Supporting Online Material  
Manuscript 1115576  

Materials and Methods  

The data for the Cape Floristic Region (CFR) sites comes from 5x10 meter relevé plots, which had been surveyed using standard Braun-Blanquet survey methods: all vascular plant species had been identified to species and assigned a categorical percent cover score (S1). From the complete survey data, “canopy” shrubs and small trees, as well as smaller shrubs and shrublets were selected. These woody plants cover a range of relative sizes roughly equivalent to the range in relative sizes for rain forest canopy and understory trees surveyed by Pitman et al. (S2, S3) and Condit et al. (S4). All CFR plots are located in fynbos regions in areas undisturbed by land use transformation. Species names were standardized using the reference Cape Flora (S5) and, as for the Amazon data (S3), individuals not identified to species were removed.

Since neutral theory describes ecological dynamics at the individual level, it was necessary to convert the areal or cover-based Braun-Blanquet scores for the CFR data to numbers of individuals. We used an independent data set that contains both counts of individuals and cover estimates across a range of typical fynbos plots to calibrate the number of individuals per square meter (S6). The conversion factors were estimated separately for each growth form classification: trees (0.1 individual/m²), large shrubs (0.25 individual/m²), shrubs (4 individuals/m²), and shrublets (8 individuals/m²). Sensitivity of parameter estimates to these conversion factors was assessed, and the results presented are qualitatively unaffected across a reasonable range of values. Despite the difference in methodology (cover-based sampling versus counts of individuals), mean plant size and plot size, the total sampling effort for each sub-regional site was approximately equivalent for the two data sets (see Table 1). Both data sets consist of a set of plots distributed non-randomly across identifiable sub-regions to capture the range of typical vegetation sub-types within the sub-region (S2, S7-9).

Parameter estimates were obtained using a likelihood derived from the closed-form expression for the ZSM:

\[
L\{S_1, \ldots, S_n \mid \Theta, m\} = \prod_{n=1}^{J} \Phi_{n}^{S_n} = \frac{\Theta}{n} \binom{J}{n} \sum_{x=0}^{\gamma_{n} \cdot y_{n}} \frac{\Gamma_{\gamma_n - y_n}}{\Gamma \cdot \Gamma_{\gamma_n + y_n}} e^{-\frac{x}{\gamma_{n}}} dx
\]

where \( \gamma = \frac{(J - 1)m}{1 - m} \) and J is the (known) sample size (SI0-12). In this function, \( \Phi_{n} \) is the (predicted) probability that a species randomly drawn from the local community sample has abundance n, and \( S_n \) is the observed number of species of abundance n (SI0).

Parameter estimates were obtained both by maximizing the likelihood and by Bayesian posterior simulation. Maximum likelihood estimates were obtained through minimization of the negative log-likelihood. Using the full likelihood to fit the model is preferable to earlier efforts that have used \( \chi^2 \) deviance measures on \( \log_2 \) binned abundances (SI1, 13) because this approach uses the full data set rather than summarized data (SI0).

For Bayesian inference, a flat beta prior was placed on \( m \) and an uninformative gamma prior placed on \( \Theta \), with parameters selected so that 95% of the posterior mass of
the prior was below 4,000. Samples were thus drawn from the following (unnormalized) posterior distribution, where $L[S_1,...,S_n | \Theta, m]$ is taken from (1):

$$
P[\Theta, m | S_1,...,S_n, J] \propto L[S_1,...,S_n | \Theta, m] \cdot \text{Gamma}_{(1,0.006)}(\Theta) \cdot \text{Beta}_{(1,1)}(m)$$

(2)

Convergence was assessed through visual inspection of the chains and the Brooks-Gelman-Rubin statistic (S14).

Estimates of $\Theta$ for the metacommunity neutral model with no migration limitation were obtained by maximizing the likelihood of the metacommunity sampling distribution (S10):

$$
S_n = \int_0^1 \exp(-xJ) \frac{x^n(1-x)^{n-1}}{n!} dx
$$

(3)

Comparison between the one and two parameter models were done using a likelihood ratio test, since the models are nested (S10).

Supporting References and Notes: