Supplementary Material, Online Resource 1

Community Development Model

Origin of the dust bunny distribution in ecological community data
B. McCune B and H. T. Root. 2014. *Plant Ecology*
Corresponding author: B. McCune, Oregon State University, mccuneb@onid.orst.edu

We model multiple population dynamics simultaneously and as simply as possible to produce dust bunny distributions. Other existing community simulators are static and statistical rather than dynamic (e.g. COMPAS by Minchin 1987). These simulate communities according to the statistical properties of species response to environmental gradients, rather than deriving communities from dynamic population processes. In contrast, we develop communities in unoccupied space, simulating response to a community-replacing disturbance. Populations are described by A_j, the density or number of individuals of a given species *j*. The model is discrete with respect to time. One time step can be considered either a single generation (in which case we assume that all species in the community have the same generation time), or a specific time interval (e.g. 1 year).

Environmental gradient.—Species have Gaussian responses on one or more environmental gradients. For simplicity, species response curves are identical in height and spread and are evenly spaced along the gradients. In other words, for a given simulation, niche widths and maxima are set equally among species. In reality, species performances along environmental gradients take various shapes and spreads, but this added complexity was unnecessary to generate realistic data sets.

The environmental effect on species *j* at a particular point *x* on an environmental gradient *k*, the species having an optimum at $xopt_{kj}$ and spread or niche width, s_k, is modeled as a Gaussian function:

$$Z_{jk} = \exp\left[-0.5\left(\left(x_k - xopt_{jk}\right)/s_k\right)^2\right] \quad (1)$$

The range of the environmental effect is 0 (complete failure) to 1 (maximal performance).

Zero to three environmental gradients are generated independently: species optima are assigned at random along each gradient. The sampled range of each gradient is 0 to 100, but the optima are assigned from -25 to +125, such that some species optima fall outside of the sampled range.

With more than one environmental gradient, the performances on individual gradients are multiplied to achieve a combined effect **Z*** of the *q* environmental gradients on species *j*:

$$z_j^* = \prod_{k=1}^{q} z_{jk} \quad (2)$$

Optionally, individual environmental gradients can be assigned weights, *w*, ranging from 0 (no effect) to 1 (full effect), in calculating the combined effect, **Z***, of multiple environment gradients on species performance:

$$z_j^* = \prod_{k=1}^{q} 1 - w_k\left(1 - z_{jk}\right) \quad (3)$$

Although this could be useful in making the model more realistic and general, it proved unnecessary to generate dust bunnies. Similarly, each environmental gradient could be weighted individually by species. In the simulation results reported here, however, we gave the environmental gradients equal and full weights.

Competitive ability.—Competitive abilities (*C*) are assigned to species by generating random numbers with a uniform distribution ranging from zero (no competitive ability) to one (maximal competitive ability). Recognizing the inherent tradeoffs between competitive ability and reproductive effort, both immigration rates and the intrinsic rate of population growth are set to vary negatively with competitive ability (Fig. 1-1). We call this "competitive ability" because it incorporates both a competitive effect on other species (via its relationship to immigration and growth rates) and a competitive response to other species (via its effective carrying capacity; see below).

Immigration.—The number of immigrants in the population incorporated a random draw that can be thought of as the number of propagules arriving; this was then filtered by the favorability of the habitat, $\mathbf{Z}^*$, at the site such that immigrants were more likely to survive in suitable habitats.

Immigration (I) is treated as a species-specific stochastic Poisson process. We chose a Poisson model because it counts the number of discrete and independent random occurrences per unit time. In this case the discrete event is the arrival of an individual and, for sessile organisms, its subsequent colonization. The Poisson parameter, μ, for a given species can be thought of as the immigration pressure from that species. For simplicity, we assumed that the immigration pressure, μ, varies linearly and negatively with competitive ability: $\mu = -C + 1$ (Fig. 1-1). Furthermore, we assume that immigration pressure is regional, rather than from other sample units in a given data set.

