. & Maritan, A. (2003). Neutral theory and relative species abundance in ecology. *Nature* **424**:1035–1037.

**Fig. S1.** Abundance data on Williams' Lepidoptera. Absolute number of species at each abundance interval. Notice that the second abundance class is overestimated by the theory, while the third abundance class is underestimated by the theory. The same kind of fluctuation is to be observed when we compare species within abundances in the sixth and seventh bin against predicted values. Therefore, data seem to be very noisy, although there is no trend in the residuals to be observed, because they are scattered at random. The point in the last bin is caused by only one super-abundant species (*Agrotis exclamationis*) represented by 2349 individuals in the sample.

**Fig. S2.** Monte Carlo test.  $\chi^2$  probability functions with 10 degrees of freedom (A) and Monte Carlo generated (B). The density distribution of the Monte Carlo  $\chi^2$  statistic is also shown (C).

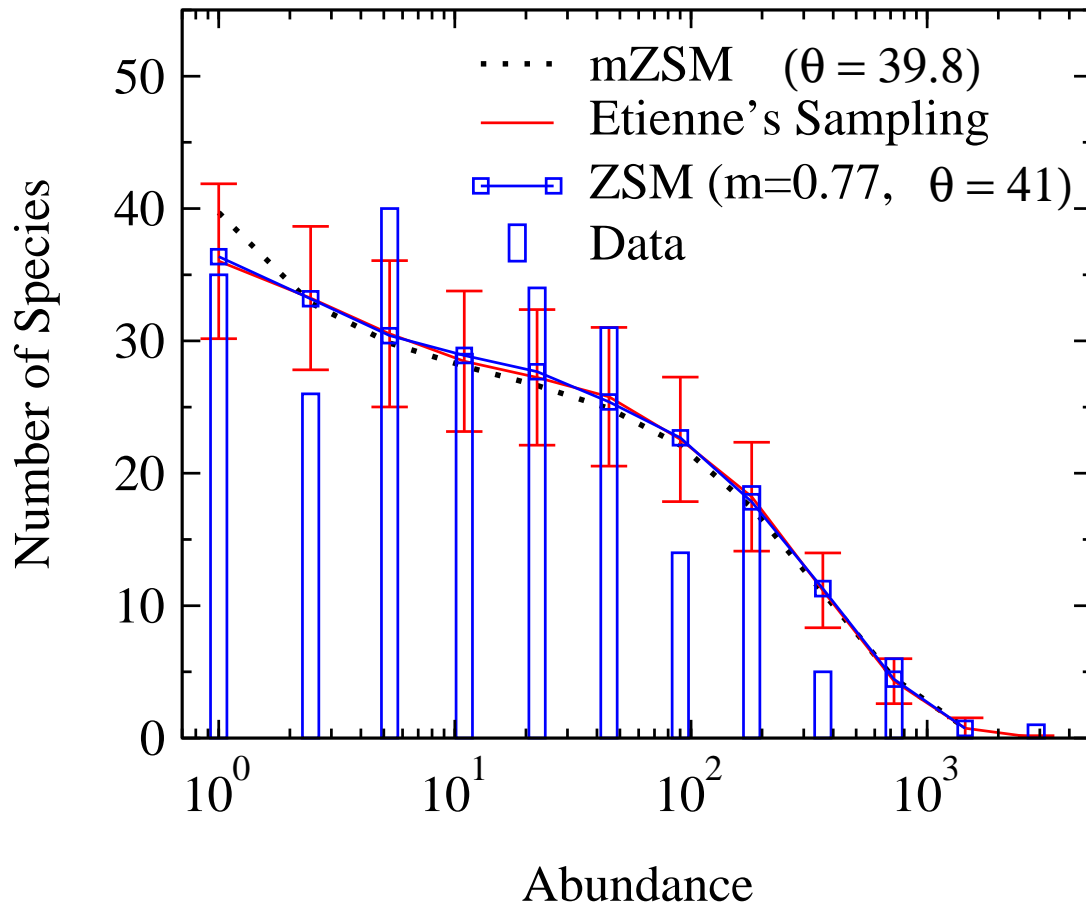


Fig. S1. Abundance data on Lepidoptera in light traps.

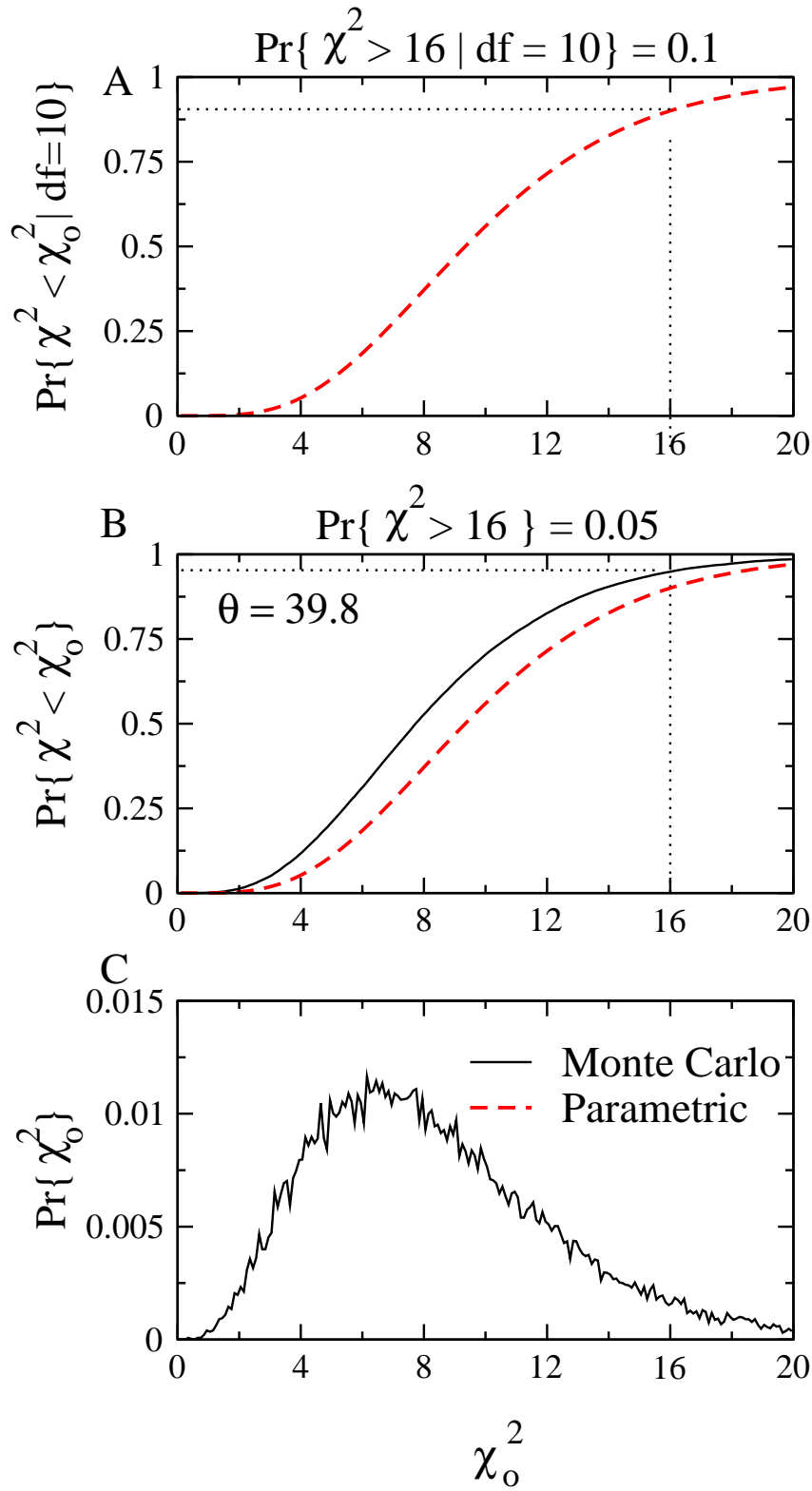


Fig. S2. Lepidoptera Community. Monte Carlo test.