Chapter 3

A sampling formula for ecological communities with multiple dispersal syndromes

Thijs Janzen, Bart Haegeman, Rampal S. Etienne

Over the past decade, the neutral theory of biodiversity has stirred up community assembly theory considerably by suggesting that stochasticity in the form of ecological drift is an important factor determining community composition and community turnover. The neutral theory assumes that all species within a community are functionally equivalent (the neutrality assumption), and therefore applies best to communities of trophically similar species. Evidently, trophically similar species may still differ in dispersal ability, and therefore may not be completely functionally equivalent. Here we present a new sampling formula that takes into account the partitioning of a community into two guilds that differ in immigration rate. We show that, using this sampling formula, we can accurately detect a subdivision into guilds from species abundance distributions. We apply our sampling formula to tropical tree data from Barro Colorado Island, Panama. Tropical trees are divided depending on their dispersal mode, where biotically dispersed trees are grouped as one guild, and abiotically dispersed trees represent another guild. We find that adding guild structure to the neutral model significantly improves the fit to data and provides a better understanding of community assembly on BCI. Our findings are thus an important step towards an integration of neutral and niche theory.

Journal of Theoretical Biology, in review
Introduction

The astonishing biodiversity around the globe, especially in the tropics, makes one wonder how this biodiversity has originated and how it can be maintained. Traditionally, species composition in an ecological community is explained by species-specific traits and species requirements. By contrast, the more recent neutral theory (Hubbell 2001; Etienne & Ollf 2004; Rosindell et al. 2011) explains species composition in an ecological community by stochastic demography and dispersal. This theory deliberately neglects species-specific differences (the neutrality assumption). It oversimplifies ecology in order to emphasize that ecological drift is an important factor in community assembly (Rosindell et al. 2011; Wennekes et al. 2012). Despite this simplification the model can explain various biodiversity patterns avidly, suggesting that indeed ecological drift is an important factor in community assembly (Etienne & Ollf 2004; Alonso et al. 2006; Rosindell et al. 2011).

The neutrality assumption states that all the individuals within an ecological community have the same birth rates, death rates, dispersal rates and speciation rates, irrespective of the species the individuals belong to (Hubbell 2001). The ecological community is assumed to consist of individuals of functionally equivalent species that compete with each other for space in the community. As a result, patterns in abundance predicted by the theory are purely the result of drift, not the result of competitive asymmetries between the species in the local community. The neutrality assumption is the most debated assumption of the Neutral Theory of Biodiversity (McGill et al. 2006b; Purves & Pacala 2008; Turnbull et al. 2008; Gotelli et al. 2009). Most importantly, the neutrality assumption refutes the idea of the unique correspondence between a species and its niche (interpreted here as the set of conditions and requirements for a species to survive (Hutchinson 1958), although the exact meaning of the niche concept is unclear (McInerny & Etienne 2012a; b; c)). More specifically, the neutrality assumption ignores specific interactions between species and species-specific adaptations, such as habitat specialization; furthermore it ignores the effects of density dependence, ecological succession and the impact of trait differences (Purves & Turnbull 2010).

There have been several attempts to bridge the gap between niche and neutral theory. For example, Purves and Pacala (2005) introduced niche structure into the neutral model. They modelled a community as consisting of four different niches (habitats), where, in the extreme case, each niche behaved as if it was an independent neutral community. They study a range of possible interactions between these niches, such as succession (where replacement is driven through cyclical succession between the guilds), lottery (where the replacing individual belongs randomly to one of the four niches) and habitat specialization (where replacement is dependent on the identity of the local site). They find that none of these three interactions affects patterns in the species abundance distribution, provided that within-niche diversity is sufficiently high. Similarly, Chisholm and Pacala (2010) modelled niches as a part of the local or
metacommunity that contains suitable habitat for one or more species and partitioned the local community into $K$ niches. They found that when they impose strong niche structure (e.g. high values of $K$), their model explains species abundance data equally well as the standard neutral model. High values of $K$ could be interpreted as an extreme form of the model of Purves and Pacala (2005), where there are many habitats that each have low within-niche diversity. However, further analysis of this approach revealed that the independence between niches effectively removes species interactions, thereby introducing neutral-like behaviour, instead of the intended addition of niche structure (Haegeman & Etienne 2011).

Other attempts at exploring the continuum between niche and neutral models explore the effect of differences in birth and death rates, which might arise through differences in intraspecific and interspecific competition. In the fully neutral case, intraspecific and interspecific competition are identical, whereas classic coexistence theory predicts that coexistence is promoted when intraspecific competition is stronger than interspecific competition (Adler et al. 2007). Combining community assembly with classic coexistence modelling, Noble and Fagan (2011) showed that when intraspecific competition exceeds interspecific competition, patterns similar to a fully neutral model emerge. Along similar lines, Haegeman and Loreau (2011) investigated how altering the difference between intraspecific and interspecific competition affects the species abundance distribution. They focused on the parameter space where intraspecific competition exceeds interspecific competition, i.e. where classical theory predicts coexistence. They found that with increasing interspecific competition, fluctuations in local community size increase, and the local community becomes more prone to extinction. More importantly they found that altering the difference between intraspecific and interspecific competition only influenced the species abundance distribution marginally, and concluded that from species abundance data alone it might be difficult to assess the degree of intraspecific versus interspecific competition. Proceeding even further, Pigolotti and Cencini (Pigolotti & Cencini 2013) found an analytical expression for the expected species abundance distribution where the degree of intraspecific and interspecific competition can be tuned by a single parameter. Their results suggest a profound impact of the degree of intraspecific versus interspecific competition not only on the species abundance distribution, but also on the average species lifetime and on the total variation in species lifetimes in the local community.

Competitive asymmetry could also result in differences in birth rate irrespective of competition. Du and colleagues find that introducing competitive asymmetry breaks down neutral patterns (Du et al. 2011), but also find that these effects can be counteracted by negative density dependence: communities with intermediate competitive asymmetry and intermediate levels of negative density dependence show species abundance distributions that are indistinguishable from neutral distributions, suggesting that neutral patterns can emerge from non-neutral assumptions.

Breaking neutrality through the introduction of differences in dispersal rather than birth and death rates has been less well studied. Turnbull and colleagues (Turnbull et
al. 2008) investigated the effect of an equalizing trade-off between seed mass and seed number on neutrality. They found that after including such a trade-off, neutral patterns break down as soon as seed arrival becomes stochastic. Liu and colleagues (2011) relaxed the neutrality assumption by introducing stochastic differences in dispersal ability between species. As the standard deviation of the Gaussian distribution governing these differences increases, the neutral patterns break down and community assembly becomes deterministic, where species with a high dispersal ability tend to dominate the local community. Liu and colleagues compared the effect of differences in dispersal ability to data generated with the neutral model without these differences, but did not confront their model with empirical data.

Trophically similar species may come close to fitting the neutrality assumption, but differences in dispersal may prevent them from being functionally equivalent. Differences in dispersal might arise through differences in seed size (Muller-Landau & Hardesty 2005), differences in fruit size (Seidler & Plotkin 2006) but might also manifest themselves as differences in flight prowess (Valtonen et al. 2013) or differences in pelagic larval duration in coral reef fish (Victor & Wellington 2000; Almany et al. 2007). In this paper we will study such differences in dispersal, focusing on tropical trees. The majority of tropical tree species (73%) disperse through animal means (Muller-Landau & Hardesty 2005), such as bats, birds, mammals, ants and sometimes even fish. The other 27% of tree species relies on abiotic factors to disperse their seeds, such as wind, water or ballistics.