Immigration in a given time step is determined by a draw from a uniform random number generator along with the probabilities of I immigrants:

$$p_j(I) = \frac{\mu^I e^{-\mu}}{I!} \quad (4)$$

For example, if $\mu = 0.05$, then $p_j(0) = 0.905$, $p_j(1) = 0.090$, $p_j(2) = 0.005$. and $\sum_y^\infty p_j(I = y) = 1$. We calculated $p_j(I = y)$ for $I \leq 15$, because for our choices of μ, p_j is negligibly small for $I > 15$.

Immigrants cannot colonize successfully outside of their environmental tolerance, so before taking a random draw to determine I, these probabilities, p_j, are first adjusted to p_j' by the environmental effect z_j^* as derived above, for example: $p_j'(1) = p_j(1) \cdot z^*_j$.

Cumulatively the environmental effect deflates the sum of the Poisson distribution for nonzero values by the proportion z^*_j, ranging from $z^*_j = 1$ at a species optimum to $z^*_j = 0$ for complete failure. We correspondingly increase the frequency of zero values as follows. Since $p_j(I \neq 0) = 1 - p_j(0)$ and $p_j'(0) = 1 - z^*_j \cdot p_j(I \neq 0)$, then $p'_j(0) = 1 - z^*_j(1 - p_j(0))$. The idea is that we know the probability of a positive value of I by subtracting $p_j(0)$ from 1. Then we deflate $\sum_y^\infty p_j(I = y)$ by z^*_j. The amount of deflation is added back in to $p'(0)$, so that $\sum_y^\infty p'_j(I = y) = 1$.

For example, assume $p_j(1) = 0.08$, and species j is tolerant to the environment at a given location, with $z^*_j = 0.5$ indicating half of its optimum performance, then $p'_j(1) = 0.08(0.5) = 0.04$. Conversely, if a species has this same immigration pressure, but the habitat is completely unfavorable with $z^*_j = 0$, then $p'_j(1) = 0.08(0.00) = 0$, and the species will not be observed there, regardless of the immigration pressure.

After generating p'_j ($I = 0$), p'_j ($I = 1$), p'_j ($I = 2$), etc., I is assigned accordingly by a random number draw. In computation, as soon as a random number from a uniform pseudorandom number generator in the interval (0,1) exceeds the cumulative sum of p'_j as I is incremented, I is set to that integer.

Intrinsic Growth Rate.—Assume that the relationship between intrinsic growth and competitive ability has a fixed, negative tradeoff (Fig. 1-1):

$$R = -1C + 2 \quad (5)$$

This is a simple expression of the basic relationship proposed by Pianka (1970), Grime (1977), and others. Then by algebra, species with higher intrinsic growth rates, R, tend to have higher immigration pressure, μ (Fig. 1-1):

$$R = \mu + 1 \quad (6)$$

Tuning Immigration Pressure against Competitive Ability.—To control the system-wide balance between growth rates and dispersal limitations, we introduce a dispersal limitation factor, d. This single factor is used to change the relative slopes of μ vs. C, R vs. μ, and R vs. C, with the last slope arbitrarily held constant at -1 (Fig. 1-1).

The parameter d is held constant for a given simulation, but can be varied to increase or decrease the dispersal limitation built into the whole community. Increasing d increases the dispersal limitation by selecting a lower immigration pressure for a given competitive ability and intrinsic growth.

For μ vs. C,

$$\mu = (1 - C)/d \quad (7)$$

Similarly, for R vs. μ,

$$R = d * \mu + 1 \quad (8)$$

Stronger dispersal limitation at the community level means that immigration pressure is low for a given combination of intrinsic growth rates, *R*, and competitive ability, *C*.