Current neutral models fail to include differences in dispersal between species that share the same local community and metacommunity. Here we present a model where we classify species according to their dispersal syndrome. We will call the resulting classes guilds. This is a simple, but important step towards incorporating differences between species without needing to explicitly quantify these differences for every species in the community. Instead we only need to quantify the differences between guilds, and assess the importance of these differences for community assembly. Our model differs from the standard neutral model (Hubbell 2001; Etienne & Alonso 2005) by subdividing the community into two guilds, where each guild is a group of species that have the same dispersal rate. Between guilds, dispersal rates may differ, but the speciation rate, birth and death rates are identical. We show that our model can accurately distinguish between datasets including a guild structure, and datasets that do not have any guild structure. Our model is able to detect signatures of guild structure from the species abundance distribution, refuting the idea that the species abundance distribution does not contain sufficient information to draw conclusions about underlying community assembly mechanisms. Secondly, we show that parameter estimates obtained with our model are accurate and differ considerably from estimates obtained using the standard neutral model without guild structure. Lastly we illustrate the model by applying it to the tropical tree dataset of Barro Colorado Island (BCI).
Model

We assume that there are two guilds $X$ and $Y$ that differ in their immigration parameter $m_i$ ($i = X,Y$); all species within each guild share the same migration parameter $m_i$. All species, regardless of the guild they belong to, have the same fundamental biodiversity number $\theta$, as in the standard neutral model. In the metacommunity, every time step one individual dies and is replaced by an individual from either guild $X$ or guild $Y$. With probability $\nu$ a speciation event occurs resulting in a new species that belongs to guild $X$ with probability 0.5, and to guild $Y$ with probability 0.5. With probability $1 - \nu$ no speciation event occurs; then the new individual belongs to guild $X$ or $Y$ depending on the relative abundance of guilds $X$ and $Y$ in the metacommunity. Over time the relative abundances of both guilds reach a dynamical equilibrium.

In the local community, a deceased individual can be replaced by either an individual from the local community with probability $1 - m_X - m_Y$ or by an immigrant from the metacommunity; this is an individual from guild $X$ with probability $m_X$ or an individual of guild $Y$ with probability $m_Y$. The migration probability $m_i$ of guild $i$ depends on the dispersal ability of guild $i$, here called $\alpha_i$, and the relative abundance of the guild in the metacommunity, $p_i$. The migration probability of guild $i$ is then given by $m_i = \alpha_i p_i$. The dispersal ability $\alpha_i$ is bounded between [0,1] where values close to zero indicate low dispersal ability and values close to one indicate good dispersal ability.

Sampling formula

In the case of a single guild, fitting the neutral community model to data makes use of the Etienne sampling formula (Etienne 2005), which is a dispersal-limited extension of the Ewens sampling formula (Ewens 1972). This formula gives the probability of a data set of species abundances in a sample as a function of the model parameters. Here we briefly describe the extension of the neutral sampling formula to the case of two guilds and we provide a more detailed derivation in the Appendix.

The stationary abundance distribution of the neutral community depends on the fundamental biodiversity number $\theta$ and on the fundamental dispersal number $I$, which are defined as

$$\theta = \frac{\nu(J-1)}{1-\nu} \quad \text{and} \quad I = \frac{m(J-1)}{1-m}$$

where $\nu$ is the speciation rate, $J_M$ is the metacommunity size, $J$ is the local community size, and $m$ is the migration probability. Likewise, the stationary abundance distribution of the two-guild neutral community can be expressed in terms of guild-specific biodiversity numbers $\theta_X$ and $\theta_Y$ and guild-specific dispersal numbers $I_X$ and $I_Y$ (see Appendix S4A and S4B). We define speciation to be equally likely between guilds, such that half of the speciation events result in a new species of guild $X$ and half of the speciation events result in a new species of guild $Y$, we obtain: $\theta_X = \theta_Y = \theta/2$. We define migration from the metacommunity to the local community as the product of dispersal ability and the relative frequency of the guild in the metacommunity, such that for guild
i: \( m_i = \alpha_i p_i \) where \( \alpha_i \) is the dispersal ability of guild \( i \) and \( p_i \) is the relative frequency of guild \( i \) in the metacommunity. For the dispersal numbers, using the guild-specific immigration probabilities \( m_i = \alpha_i p_i \) we obtain \( I_X \) and \( I_Y \):

\[
I_X = \frac{\alpha_X p_X (J-1)}{1-\alpha_X p_X - \alpha_Y p_Y} \quad \text{and} \quad I_Y = \frac{\alpha_Y p_Y (J-1)}{1-\alpha_X p_X - \alpha_Y p_Y}
\]

(1)

Note that \( \theta = \theta_X + \theta_Y \) and \( I = I_X + I_Y \), that is, the speciation and immigration processes are split out over the two guilds. Using the guild-specific biodiversity and dispersal numbers, the two guilds abundance distribution is (see equation (S8)):

\[
\]

(2)

where vector \( D_X \) contains the species abundances in guild \( X \) and vector \( D_Y \) contains the species abundances in guild \( Y \). The second factor in the right-hand side, \( P(D_X | \theta_X, I_X, J_X) \), is the one-guild Etienne sampling formula of guild \( X \) as if it was isolated from guild \( Y \) (but with the appropriate biodiversity and dispersal number). The third factor in the right-hand side, \( P(D_Y | \theta_Y, I_Y, J_Y) \), is the Etienne sampling formula of guild \( Y \), as if it was isolated from guild \( X \). The two isolated guild abundance distributions are combined through the probability distribution of the guild sizes (see equation (S6)),

\[
P(J_X, J_Y | I_X, I_Y, J) = \frac{J!}{(I_X + I_Y)!} \frac{(I_X J_X)(I_Y J_Y)}{I_X! J_X! I_Y! J_Y!}
\]

(3)

Hence, as far as the stationary abundance distributions are concerned, the dependence between guilds is concentrated in the guild sizes. In other words, after conditioning the abundance distribution on the guild sizes, and for given values of \( I_X \) and \( I_Y \), the abundance distributions of guild \( X \) and guild \( Y \) are independent. This shows that the one-guild abundance distributions are the fundamental building blocks of the abundance distribution of a community consisting of two (or more) neutral guilds.

Equation (2) is not yet the full sampling formula, because \( I_X \) and \( I_Y \) depend on \( p_X \) and \( p_Y \) (see equation (1)), which are variables, not parameters. The distribution of the metacommunity guild sizes \( p_X \) and \( p_Y = 1 - p_X \) is a beta distribution (see equation (S10)),

\[
\rho(p_X | \theta) = \frac{(I(\theta))}{I(\theta)^2} \left[ \frac{\theta}{2} - 1 \right]^{-1} \left[ \frac{1 - \theta}{2} - 1 \right]^{-1}
\]

(4)

Hence, we find the full two-guild sampling formula by integrating over all possible values of \( p_X \),

\[
P(D_X, D_Y | \theta, \alpha_X, \alpha_Y) = \int_0^1 P(J_X, J_Y) P(D_X | \theta, I_X, J_X) P(D_Y | \theta, I_Y, J_Y) \rho(p_X | \theta) dp_X
\]

(5)

Code to calculate this sampling formula for a dataset is available in the GUILDS package for R.
Conditioning on guild size

Using equation (3), we can calculate the expected guild sizes, given $\alpha_x$, $\alpha_y$ and $J$:

\[
E(J_X|\alpha_x, \alpha_y, J) = \frac{(J_X)}{(I_x + I_y)} = \frac{J\alpha_x p_x}{\alpha_x p_x + \alpha_y p_y} \approx \frac{J\alpha_x}{\alpha_x + \alpha_y}
\]

\[
E(J_Y|\alpha_x, \alpha_y, J) = \frac{(J_Y)}{(I_x + I_y)} = \frac{J\alpha_y p_y}{\alpha_x p_x + \alpha_y p_y} \approx \frac{J\alpha_y}{\alpha_x + \alpha_y}
\]

(6a)

(6b)

From the expected guild sizes it follows that the ratio of guild sizes is equal to the ratio of dispersal rates: $E[J_X] / E[J_Y] = \alpha_x/\alpha_y$. Explorations of the sampling formula confirmed that our estimated values for the dispersal parameter closely mimic the ratio of guild sizes (figure A1). This is also intuitively understandable, consider two guilds, with one guild being twice as large as the other guild (i.e. $J_X = 2J_Y$). In order to reach such a skewed distribution of individuals, either this distribution is already present in the metacommunity, or there is a large skew in dispersal ability. The beta distribution we assume in the metacommunity (equation (4)) does allow for some divergence from a 50/50 distribution, but on average we do not expect the metacommunity to be highly skewed towards one particular guild. As a result, we expect that given a dataset with differently sized guilds, our sampling formula will estimate differences in dispersal ability and hence assume some form of guild structure, even when these differences in guild size are not caused by guild effects. We have circumvented this problem by conditioning our sampling formula on guild sizes. This yields the probability of our data given the parameter values and the guild sizes. As a result, differences in parameters, and any detected guild effects, are independent of guild size and solely dependent on differences in dispersal ability.