Growth.—Growth, G_j, of a particular species *j* is a function of intrinsic growth rate *R* diminished by:

1. an effect of suboptimal environment, z^*_j
2. a logistic competitive effect as the sample unit approaches its carrying capacity, *K*, set here to a constant 1000.

$$G_j = R_j \cdot z_j \cdot A_j \left(1 - \frac{\sum_{j=1}^{p} A_j}{c_j K}\right) \quad (9)$$

The term c_jK makes an "effective carrying capacity," that is, a species-specific carrying capacity that is proportionate to competitive ability. This last term departs from the widely used Lotka-Volterra standard, which is

$$\left(1 - \frac{\sum \alpha_{ji} A_j}{K}\right) \quad (10)$$

with α_{ji} being the effect of each species *i* on species *j*. Here we make the simplification that each species is affected equally by equal amounts of other species (i.e. $\alpha_{j1} = \alpha_{j2} = \ldots \alpha_{ijp}$). Furthermore, rather than express negative competitive effects, α, of other species, we use its complement, c_j, the competitive ability of species *j*. This makes it easier to express a basic trait, competitive ability, in the other parts of the model. The model implies that the general competitive ability of a given species matters more than the impact of other particular species on that species.

In contrast, the traditional Lotka-Volterra formulation allows that a species might have a strong impact on one species and a weak impact on another. For example, if Species A is limited by water, and Species B by light, they may not compete strongly because A can grow well under B. Introduce Species C, which strongly competes for light but does not use much water. In the Lotka-Volterra formulation, Species C can be strongly competitive against B and have little effect on A. In our formulation of the model, however, competitive ability doesn't apply differently to different resources, except as expressed in optima along particular environmental gradients. Thus in our model, Species C would be assumed to be a strong competitor against both A and B.

Population size.—Population size for a given sample unit is incremented with a difference equation based on population size of species *j* in the previous time step ($A_{j,t}$), immigration, and growth:

$$\frac{\Delta A_j}{\Delta t} = I_j + G_j \quad (11)$$

$$A_{j,t+1} = A_{j,t} + \frac{\Delta A_j}{\Delta t} \quad (12)$$

Community matrices.— Each sample unit × species matrix was assembled by running the model once for each sample unit, choosing the following parameters, as specified in Table 2: degree of dispersal limitation, niche width, number of environmental gradients, and number of time steps (or generations). Sample units in a given matrix vary in position on one or more environmental gradients. Species in a given matrix vary in position of optima on those gradients, degree of dispersal limitation, competitive abilities, and intrinsic growth rates.

Fig 1-1. Simple functions relating competitive ability (*C*), the intrinsic population growth rate (*R*), and the Poisson parameter (μ) representing immigration pressure. The dispersal limitation factor (*d*) controls the balance between population growth rates and dispersal limitation.

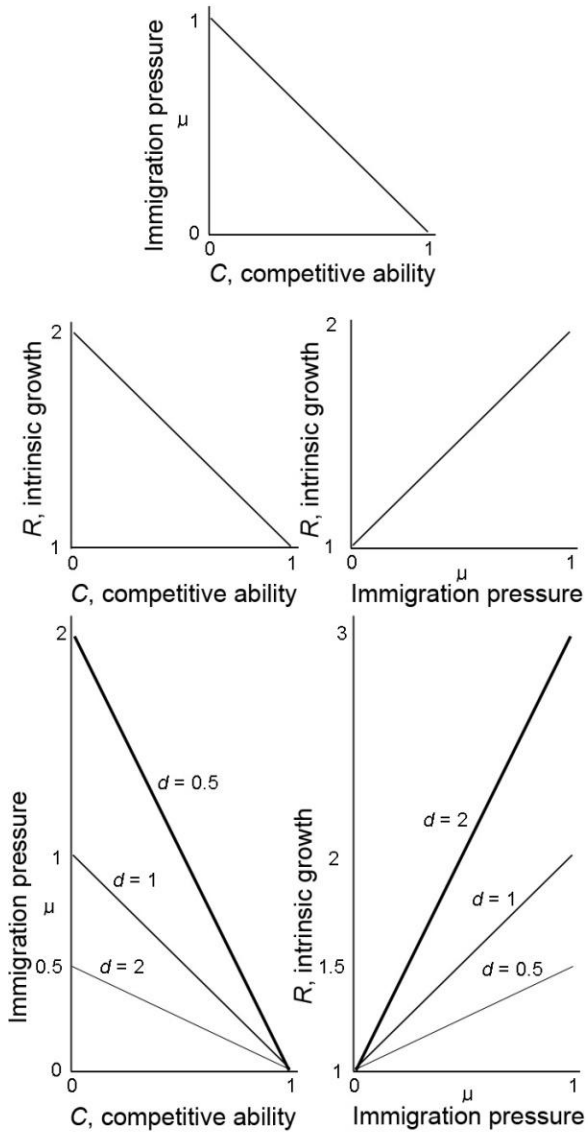

Table 1-1 Community simulation model parameters, inputs, and outputs. Manipulated inputs are constant for generating a given data set ("manipulated by data set") or for generating a given community sample unit ("manipulated, by sample unit"). Each data set consisted of 200 sample units × 30 species.

Model inputs and outputs, units	Symbol	Source or type	Min	Max
Environment and disturbance descriptors				
Position on environmental gradient *k*, arbitrary environmental gradient units	x_k	random	0	100
Number of environmental gradients, count	q	manipulated, by data set	0	3
Maximum number of time steps (or number of generations) since disturbance, count	t_{max}	manipulated, by data set	1	10
Individual species descriptors				
Species optimum on environmental gradient, arbitrary environmental gradient units	$xopt_{jk}$	random	-25	125
Niche width, standard deviations of environmental gradient units	s	manipulated, by data set	15	50
Environmental effect, unitless proportion of maximum performance	z^*_j	function of x, s, $xopt_{jk}$	0	1
Immigration pressure (Poisson parameter)	μ	random	0	2
Probability of y immigrants for species j	$p_j\ (I = y)$	Poisson distribution		
Competitive ability, density/density	C	random	0	1
Dispersal limitation, unitless slope of μ vs. C	d	manipulated	1	20
Intrinsic growth rate, density/density	R	fixed linear function of C and μ	1	2
Carrying capacity, community-level but applied to individual species, density	K	constant	1000	1000
Population growth, density	G	modified logistic function of R, z, A, C, and K	0	unbounded
Inputs varied by time step				
Immigration, count	I	Poisson distribution applied to random draw	0	unbounded
Probability of y immigrants for species j	$p_j\ (I = y)$		0	1
Outputs				
Population size, number of individuals (or density) of species j	A_j	function of I, G, and A in previous time step	0	unbounded

Additional inputs available in software, but not used here				
Range across sample units in time steps since disturbance, count	t_{range}	constant for these simulations	0	0
Maximum number of time steps since disturbance, for sample unit *i*, count	$t_{max,i}$	random within t_{range}	1	10
Environmental gradient weighting option, categorical	E_{opt}	0=equal, 1=descending linear series, 2=descending series, broken stick	0	2
Weight of environmental gradient *k*, proportion of maximum	w_k	function of E_{opt}	0	1

Sensitivity Analysis

Sensitivity analysis measured the importance to the dust bunny intensity (DBI) of the four factors that we varied as inputs to the population models: number of environmental gradients, niche width, number of generations since community-replacing disturbance, niche width, and dispersal limitation. Sensitivities were analyzed by first fitting a multidimensional response surface for the DBI = *f*(input variables), where *f* is an unspecified smooth function derived with a kernel smoother, and each data point is one of the 19,200 simulated data sets. We used a multiplicative Gaussian kernel with a local linear model for nonparametric multiplicative regression (NPMR, McCune 2006), as implemented in HyperNiche (McCune and Mefford 2009). The multiplicative kernel automatically accommodates interactions among input variables. Smoothing parameters ("tolerances" in HyperNiche) were arbitrarily set to 5% of the range of the input variables; this yielded cross-validated R^2 of 0.982 to 0.985 for the 5-dimensional response surface of DBI to the four factors.

Given an NPMR representation of the response surface, sensitivity analysis proceeded by nudging the input values up and down by 5% of the range for individual variables, then measuring the resulting change in the dust bunny indices, data point by data point. The change in the response was measured as a proportion of the observed range of the response variable. Scaling the differences in response and differences in predictors to their respective ranges allows a sensitivity measure, *Q*, that is an easily interpreted ratio, independent of the units of the variables.