The joint probability of a combination of guild sizes is given by (see appendix equation (S12)):

\[
P(J_X, J_Y|\theta, \alpha_x, \alpha_y, J) = \int_0^1 P(J_X, J_Y|I_X, I_Y, J)\rho(p_X|\theta)dp_X
\]

(7)

We condition by dividing our sampling formula by this likelihood, and thus obtain:

\[
P(D_X, D_Y|\theta, \alpha_x, \alpha_y, J_X, J_Y) = \frac{\int_0^1 P(J_X, J_Y)P(D_X|\theta, \alpha_x, J_X)P(D_Y|\theta, \alpha_y, J_Y)\rho(p_X|1-p_X)dp_X}{\int_0^1 P(J_X, J_Y|I_X, I_Y, J)\rho(p_X|\theta)dp_X}
\]

(8)

The conditioned sampling formula no longer results in a relation between guild size and estimated dispersal ability (figure A1). Code to compute the conditioned sampling formula for a dataset is available in the GUILDS package for R.

Testing on artificial data

The two-guilds sampling formula can be reduced to the Etienne sampling formula of the standard, single-guild, neutral model by setting the dispersal ability of both guilds to the same value ($\alpha_x = \alpha_y$). Throughout the text we will refer to this model as $D0$. As an
alternative model, we allow the dispersal rates to differ between guilds ($\alpha_X \neq \alpha_Y$). We will refer to this model as $D1$. To assess how well we can distinguish the two models from each other, we generated 100 replicate datasets for all unique combinations of $\theta = \{30, 100, 300\}$ and $\alpha = \{0.001, 0.01, 0.1\}$. There are 9 different combinations for $D0$ (all combinations of $\theta$ and $\alpha$, $3 \times 3$), and 9 different combinations for $D1$ (3 different $\theta$ values with one of three $\alpha$ combinations: $[0.001, 0.01], [0.001, 0.1]$ or $[0.1, 0.1]$, which again yields $3 \times 3$ combinations). Community size was set at 20,000 individuals. We generated artificial datasets using a three-step procedure: first, the sizes of guilds $X$ and $Y$ in the metacommunity (assuming that the metacommunity size is infinite), were drawn from a beta distribution with parameter $\theta$ (equation (4)). Secondly, the total number of individuals ($J_i$) of each guild $i$ in the local community was drawn from equation (2) with parameters $J$, $I_i$. The species abundance distribution of each guild was then generated using the urn scheme as described in Etienne (2005) with parameters $J$ and $I$. Code to generate a local community according to the aforementioned procedure is available in the GUILDS package for R.

For every artificial dataset we performed maximum likelihood estimation for the two models ($D0, D1$), where the likelihood maximization was started at the parameter values used to generate the data with. The obtained likelihood values for the maximum likelihood optimum were used to calculate the Akaike Information Criterion (AIC) (Akaike 1974).

$$AIC = 2k - 2\ln(L)$$

where $k$ is the degrees of freedom of the model, and $L$ is the Maximum Likelihood of the model. The number of degrees of freedom $k$ depends on the number of free parameters in the model, which is 2 for the $D0$ model, ($\theta$ and $\alpha$) and 3 for the $D1$ model ($\theta, \alpha_X$, and $\alpha_Y$). After calculation of our AIC values, we compared the AIC scores with AIC weights (Wagenmakers & Farrell 2004):

$$w_i(AIC) = \frac{\exp(-\frac{1}{2} \Delta_i AIC)}{\sum_{k=1}^{K} \exp(-\frac{1}{2} \Delta_k AIC)}$$

where $\Delta_i AIC = AIC_i - \min(AIC)$, and $K$ is the total number of models compared (in this case, 2). AIC weight $w_i$ can be interpreted as the probability of model $i$ being the best model among the models considered.

To assess the accuracy of our parameter estimates we performed Maximum Likelihood estimation for the same simulated communities, but now starting at a grid of $2^d$ initial parameter combinations (with $d$ being the number of free parameters in the model, 2 for $D0$, 3 for $D1$), not necessarily including the values used to generate the data. The initial values contained all possible combinations for $\theta$ of $[30,300]$ and $\alpha$ of $[0.001, 0.1]$. Using the 100 obtained Maximum Likelihood estimates we calculated the 25th, 50th and 75th percentiles.
Empirical data
To illustrate the application of our sampling formula, we performed both model selection and parameter estimation techniques on a well-studied dataset of tropical forest trees: the Neotropical community dataset of Barro Colorado Island (BCI), Panama (Condit et al. 1996, 2002; Hubbell 2001; Volkov et al. 2003; Etienne 2005). The dataset consists of the abundance of all free-standing woody plants with >10 cm diameter at breast height in 50 ha of forest. We analyzed censuses from 1982, 1985, 1990, 1995, 2000 and 2005. The resulting dataset consists of recorded abundances of 6 different years, 252 woody plant species, with a summed total over 20,000 individuals per census. Tree species in this data set are grouped according to their dispersal syndrome, where all biotically dispersed (i.e. via birds (171 species), bats (37 species) and mammals (194 species)) trees are grouped together in one guild, and all abiotically dispersed trees (i.e. wind (33 species), water (1 species), ballistic means (10 species)) are grouped together in another guild (Muller-Landau & Hardesty 2005).

Posterior analysis
To elucidate the differences between the models, we calculated the expected species abundance distribution for every dataset, using a hybrid approach of simulation and exact calculation. The expected species abundance distribution was approximated as follows: Given \(J_X\) and \(J_Y\) we obtained \(I_X\) and \(I_Y\) by first drawing \(p_X\) from:

\[
P(p_X|\theta, \alpha_X, \alpha_Y, J_X, J_Y) = \frac{P(J_X, J_Y| I_X, I_Y, J_X, J_Y)\rho(p_X|\theta)}{\int_0^1 P(J_X, J_Y| I_X, I_Y, J_X, J_Y)\rho(p_X|\theta)dp_X}
\]

and then calculated \(I_X\) and \(I_Y\) using equation (1). We then calculated the expected number of species in guild \(i\) with \(n\) individuals using equation 6 from Etienne and Alonso (2005):

\[
E[S_n| \theta, I, J] = \frac{\theta}{(I)_j} \int (I)_n[I(1-x)]_{j-n} \frac{(1-x)^{\theta-1}}{x} dx
\]

Because drawing from the distribution in equation (9) is inherently stochastic, we averaged over 100 replicates to obtain the final expected abundance distribution.

Furthermore we studied the power of the imposed guild structure on the data, i.e. we determined whether adding guild structure to the data adds information. We did this by randomizing the datasets 100 times, by randomly assigning species to a guild (with equal probability), thus removing any guild structure. For every randomized dataset we then used Maximum Likelihood Estimation with both models, initialized at a grid of initial values. Using AIC weights we estimated which of the two models best explained the data. If the imposed guild structure provides additional information on the dataset, we would expect that after randomization any signals of guild structure are lost, and the \(D0\) model is favored. Conversely, if any random subdivision into two groups would also cause detection of guild structure and the \(D1\) model would be favored after randomization the detection of guild structure in the original data set is ecologically meaningless.
Lastly we used our sampling formula to evaluate the goodness-of-fit of the neutral model with guild structure to empirical data, by performing an ‘Exact’ test of neutrality (Etienne 2007). Using the parameter estimations obtained with Maximum Likelihood Estimation, we generated 100 different data sets (using the maximum likelihood estimates for \( \theta \) & \( \alpha \)). Datasets were generated by first drawing \( p_X \) from equation (9), then using \( f_X \) and \( f_Y \) from the data and our obtained maximum likelihood estimates for \( \theta \) and \( \alpha \) we generated the species abundance distribution for each guild using the urn scheme as described in Etienne (2005). For these 100 replicate datasets we then again calculated the model parameters and likelihood using Maximum Likelihood Estimation as described in the previous section (this is different from Etienne 2007 who did not maximize the likelihood – the procedure used here is less conservative (Efron & Tibshirani 1994)). If the likelihood of the empirical data is smaller than the obtained frequency distribution of likelihood from the replicate datasets it is unlikely that neutral processes were responsible for the observed community. If however the likelihood of the empirical data is not different from the obtained frequency distribution, neutrality cannot be rejected as a process responsible for the observed abundance distribution.