The general concept is:

$$Q = \frac{\text{mean difference in response / range in response}}{\text{difference in predictor / range in predictor}} \qquad (13)$$

If $\hat{y}_{i+}$ is the estimate of the response variable for case *i*, having *increased* the input variable *j* by an arbitrarily small value Δ (say 0.05 of the range of the input variable), and $\hat{y}_{i-}$ is the estimate of the response variable for case *i*, having *decreased* the input variable by the same small value Δ, then the numerator for Eqn. 13, the scaled difference in response, can be written as:

$$\text{scaled difference in response} = \frac{\sum_{i=1}^{n} \left| \hat{y}_{i+} - \hat{y}_{i-} \right|}{2n \left| y_{\max} - y_{\min} \right|} \qquad (14)$$

Dividing by two expresses the sensitivity as a response to a single nudging, since two nudgings are used to calculate the numerator. The denominator, the scaled difference in the predictor, is Δ. This is the amount by which we choose to nudge the predictor expressed as a proportion of the predictor's range. Combining these yields the sensitivity measure *Q*:

$$Q_j = \frac{\left(\sum_{i=1}^{n} \left| \hat{y}_{i+} - \hat{y}_{i-} \right| \right)}{2n \left| y_{\max} - y_{\min} \right| \Delta} \qquad (16)$$

By accumulating Q_j across all of the data points for each input variable j, we evaluated the sensitivity of the modeled response to each input variable. The greater the sensitivity, the more influence that variable has in the model. If $Q = 1.0$, then, on average, nudging a predictor results in a change in response of equal magnitude to the change in predictor, both changes measured relative to their ranges. If $Q = 0.0$, then nudging a predictor has no detectable effect on the response.

References

McCune B (2006) Non-parametric habitat models with automatic interactions. J Veg Sci 17:819-830.

McCune B, Mefford MJ (2009) HyperNiche. Nonparametric Multiplicative Habitat Modeling. Version 2.21. MjM Software, Gleneden Beach, Oregon, USA.

Minchin PR (1987) Simulation of multidimensional community patterns: towards a comprehensive model. Vegetatio 71:145-156.

Supplementary Material, Online Resource 2

Content, Source and Characteristics of Real Data Sets For Comparison

Origin of the dust bunny distribution in ecological community data
B. McCune B and H. T. Root. 2014. *Plant Ecology*
Corresponding author: B. McCune, Oregon State University, mccuneb@onid.orst.edu

Table 2-1 Content, source, beta diversity, and dust bunny indices for real data sets.

		Raw data		log(x+1) transformed		
Matrix	% zero	β_d	DBI	β_d	DBI	Source
AlaskaEpiphytes: 50 plots × 29 lichen spp, cover classes on arscine-squareroot scale, averaged across many subplots	61.2	1.0	0.873	0.8	0.728	Derr et al. (2007)
HydroPlants: 50 plots × 88 plant species, % cover	82.1	2.1	0.929	1.7	0.881	Otting (1998)
IntertidalWhelks: 53 plots × 17 invertebrate species, rocky intertidal, % cover averaged across dates	45.7	0.6	0.855	0.4	0.655	Navarrete and Menge (1996)
MammothTrees: 78 plots × 22 tree species, basal areas per hectare	87.5	3.0	0.956	2.3	0.911	McCune and Henckel (1993)
OakWoods: 47 stands × 189 plant species, various abundance measures all relativized by species maximum	81.8	2.0	0.931	1.4	0.889	Thilenius (1968)
OceanMicrobes: 369 samples × 830 T-RFLP peak heights	84.6	1.5	0.959	1.3	0.907	Treusch et al. (2009)
PondBirds: 130 ponds × 87 bird species, counts	83.4	1.5	0.954	1.2	0.933	Harris (2001)
SoilFungi: 96 samples × 283 OTUs, frequency of DNA sequence reads	80.8	2.0	0.961	1.0	0.881	Hesse and Spatafora (2013)
StreamInverts: 108 sites × 78 taxa, counts	67.3	1.3	0.939	0.7	0.797	Miller et al. (2007)
XmasBirds: 41 Area-Year combinations × 145 species, counts	62.8	1.9	0.902	0.8	0.780	Hoyer (1994, 1995).