**Results**

The ability to accurately select the correct model is essential for the implementation of our sampling formula. Not only should our sampling formula favor the more complex model when the data warrants it, it should also reject the more complex model if the data shows no sign of guild structure. We tested the ability of the sampling formula to detect guild structure by confronting it with artificially generated data. The artificial data contained either no guild structure at all, or was generated including different degrees of difference in dispersal limitation between guilds. Data generated using the \( D0 \) model (no guild structure), was correctly identified as having no guild structure in 88% of all simulated datasets (794 out of 900 datasets were correctly identified as \( D0 \)). Data generated using the \( D1 \) model was correctly identified as having guild structure in 85% of all simulated datasets (765 out of 900 datasets were correctly identified as \( D1 \)). Hence, type I (12%) and Type II (15%) errors are very similar, and the model adequately detects guild structure in the majority of the simulated datasets we analyzed.

Using artificial data generated with either the \( D0 \) or the \( D1 \) model, we tested the precision and bias of the new guilds sampling formula. We report the 25th, 50th and 75th percentiles of 100 replicates (Table 3.1). For the \( D0 \) model, the parameter value used to generate the data fell between the 25th and 75th percentiles for 8 out of 9 parameter combinations. For the \( D1 \) model, the parameter values used to generate the data all fell between the 25th and 75th percentiles. The bias of the \( D0 \) model was small: the 50th percentiles of maximum likelihood estimates for datasets simulated with high
Table 3.1. Bias and precision of the maximum-likelihood estimates as shown by the median and the 25\textsuperscript{th} and 75\textsuperscript{th} percentiles of the estimated parameter values of 100 simulated data sets per parameter combination.

<table>
<thead>
<tr>
<th>Parameters used to generate data</th>
<th>Estimated parameter values (25\textsuperscript{th}, 50 and 75\textsuperscript{th} percentiles)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>\theta</td>
</tr>
<tr>
<td>-------</td>
<td>--------</td>
</tr>
<tr>
<td>D0</td>
<td>30</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>100</td>
</tr>
<tr>
<td></td>
<td>100</td>
</tr>
<tr>
<td></td>
<td>300</td>
</tr>
<tr>
<td></td>
<td>300</td>
</tr>
<tr>
<td>D1</td>
<td>30</td>
</tr>
<tr>
<td></td>
<td>30</td>
</tr>
<tr>
<td></td>
<td>100</td>
</tr>
<tr>
<td></td>
<td>100</td>
</tr>
<tr>
<td></td>
<td>100</td>
</tr>
<tr>
<td></td>
<td>300</td>
</tr>
<tr>
<td></td>
<td>300</td>
</tr>
</tbody>
</table>
dispersal values ($\alpha > 0.001$) were close to those used to simulate the datasets. Combinations with low dispersal ($\alpha = 0.001$) tended to have a median slightly underestimating $\theta$, but an accurate estimate of $\alpha_X$, except for the combination [300,0.001], for which none of the percentiles included the correct $\alpha_X$ value. Precision of the $D0$ model was high, with the overall spread of estimated parameter values closely clustered around the median value, with a notable exception for the combination [300,0.001], where estimates for $\alpha_X$ have a large spread. The $D1$ model had a similarly low bias as the $D0$ model and median estimates were close to parameters used to generate the data. Precision of the $D0$ model was high, with the 25th and 75th percentile generally close to each other, except for one combination: [30, 0.1, 0.01], where the 75th percentile of the estimate for $\alpha_X$ was 1.

For the empirical dataset, the $D1$ model had a much higher likelihood than the $D0$ model, for all censuses. After penalizing the likelihood for added complexity and calculating the corresponding AIC score and AIC weights, the $D1$ model was convincingly selected. Considering the ability of our sampling formula to detect or reject guild structure in artificial datasets, we conclude that using our guild sampling formula we have convincingly detected guild structure based on dispersal syndromes for all six BCI censuses.

For the six BCI censuses, we found parameter estimates using both the $D0$ and the $D1$ model (Table 3.2). The $D0$ model has been shown to have two competing optima (Etienne et al. 2006), one with a high value for $\theta$ and low value for $\alpha$ and another with a low value for $\theta$ and a high value for $\alpha$. For the $D1$ model we found two competing optima as well. One of the two optima combines high diversity with high dispersal limitation and typically has a high $\theta$ value ($\sim200$), combined with low $\alpha$ values ($\sim0.005$ and $\sim0.0008$), whilst the other optimum combines low diversity with low dispersal limitation and has a lower $\theta$ value ($\sim53$) combined with one extreme $\alpha$ value (of 1.0) and one much lower $\alpha$ value ($\sim0.0006$).

For three out of six BCI censuses, the extreme dispersal optimum is favored over the high diversity, high dispersal limitation optimum. For the three BCI censuses where the extreme dispersal optimum was not favored, the dispersal ability of the guild that relies on biotic dispersal is 5.74 (standard deviation 0.17) times higher than the dispersal ability of the guild that relies on abiotic dispersal.

We performed the ‘exact’ test of neutrality as described in the methods section, to estimate whether the observed data is the result of a guild structured neutral process, or whether perhaps the observed data is the result of a different process. All optima for the $D1$ model have non-significant $p$-values for the ‘exact’ test of neutrality (Table 3.2), and hence we cannot distinguish patterns in these communities from those generated with our model with two neutral guilds. All optima for the $D0$ model have significant $p$-values, except for the census of 1990, which has $p$-values of 0.05 and 0.06, which are barely non-significant. Thus, the empirical data are unlikely to be generated with the neutral model without guild structure. The combined results for the $D0$ and $D1$ models therefore strongly suggest that guild structure is an important aspect of the empirical data.
Table 3.2. Parameter estimates for six different censuses of Barro Colorado Island. The \( D0 \) model does not take into account differences in dispersal between the guilds, the \( D1 \) model does take these differences into account. Guild \( X \) represents tree species with biotic dispersal, and guild \( Y \) represents tree species with abiotic dispersal. The \( p \)-value of the ‘exact’ test of neutrality is reported in the last column.

<table>
<thead>
<tr>
<th>Census</th>
<th>General statistics</th>
<th>Parameter estimates</th>
<th>Model fit</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>( J )</td>
<td>( J_X )</td>
<td>( J_Y )</td>
</tr>
<tr>
<td>1982</td>
<td>20914</td>
<td>18321</td>
<td>2593</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1985</td>
<td>20742</td>
<td>18203</td>
<td>2539</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1990</td>
<td>21253</td>
<td>18641</td>
<td>2612</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1995</td>
<td>21460</td>
<td>18812</td>
<td>2648</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2000</td>
<td>21205</td>
<td>18607</td>
<td>2598</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2005</td>
<td>20860</td>
<td>18321</td>
<td>2539</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Randomization tests revealed that for all six datasets, randomization removed any signal of guild structure (Figure 3.1). AIC weight was higher for the D0 model than for the D1 model for all 100 replicates for censuses 1985, 1990, 1995, 2000 and 2005. The 1982 census had 90 out of 100 replicates for which the AIC weight of the D0 model was larger than the AIC weight of the D1 model, retaining a guild signal after randomization in 10% of the replicates. Together these results indicate that the subdivision based on ecological data conveys more information than a random subdivision in guilds does.

For all datasets we plotted the empirical species abundance distribution versus the expected abundance distribution under the Maximum Likelihood Estimates (Figure 3.2). For all datasets we observe that the D0 model tends to underestimate abundances for guild X, whilst overestimating abundances for guild Y. It appears that, in an attempt to fit best to both guilds, neither of them is fitted well, which explains the poor performance of the D0 model on our empirical data. For the D1 model we observe that for both the high diversity, high dispersal limitation optima (1982, 1985, 200) and the low diversity, low dispersal limitation optima (1990, 1995, 2000) expected abundance distributions closely match the empirical data (Figure 3.2). For guild Y both optima show similar patterns, whereas for the larger guild X, the high diversity, high dispersal limitation optima tend to expect a higher number of rare species than the low diversity, low dispersal limitation optimum.
Figure 3.2. Empirical (grey bars) and expected (solid curve: D1 model, dashed curve: D0 model) species abundance plots for all BCI censuses. Histograms on the left hand side represent the biotically dispersing guild, histograms on the right hand side the abiotically dispersing guild. Abundances are binned in log2 bins.
Discussion

In this paper we have presented a novel sampling formula that extends the neutral model to a non-neutral setting of two guilds with different dispersal modes. The purpose of our sampling formula is two-fold: 1) to assess whether a subdivision into two guilds, based on ecological information regarding dispersal, amounts to a significant difference in community structure and 2) to illustrate how to determine, for empirical data sets, to what extent the two guilds differ in their dispersal ability.