References

Derr CC, McCune B, Geiser LH (2007) Epiphytic macrolichen communities in *Pinus contorta* peatlands in southeastern Alaska. Bryol 110:521-532.

Harris PD (2001) Bird Community Patterns of Spring-seasonal and Semi-permanent Wetlands in the Sacramento Valley, California. MS Thesis, Oregon State University.

Hesse CN, Spatafora JW (2012) Soil fungal communities associated with the mat-forming ectomycorrhizal genera *Piloderma* and *Ramaria* determined by ITS amplicon pyrosequencing. ms.

Hoyer R (1994-5) The Chat 23(5), 24(5).

McCune B, Henckel CL (1993) Tree mortality rates in two contrasting forests in Mammoth Cave National Park. Nat Areas J 13:115-123.

Miller SW, Wooster D, Li J (2007) Resistance and resilience of macroinvertebrates to irrigation water withdrawals. Freshw Biol 52:2492-2510.

Navarrete SA, Menge BA (1996) Keystone predation and interaction strength: interactive effects of predators on their main prey. Ecol Monogr 66:409-429

Otting NJ (1998) Ecological characteristics of montane floodplain plant communities in the Upper Grande Ronde basin, Oregon. MS Thesis, Oregon State University.

Thilenius JF (1968) The *Quercus garryana* forests of the Willamette Valley, Oregon. Ecol 49:1124-1133.

Treusch AH, Vergin KL, Finlay LA, Donatz MG, Burton RM, Carlson CA, Giovannoni SJ (2009) Seasonality and vertical structure of microbial communities in an ocean gyre. Int Soc for Microb Ecol J 3:1148–1163.

Supplementary Material, Online Resource 3

Relationships among measures of beta diversity and dust bunny intensity

Origin of the dust bunny distribution in ecological community data
B. McCune B and H. T. Root. 2014. *Plant Ecology*
Corresponding author: B. McCune, Oregon State University, mccuneb@onid.orst.edu

The intensity of a dust bunny is, in practice, strongly related to the beta diversity of a data set (Fig. 3-1). They are not, however, identical. Beta diversity measures the amount of heterogeneity in a community sample. The dust bunny intensity measures a particular kind of heterogeneity, where points tend to lie along the edges of a high dimensional species space. It is measured by the DBI, which is based on the community matrix mean after relativization by species maxima, and by the proportion of zeros in the community matrix.

The relationships among these measures reveal some of their basic properties:

1. Whittaker's beta diversity (BetaDivW = β_W = gamma/alpha – 1) is a simple hyperbolic function of the proportion of zeros in the community matrix (PctZeros; see text).
2. Beta diversity measured in half changes, β_D, is an exponential function of the average Sørensen (Bray-Curtis) distance among sample units, by definition.
3. DBI based on the log transformed data matrix (DBIlog) is more closely related to the proportion of zeros in the matrix (PctZeros) than is the DBI based on the untransformed matrix. Either DBI log or PctZeros is therefore usable as a measure of the intensity of the dust bunny distribution.
4. Average Sørensen (Bray-Curtis) distance of the log-transformed matrix (AveDistL) has a nearly linear relationship to the proportion of zeros in the matrix (PctZeros; Fig. 3-1).

Fig 3-1 Scatterplots of beta diversity and dust bunny measures in simulated data sets. Variable abbreviations: DBI = dust bunny intensity on untransformed matrix; PctZeros = percentage of zeros in the matrix; AveSorDi = average Sorensen distance among sample units, untransformed; BetaDivH = β_D = beta diversity in half changes, untransformed matrix (eqn. 18 in text); BetaDivW β_W = Whittaker's beta diversity; DBIlog = dust bunny intensity on log transformed matrix; AveDistL = average Sørensen distance among sample units, after log transformation, BetaHClo = β_D = beta diversity in half changes on log transformed matrix.