Using simulated data we have shown that our sampling formula can detect guild structure from data generated including guild structure and reject guild structure when guild structure was not imposed on the simulated data. Furthermore, the simulation results showed parameter estimates obtained using our sampling formula to be unbiased (i.e. close to the parameters used to generate the data), and precision of our parameter estimates was generally high (i.e. spread in parameter estimates was low).

Our new guilds sampling formula allowed us to conclude that for all six censuses of tropical tree communities in BCI, Panama, inclusion of guild structure was favoured and tree species relying on biotic dispersal tend to be less dispersal limited than tree species relying on abiotic dispersal.

It has been suggested that the species abundance distribution contains insufficient information to distinguish between competing models (Cohen 1968; Mcgill 2003; McGill et al. 2006b, 2007; Ricklefs 2006), and that additional data are needed to test validity of community assembly models, for instance in the form of phylogenetic diversity or spatial abundance patterns (McGill et al. 2006b; Jabot & Chave 2009). Here we show that we can distinguish between competing models, using the species abundance distribution combined with ecological information about dispersal mode. However, including information on dispersal and guild structure does not resolve the multiple optima problem of the Etienne Sampling Formula. The Etienne Sampling Formula can potentially yield multiple optima with similar likelihood values. Situated at opposite ends of the parameter continuum, one optimum is typically associated with a high value for $\theta$ and a low value for $l$ (or $m$) and the other optimum with a low value for $\theta$ and a high value for $l$ (or $m$). Additional information about the local community tends to favour one of these two optima. Jabot and colleagues (2009) combined abundance data and phylogenetic data within an approximate Bayesian framework and recovered only one optimum, with high $\theta$ and low $l$. In another approach, Etienne (2007) combined information on multiple local communities to obtain estimates for the neutral model and also found only a single optimum. Here we have included information on guild structure, based on ecological information about dispersal, and recover two competing optima. However, one of these optima seems to be a mathematical abnormality, which is situated at the very limit of parameter space. Parameter estimates for this optimum reflect limited diversity, but extremely low dispersal limitation for one guild ($\alpha = 1$) and high dispersal limitation for the other guild ($\alpha < 0.001$). Expected abundance distributions for these extreme dispersal optima seem to reflect
the empirical abundance distributions well, although the ecological interpretation of unlimited dispersal remains problematic.

Parameter estimates for the Tropical Tree datasets from BCI suggest high values for $\theta$ (average value of 215.42) and low values for $\alpha$ (average values of 0.0050 and 0.00077 for the biotic and abiotic dispersing guilds respectively, ignoring the optima with extreme $\alpha$ values), implying that the tropical tree ecosystem in BCI is highly diverse and fairly dispersal limited. The guild that relies on biotic dispersal (e.g. through birds, bats and mammals) consistently has a higher estimated dispersal ability than the guild relying on abiotic dispersal, such as dispersal through ballistics, gravity, wind and water. This is in line with previous findings (Muller-Landau & Hardesty 2005; Seidler & Plotkin 2006), where animal dispersed trees on average dispersed further than wind dispersed trees. Although wind dispersed trees could potentially disperse over long distances, the tight canopy of tropical forests restricts air movement and generally abiotically dispersed trees tend to disperse over shorter distances than animal dispersed trees (Seidler & Plotkin 2006; Beaudrot et al. 2013).

Our current subdivision in guilds has lumped together trees with fairly different modes of dispersal; we have for instance lumped tree species dispersed by birds as well as tree species dispersed by small mammals in the same guild (biotically dispersed). We expect however that these differences will be less important than the differences between guilds, that is, differences in dispersal between those species relying on animals and species that rely on wind, water or ballistic means. Extending the sampling formula towards more than two guilds is fairly straightforward, but it remains questionable whether this will yield additional understanding of the system. We expect that a larger total sample size is needed to reveal differences in dispersal ability with an increased number of guilds.

An important question that automatically arises when looking at guild structured data is whether the suggested dichotomy introduces more information and structure to the data than a random subdivision into two guilds would. This would quantify the importance of including guild structure in the analysis of community assembly. In our analysis we have tried to approach this question by randomly assigning species to a guild, and assessing which model best explains the (now randomized) data. We found that after randomization the signal of guild structure was almost always lost. This randomization, however, requires making an a priori choice about how to divide species over guilds (either 50/50 or some other distribution). An alternative to randomly assigning species to different guilds would be to randomly assign individuals to different guilds (whilst keeping the total number of individuals per guild constant). The number of individuals and number of species are tightly linked however, and it appears non-trivial how to correctly assign species to the randomized individuals without assigning the same species label to individuals in both guilds. Ultimately, validating the guild structure thus lies not so much in finding a randomization that can test the added value of the imposed guild structure, but in validating the ecological causes that determine why species belong to different guilds.
In a recent paper, Humphreys and Barraclough (2014) also considered a metacommunity divided into multiple “guilds”, and studied the effect of differences in dispersal. Dispersal in their model is not defined as dispersal between a local community and a metacommunity, but rather defined as the connectivity between the two guilds – dispersal represents here the probability of a species from one guild to disperse towards the other guild. This would be analogous to a speciation event of a species from one guild speciating into a species from another guild in our model. Humphreys and Barraclough focus on the emergence of higher Evolutionary Significant Units (hESU’s) as the result of a lack of dispersal between guilds and show that when the exchange of species between guilds is low, this leads to a clear phylogenetic pattern, where both guilds cluster into two distinct clades separated by long external branches. In our sampling formula we have chosen not to focus on speciation dynamics in the metacommunity in favour of unraveling the effects of differences in dispersal. It would be very interesting to look into a guild structured model where both within-guild speciation (e.g. an individual of guild X speciates into a new species belonging to guild X) and between-guild speciation (e.g. an individual of guild X speciates into a new species belonging to guild Y) is modeled. However, because this would introduce at least three new parameters to estimate, we doubt whether such a large number of parameters can be accurately estimated using species abundance data and information on guild structure; perhaps this requires the inclusion of additional information about phylogeny.

Our sampling formula resembles the multiple samples sampling formula presented by Etienne (2007). That sampling formula considers multiple local communities with independent migration, which all share the same metacommunity (with one single estimate for \( \theta \)). If we interpret these different local communities as different guilds, the multiple samples model closely resembles our multiple guilds model. An importance difference, however, is hidden in the metacommunity structure. The multiple samples metacommunity consists of one single metacommunity, without any structure. Our multiple guilds metacommunity is explicitly structured such that there are two separate guilds in the metacommunity that have independent dispersal towards the local community. Due to their independent dispersal, guild sizes and number of species in the local community can differ from each other, whereas the linked local communities from the multiple samples model all sample from the same species pool.

Our guilds sampling formula disentangles migration, dispersal ability, and metacommunity abundance. In classical neutral theory, dispersal limitation between the local and metacommunity is governed by one single parameter, \( m \) (migration) (Hubbell 2001). This can be interpreted as the combined effects of dispersal, recruitment, establishment and metacommunity abundance. In our sampling formula we made the link between dispersal and metacommunity abundance more explicit by defining migration as the product of dispersal and relative metacommunity abundance: \( m_i = \alpha_i p_i \); our newly defined dispersal ability \( \alpha \) still includes dispersal, recruitment and establishment. Because we have redefined the migration parameter, and have focused on estimating \( \alpha \), estimates of our model cannot be directly compared with previously obtained.
estimates of immigration (Etienne 2007; Jabot & Chave 2009). Inferences with our D0 model however, provide a good reference point, as this model assumes no guild structure and reduces to the Etienne Sampling Formula (Etienne 2005) with the migration parameter substituted by our new dispersal parameter and the relative metacommunity abundance. Although our sampling formula is not the first to acknowledge the influence of metacommunity abundance on migration (Rosindell & Harmon 2013), our sampling formula is the first to focus on the effect of dispersal itself and to disentangle dispersal from metacommunity abundance.

In our model we assumed independence of migration and speciation ability. We assumed that in the metacommunity there are no differences between guilds with respect to speciation and have only focused on differences of migration between the metacommunity and local community. If there are profound differences in dispersal ability between guilds, however, we would expect this to also influence the probability of speciation. This is a general problem of two-scale neutral models (Leigh 2007). A lack of dispersal generally tends to lead to a more patchy distribution and can facilitate geographical isolation of populations. As a result we expect an interaction between speciation and dispersal ability. Correctly implementing this interaction would require extending our current model towards a spatially explicit form. Rosindell and Phillimore (Rosindell & Phillimore 2011) took a first step towards a further integration of dispersal and speciation by identifying the difference between in situ speciation on an island (cladogenesis) and speciation through drift over time, where an immigrant on an island diverges from its ancestor on the mainland (anagenesis). Future work could focus on a more direct connection between dispersal and speciation and could provide a more explicit link between spatially explicit processes driving both dispersal limitation and speciation.

Our new sampling formula is a first step towards incorporating ecological reality into a neutral approach of community assembly. It enhances our understanding and appreciation of the interplay between stochasticity, dispersal and species specific requirements that govern the patterns we observe in ecological communities and the underlying processes of community assembly.

Acknowledgements

We thank Joe Wright, Helene Muller-Landau and Denise Hardesty for help with categorizing BCI tree species according to dispersal mode. Financial support for BH was provided by the TULIP Laboratory of Excellence (ANR-10-LABX-41). TJ and RSE thank the Netherlands Organisation for Scientific Research (NWO) for financial support. Computer code for this work has been made available as the R package “GUILDS”.
Table A1. Parameter estimates for the six different censuses of BCI, including both found optima for the D0 model. Guild X represents tree species with biotic dispersal, and guild Y represents tree species with abiotic dispersal. The p-value of the Neutrality test is reported in the last column.

<table>
<thead>
<tr>
<th>Census</th>
<th>J</th>
<th>J_X</th>
<th>J_Y</th>
<th>S_X</th>
<th>S_Y</th>
<th>Model</th>
<th>( \theta )</th>
<th>( \sigma_X )</th>
<th>( \sigma_Y )</th>
<th>LL</th>
<th>AIC</th>
<th>( \Delta \text{AIC} )</th>
<th>AIC W</th>
<th>p value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1982</td>
<td>20914</td>
<td>18321</td>
<td>2593</td>
<td>196</td>
<td>46</td>
<td>D0</td>
<td>111.04</td>
<td>0.0041</td>
<td>0.0041</td>
<td>-398.65</td>
<td>801.31</td>
<td>55.72</td>
<td>0</td>
<td>0.00</td>
</tr>
<tr>
<td>1982</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>D0</td>
<td>33.663</td>
<td>0.0515</td>
<td>0.0515</td>
<td>-398.24</td>
<td>800.48</td>
<td>54.89</td>
<td>0</td>
<td>0.02</td>
</tr>
<tr>
<td>1982</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>D1</td>
<td>255.64</td>
<td>0.0047</td>
<td>0.0008</td>
<td>-369.79</td>
<td>745.58</td>
<td>0</td>
<td>0.87</td>
<td>0.58</td>
</tr>
<tr>
<td>1982</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>D1</td>
<td>55.77</td>
<td>1.0000</td>
<td>0.0006</td>
<td>-371.70</td>
<td>749.39</td>
<td>3.81</td>
<td>0.13</td>
<td>0.91</td>
</tr>
<tr>
<td>1985</td>
<td>20742</td>
<td>18203</td>
<td>2539</td>
<td>197</td>
<td>44</td>
<td>D0</td>
<td>32.83</td>
<td>0.0569</td>
<td>0.0569</td>
<td>-400.20</td>
<td>804.41</td>
<td>59.00</td>
<td>0</td>
<td>0.00</td>
</tr>
<tr>
<td>1985</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>D0</td>
<td>115.51</td>
<td>0.0040</td>
<td>0.0040</td>
<td>-400.70</td>
<td>805.40</td>
<td>59.98</td>
<td>0</td>
<td>0.00</td>
</tr>
<tr>
<td>1985</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>D1</td>
<td>56.00</td>
<td>1.0000</td>
<td>0.0006</td>
<td>-372.36</td>
<td>750.72</td>
<td>5.31</td>
<td>0.07</td>
<td>0.90</td>
</tr>
<tr>
<td>1985</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>D1</td>
<td>285.24</td>
<td>0.0045</td>
<td>0.0008</td>
<td>-369.71</td>
<td>745.41</td>
<td>0</td>
<td>0.93</td>
<td>0.28</td>
</tr>
<tr>
<td>1990</td>
<td>21253</td>
<td>18641</td>
<td>2612</td>
<td>189</td>
<td>42</td>
<td>D0</td>
<td>31.10</td>
<td>0.0541</td>
<td>0.0541</td>
<td>-392.60</td>
<td>789.21</td>
<td>50.93</td>
<td>0</td>
<td>0.05</td>
</tr>
<tr>
<td>1990</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>D0</td>
<td>82.52</td>
<td>0.0049</td>
<td>0.0049</td>
<td>-395.84</td>
<td>795.67</td>
<td>57.39</td>
<td>0</td>
<td>0.06</td>
</tr>
<tr>
<td>1990</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>D1</td>
<td>53.45</td>
<td>1.0000</td>
<td>0.0006</td>
<td>-366.14</td>
<td>738.28</td>
<td>0</td>
<td>0.72</td>
<td>0.87</td>
</tr>
<tr>
<td>1990</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>D1</td>
<td>198.29</td>
<td>0.0050</td>
<td>0.0008</td>
<td>-367.06</td>
<td>740.12</td>
<td>1.84</td>
<td>0.28</td>
<td>0.59</td>
</tr>
<tr>
<td>1995</td>
<td>21460</td>
<td>18812</td>
<td>2648</td>
<td>188</td>
<td>41</td>
<td>D0</td>
<td>30.90</td>
<td>0.0521</td>
<td>0.0521</td>
<td>-403.60</td>
<td>811.19</td>
<td>52.16</td>
<td>0</td>
<td>0.00</td>
</tr>
<tr>
<td>1995</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>D0</td>
<td>81.68</td>
<td>0.0048</td>
<td>0.0048</td>
<td>-406.48</td>
<td>816.97</td>
<td>57.94</td>
<td>0</td>
<td>0.00</td>
</tr>
<tr>
<td>1995</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>D1</td>
<td>53.19</td>
<td>1.0000</td>
<td>0.0005</td>
<td>-376.51</td>
<td>759.03</td>
<td>0</td>
<td>0.64</td>
<td>0.66</td>
</tr>
<tr>
<td>1995</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>D1</td>
<td>197.21</td>
<td>0.0049</td>
<td>0.0007</td>
<td>-377.08</td>
<td>760.15</td>
<td>1.12</td>
<td>0.36</td>
<td>0.17</td>
</tr>
<tr>
<td>2000</td>
<td>21205</td>
<td>18607</td>
<td>2598</td>
<td>187</td>
<td>42</td>
<td>D0</td>
<td>30.44</td>
<td>0.0567</td>
<td>0.0567</td>
<td>-392.25</td>
<td>788.50</td>
<td>49.25</td>
<td>0</td>
<td>0.04</td>
</tr>
<tr>
<td>2000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>D0</td>
<td>81.04</td>
<td>0.0050</td>
<td>0.0050</td>
<td>-395.68</td>
<td>795.35</td>
<td>56.11</td>
<td>0</td>
<td>0.03</td>
</tr>
<tr>
<td>2000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>D1</td>
<td>52.86</td>
<td>1.0000</td>
<td>0.0006</td>
<td>-366.60</td>
<td>739.25</td>
<td>0</td>
<td>0.72</td>
<td>0.93</td>
</tr>
<tr>
<td>2000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>D1</td>
<td>189.58</td>
<td>0.0051</td>
<td>0.0008</td>
<td>-367.59</td>
<td>741.18</td>
<td>1.92</td>
<td>0.28</td>
<td>0.53</td>
</tr>
<tr>
<td>2005</td>
<td>20860</td>
<td>18321</td>
<td>2539</td>
<td>188</td>
<td>43</td>
<td>D0</td>
<td>30.25</td>
<td>0.0630</td>
<td>0.0630</td>
<td>-390.61</td>
<td>785.22</td>
<td>52.38</td>
<td>0</td>
<td>0.04</td>
</tr>
<tr>
<td>2005</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>D0</td>
<td>104.84</td>
<td>0.0040</td>
<td>0.0040</td>
<td>-392.17</td>
<td>788.34</td>
<td>55.50</td>
<td>0</td>
<td>0.01</td>
</tr>
<tr>
<td>2005</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>D1</td>
<td>53.32</td>
<td>1.0000</td>
<td>0.0006</td>
<td>-365.05</td>
<td>736.11</td>
<td>3.26</td>
<td>0.16</td>
<td>0.97</td>
</tr>
<tr>
<td>2005</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>D1</td>
<td>232.01</td>
<td>0.0046</td>
<td>0.0008</td>
<td>-363.42</td>
<td>732.85</td>
<td>0</td>
<td>0.84</td>
<td>0.48</td>
</tr>
</tbody>
</table>
Deriving the sampling formula

One-guild sampling formula
First, we sketch a derivation of the one-guild sampling formula. Consider a community of size $J$, which is governed by Hubbell’s neutral model dynamics. Individuals die at a constant rate and are replaced with probability $1 - m$ by offspring from within the community, or with probability $m$ by an immigrant from outside the community. We denote the relative abundances of the immigrants (that is, the metacommunity) by $p_1, p_2 \ldots p_s$ and the absolute abundances in the community by $N_1, N_2 \ldots N_S$. Note that

$$\sum_{i=1}^{s} p_i = 1 \quad \text{and} \quad \sum_{i=1}^{s} N_i = J$$

Then, using the fundamental dispersal number $l = \frac{m(J-1)}{1-m}$, the stationary distribution of the community abundances is

$$P(\vec{N} | \vec{p}, I, J) = \frac{J!}{(I)_J} \frac{(lp_1)N_1 \ldots (lp_S)N_S}{N_1! \ldots N_S!} \quad (S1)$$

This formula describes the abundance distribution in the local community. It can also be used to obtain the abundance distribution in the metacommunity. To do so, the migration probability $m$ must be interpreted as the speciation probability $v$, the metacommunity size is denoted by $J_M$ and the fundamental dispersal number $l$ is replaced by the fundamental biodiversity number $\theta = \frac{v(J_M-1)}{1-v}$.

Etienne (2005) used equation (S1) to derive the sampling formula of Hubbell’s neutral model, that is, the probability that a sample taken from Hubbell’s neutral community has abundance vector $D$. The sampling formula is given by:

$$P(D | \theta, I, J) = \frac{J!}{\prod_{i=1}^{s} J_i \prod_{j=1}^{J} (S_j)!} \frac{(\theta)^S}{(I)_J} \sum_{A=\lambda}^{J} \left( K(D, A) \frac{(I)_{\lambda}}{(\theta)_{\lambda}} \right) \quad (S2)$$

where $D$ is a vector of the number of individuals per species, $S$ is the number of species in vector $D$ and $J$ is the total number of individuals in vector $D$. $K(D, A)$ is defined as follows:

$$K(D, A) := \sum_{\{a_1 \ldots a_S\}_{a=1}^{S} a_{a=1}^{S}} \frac{\bar{s}(n_i a_i)}{\bar{s}(n_i 1)}$$

where $n_i$ is the number of individuals of species $i$ and $\bar{s}(n_i a_i)$ is the unsigned Stirling number of the first kind. The notation $(x)_y$ is the Pochhammer symbol defined as:

$$(x)_y = \prod_{i=1}^{y}(x + i - 1)$$

Two-guild sampling formula
Next, we derive the two-guilds sampling formula. Species belong to one of two guilds $X$ and $Y$ with different dispersal ability $\alpha_X$ and $\alpha_Y$. Total community size $J$ is a fixed
parameter, but guild sizes $J_X$ and $J_Y$ are dynamic variables. The species relative abundances in the metacommunity are $p_{X,i} (i = 1, \ldots, S_X)$ for guild $X$ and $p_{Y,i} (i = 1, \ldots, S_Y)$ for guild $Y$, so that the guild relative abundances are

$$p_X = \sum_{i=1}^{S_X} p_{X,i} \quad \text{and} \quad p_Y = \sum_{i=1}^{S_Y} p_{Y,i} \quad \text{with} \quad p_X + p_Y = 1$$

We denote the local community abundances by $N_{X,i} (i = 1, \ldots, S_X)$ for guild $X$ and by $N_{Y,i} (i = 1, \ldots, S_Y)$ for guild $Y$ so that

$$J_X = \sum_{i=1}^{S_X} N_{X,i} \quad \text{and} \quad J_Y = \sum_{i=1}^{S_Y} N_{Y,i} \quad \text{with} \quad J_X + J_Y = 1$$

As in the case of a single guild, dead individuals are replaced with probability $1 - m$ by local offspring and with probability $m$ by immigration. In contrast to the case of a single guild, the immigration probability of a specific species is not only determined by its metacommunity abundance, but also by the guild it belongs to. In particular,

Immigration by species $i$ of guild $X$ has probability $\alpha_X p_{X,i}$

Immigration by species $i$ of guild $Y$ has probability $\alpha_Y p_{Y,i}$

so that

$$m = \sum_{i=1}^{S_X} \alpha_X p_{X,i} + \sum_{i=1}^{S_Y} \alpha_Y p_{Y,i} = \alpha_X p_X + \alpha_Y p_Y$$

We use equation (S1) to compute the stationary community composition. To do so we construct a virtual metacommunity with relative abundances

Species $i$ of guild $X$ has relative abundance $\frac{\alpha_X p_{X,i}}{\alpha_X p_X + \alpha_Y p_Y}$

Species $i$ of guild $Y$ has relative abundance $\frac{\alpha_Y p_{Y,i}}{\alpha_X p_X + \alpha_Y p_Y}$

We then consider neutral immigration from this virtual metacommunity with immigration probability $m$. Explicitly,

Immigration by species $i$ of guild $X$ has probability $m \times \frac{\alpha_X p_{X,i}}{\alpha_X p_X + \alpha_Y p_Y} = \alpha_X p_{X,i}$

Immigration by species $i$ of guild $Y$ has probability $m \times \frac{\alpha_Y p_{Y,i}}{\alpha_X p_X + \alpha_Y p_Y} = \alpha_Y p_{Y,i}$

Comparing these immigration probabilities with the previous ones, we see that neutral immigration from this virtual metacommunity is equivalent with the original immigration process. Therefore, we can apply equation (S1) to the virtual metacommunity to obtain the abundance distribution of the original community,
\[ P(\tilde{N}_X, \tilde{N}_Y | \tilde{p}_X, \tilde{p}_Y, \alpha_X, \alpha_Y, I, J) \]
\[ = \frac{j!}{(I)_j} \left( \frac{\alpha_X p_X}{\alpha_X p_X + \alpha_Y p_Y} \right)_{N_{X,1}} \ldots \left( \frac{\alpha_X p_X}{\alpha_X p_X + \alpha_Y p_Y} \right)_{N_{X,S}} \times \frac{(I_Y p_Y)_{N_{Y,1}}}{(I_Y)_{J_Y}} \ldots \frac{(I_Y p_Y)_{N_{Y,S}}}{(I_Y)_{J_Y}} \]
\[ = \frac{j!}{(I)_j} \frac{(I_X)_{I_X}(I_Y)_{J_Y}}{I_X! J_Y!} \]

Introducing the guild fundamental dispersal numbers

\[ I_X = \frac{\alpha_X p_X}{\alpha_X p_X + \alpha_Y p_Y} \quad l = \frac{\alpha_X p_X(j - 1)}{1 - \alpha_X p_X + \alpha_Y p_Y} \]  
\[ I_Y = \frac{\alpha_Y p_Y}{\alpha_X p_X + \alpha_Y p_Y} \quad l = \frac{\alpha_Y p_Y(j - 1)}{1 - \alpha_X p_X + \alpha_Y p_Y} \]

we get

\[ P(\tilde{N}_X, \tilde{N}_Y | \tilde{p}_X, \tilde{p}_Y, \alpha_X, \alpha_Y, I, J) \]
\[ = \frac{j!}{(I)_j} \frac{(I_X p_X)_{N_{X,1}}}{N_{X,1}! \ldots N_{X,S}!} \ldots \frac{(I_X p_X)_{N_{X,S}}}{N_{X,1}! \ldots N_{X,S}!} \times \frac{(I_Y p_Y)_{N_{Y,1}}}{(I_Y)_{J_Y}} \ldots \frac{(I_Y p_Y)_{N_{Y,S}}}{(I_Y)_{J_Y}} \]
\[ = \frac{j!}{(I)_j} \frac{(I_X)_{I_X}(I_Y)_{J_Y}}{I_X! J_Y!} \]

In the last equality, the last line gives the abundance distribution of the species belonging to guild X, which is an instance of the one-guild formula (S1). Hence, the remaining factors on the first line give the probability distribution of the guild sizes \( J_X \) and \( J_Y \),

\[ P(J_X, J_Y | I_X, I_Y, J) = \frac{j!}{(I_X + I_Y)^j} \frac{(I_X)_{I_X}(I_Y)_{J_Y}}{I_X! J_Y!} \]  

As a result,

\[ P(\tilde{N}_X, \tilde{N}_Y | \tilde{p}_X, \tilde{p}_Y, \alpha_X, \alpha_Y, I, J) = P(J_X, J_Y | I_X, I_Y, J) P(\tilde{N}_X | \tilde{p}_X, I_X, J_X) P(\tilde{N}_Y | \tilde{p}_Y, I_Y, J_Y) \]  

(S7)

The product structure shows that, given guild sizes \( J_X \) and \( J_Y \) and parameters \( I_X \) and \( I_Y \), the abundance distributions of guilds \( X \) and \( Y \) are independent.

Equation (S7) describes the abundance distribution in the local community. In particular, the product structure of equation (S7) carries over to the sampling formula for dispersal-limited sample from the metacommunity for given values of the metacommunity relative abundances of the two guilds,
\[
P(D_X, D_Y | \theta, l_X, l_Y, J) = P(J_X, J_Y | l_X, l_Y, J) P(D_X | \theta, J_X) P(D_Y | \theta, l_Y, J_Y)
\]  
(S8)

which is equation (2) in the main text.

To obtain the full sampling formula, we need to integrate over all possible values of the relative abundances of the two guilds appropriately weighted by the probability density of these relative abundances. Equation (S7) can also be used to obtain this probability density of the guilds' relative abundances in the metacommunity. Given the relative abundances \(p_X\) and \(p_Y\) of guilds \(X\) and \(Y\), the abundance distributions of guilds \(X\) and \(Y\) are independent. To compute the distribution of \(p_X\) and \(p_Y\), we first lift the expression \(P(J_X, J_Y | l_X, l_Y, J)\) from the local community to the metacommunity. Using \(J_X \rightarrow J_{M,X}, J_Y \rightarrow J_{M,Y}, J \rightarrow J_M, l_X \rightarrow \theta_X = \frac{\theta}{2}\) and \(l_Y \rightarrow \theta_Y = \frac{\theta}{2}\), we get

\[
P(J_{M,X}, J_{M,Y} | \theta, J_M) = \frac{J_M^1}{(\theta)_{J_M}} \frac{\left(\frac{\theta}{2}\right)_{J_{M,X}} \left(\frac{\theta}{2}\right)_{J_{M,Y}}}{J_{M,X}^1 J_{M,Y}^1}
\]

or

\[
P(J_{M,X} | \theta, J_M) = \frac{J_M^1}{(\theta)_{J_M}} \frac{\left(\frac{\theta}{2}\right)_{J_{M,X}} \left(\frac{\theta}{2}\right)_{J_{M,Y}}}{J_{M,X}^1 (J_M - J_{M,X})!}
\]  
(S9)

Then we take the coupled limit \(J_{M,X} \rightarrow \infty\) and \(J_M \rightarrow \infty\) with \(J_{M,X} = p_X J_M\). That is, we transform the discrete probability distribution \(P(J_{M,X})\) of absolute abundances to a continuous distribution \(\rho(p_X)\) of relative abundances. We have

\[
P(J_{M,X} = \alpha J_M) = P\left(\frac{\alpha J_M}{J_M} \leq p_X < \frac{\alpha J_M + 1}{J_M}\right) = \int_\alpha^{\alpha+1} \rho(p_X) dp_X \approx \rho(\alpha) \frac{1}{J_M}
\]

such that

\[
\rho(p_X | \theta, J_M) = \lim_{J_M \rightarrow \infty} J_M P(p_X | J_M \theta, J_M) = \lim_{J_M \rightarrow \infty} \frac{J_M^1}{(\theta)_{J_M}} \frac{\left(\frac{\theta}{2}\right)_{p_X J_M} \left(\frac{\theta}{2}\right)_{(1-p_X) J_M}}{(p_X J_M)! ((1-J_M - p_X) J_M)!}
\]

\[
= \lim_{J_M \rightarrow \infty} \frac{\Gamma(\theta) \left(p_X J_M\right)^{\theta-1} \left(1-p_X J_M\right)^{\theta-1}}{J_M^{\theta-1} \Gamma(\frac{\theta}{2}) \Gamma(\frac{\theta}{2})}
\]

\[
= \frac{\Gamma(\theta)}{\Gamma(\frac{\theta}{2})} p_X^{\theta-1} (1-p_X)^{\theta-1} \lim_{J_M \rightarrow \infty} J_M J_M^{-\theta+1} J_M^{\theta-1} J_M^{\theta-1}
\]

\[
= \frac{\Gamma(\theta)}{\Gamma(\frac{\theta}{2})} p_X^{\theta-1} (1-p_X)^{\theta-1}
\]  
(S10)

which is equation (4) in the main text.
Combining equations (S8) and (S10), and integrating over all possible values of \( p_X \) and gives us the full sampling formula (equation (5) in the main text):

$$
P(D_X, D_Y | \theta, \alpha_X, \alpha_Y) = \int_0^1 P(J_X, J_Y | I_X, I_Y) P(D_X | \theta, I_X) P(D_Y | \theta, I_Y) \rho(p_X | \theta) dp_X \tag{S11}
$$

### Conditioning on guild size

Guild sizes \( J_X \) and \( J_Y \) are central to equation (S11) and we therefore expect that differences in guild size might disproportionately affect parameter estimates. To see to which extent parameter estimates are influenced by differences in guild size, rather than differences in dispersal ability between guilds, we plotted the ratio of guild sizes versus the ratio of dispersal ability estimates obtained using equation (S8) (we used obtained parameter estimates obtained using the procedure described under the “model selection” part of the methods section). Only the parameter estimate of the model with the highest AIC weight was taken into consideration. We found a positive relation between the ratio of dispersal abilities of both guilds and the ratio between guild sizes (\( R^2 = 0.95 \), slope = 0.927, \( p < 2 \times 10^{-16} \), figure A1). It seems thus, that equation (S8) overly emphasizes the impact of differences in guild size and negates any differences in the abundance distributions of the two guilds. A next step would be to condition our sampling formula on guild size.

To condition on guild size, we can use the likelihood of having guilds of size \( J_X \) and \( J_Y \) (equation (S6)):

$$
P(J_X, J_Y | I_X, I_Y, J) = \frac{J!}{(I_X + I_Y)J} \frac{(I_X)_{J_X} (I_Y)_{J_Y}}{J_X! J_Y!}
$$

Taking into account all possible combinations of \( I_X \) and \( I_Y \), and remembering from S4A and S4B that \( I_X \) and \( I_Y \) depend on \( p_X \) we obtain:

$$
P(J_X, J_Y | \theta, I_X, I_Y) = \int_0^1 \frac{J!}{(I_X + I_Y)J} \frac{(I_X)_{J_X} (I_Y)_{J_Y}}{J_X! J_Y!} \rho(p_X | \theta) dp_X \tag{S12}
$$

Using equation (S12), we condition equation (S11) and obtain:

$$
P(D_X, D_Y | \theta, I_X, I_Y, J_X, J_Y) = \int_0^1 \frac{J!}{(I_X + I_Y)J} \frac{(I_X)_{J_X} (I_Y)_{J_Y}}{J_X! J_Y!} \rho(p_X | \theta) dp_X \tag{S13}
$$

Using equation (S13), we repeated the procedure to obtain the ratio between guild sizes and the ratio between dispersal abilities. This time, no significant correlation between the ratio of dispersal abilities and the ratio of guild sizes was detected anymore (\( R^2 = 0.00276 \), \( p = 0.115 \), figure A1).
Figure A1. Ratio of dispersal abilities of both guilds, versus the ratio of individuals in these guilds. Left hand plot shows ratios obtained using the unconditioned sampling formula, right hand plot shows ratios obtained using the conditioned sampling formula. For the unconditioned sampling formula, a clear relation is found between the ratio of dispersal abilities and ratio of guild sizes \( R^2 = 0.95, \text{ slope } = 0.927, p < 2e-16 \). Using the conditioned sampling formula, this relationship vanishes however \( R^2 = 0.00276, p = 0.115 \).