Origin of the dust bunny distribution in ecological community data

Bruce McCune   Heather T. Root

Manuscript as accepted by the journal Plant Ecology and later published as:

McCune, B. & H. T. Root. 2015. Origin of the dust bunny distribution in ecological community data. *Plant Ecology* 216: 645-656. doi:10.1007/s11258-014.0404-1

B. McCune, Dept. of Botany and Plant Pathology, Oregon State University, Corvallis, OR 97331, USA.

email: mccuneb@onid.orst.edu

H. T. Root, Dept. of Forest Ecosystems and Society, Oregon State University, Corvallis, OR 97331, USA.

Corresponding author:
Bruce McCune
phone 541 737 1741

Abstract The distribution of sample units in multivariate species space typically departs strongly from the multivariate normal distribution. Instead of forming a hyperellipse in species space, the sample points tend to lie along high-dimensional edges of the space. This *dust bunny distribution* is seen in most ecological community data sets. The practical consequences of the distribution to the analysis of community data are well known and severe, but no one has demonstrated how population processes generate these problems. We evaluate potential causes of dust bunny distributions by simulating a large number of non-equilibrial communities under varying conditions, verifying that they resemble real data, then analyzing the relationship between the intensity of the dust bunny distribution in these data sets and the population and environmental parameters that gave rise to them. All community data sets, both simulated and real, departed strongly from multivariate normal and lognormal distributions. Four parameters influenced intensity of dust bunnies: time since community-replacing disturbance, number of environmental factors, dispersal limitation, and niche width. Samples measured soon after community-replacing disturbance had strong dust bunny distributions. Near-equilibrial communities sampled from a narrow range in environments lead to only weak dust bunnies. Community samples taken across multiple simultaneous strong environmental gradients are likely to show strong dust bunnies, regardless of the successional state, niche width of the component species, and degree of dispersal limitation. Dust bunny intensity depends not only on population processes and disturbance, but also on the properties of the sample, such as sample unit area or volume.

Keywords Community analysis, disturbance, environment, niche width, simulation model

Introduction

Ecological community data commonly consist of species abundance or presence recorded in a set of sample units. With such data, sample units can be conceptualized as points in a coordinate system defined by species abundance (species space; Goodall 1963). The distribution of sample units in multivariate species space typically departs strongly from a multivariate normal distribution. Instead of forming a hyperellipse in species space, the sample points instead tend to lie along high-dimensional edges of the space. This *dust bunny distribution* (McCune and Grace 2002) is so called because it resembles a "dust bunny," the collection of lint and dust that tends to accumulate along the edges of our living spaces (Fig. 1). The analogy is imperfect, for reasons discussed below, but our teaching has shown this to be a useful and memorable way of conveying a fundamental difference between ecological community data and the multivariate normal distribution assumed by traditional multivariate statistics. The dust bunny distribution can be considered a particular kind of zero-inflated multivariate distribution that is characteristic of ecological community data. Naming and understanding this common empirical phenomenon in community ecology can improve expectations for analyzing community datasets and inform the development of appropriate statistical models. The dust bunny is observed even for simple data sets involving a few species along a single environmental gradient. Although the dust bunny distribution is one of the most consistent statistical properties of ecological community data, the biological mechanisms producing this pattern have never been explained. Our focus here is not on how to analyze such data, but rather contributing toward the theory of its genesis.

Dust bunny distributions occur when each sample unit has only a small subset of the available species pool, leading to a large number of zeros in the community matrix. The presence of only a small number of species in each sample unit may occur because of stochastic dispersal limitations, competition among species, recent community-replacing disturbance, or environmental conditions in the sample unit being unsuitable for many of the species in the regional species pool.

Our goals diverge from studies on frequency distributions within species (e.g., Bliss and Fisher 1953; Cassie 1962) in that our concerns are multivariate and apply to both continuous data (e.g. biomass, density, cover) and discrete data (e.g., counts or cover classes). We focus on the distributional properties of sample units in species space, a high-dimensional coordinate system where each dimension of the space is an axis defined by the abundance of a single species, and each sample unit therefore occupies a point in that space (Goodall 1963). Despite the common use of non-Euclidean (city-block) distance measures in community ecology, most statistical tools taught to biologists operate in Euclidean space. Regardless of what methods we choose for analysis, community ecologists still need to be able to explain to a broader audience how the properties of their data in Euclidean space differ from a multivariate normal distribution.

The practical consequences of the dust bunny distribution to data analysis are well known and severe. Analytical tools such as principal components analysis that seek linear relationships among variables perform poorly with ecological community data; this has been known for over 40 years (e.g. Beals 1973; Whittaker and Gauch 1973). Even a simple noiseless data set of three species with Gaussian responses to a single environmental gradient shows strong nonlinearity of that gradient when sample units are plotted in species space. Remedies exist, such as nonmetric multidimensional scaling (Kruskal 1964), that can recover environmental gradients that are expressed nonlinearly in species space (Clarke 1993; McCune and Grace 2002) and using proportional city-block distance measures.

Apparently no one has specifically focused on revealing the population processes that give rise to the dust bunny distribution. Although McCune and Grace (2002) suggested the dust bunny distribution as an empirical fact for most community data sets, they did not attempt to quantify the problem nor to explain why community data are so distributed, instead focusing on the characteristics of the data and its consequences for analysis. Similarly, other authors have written at length on related issues, such as the tendency for abundance data to be strongly right skewed and zero rich (Gaston and McArdle 1994; Anderson 2001; Peck 2010), the “zero truncation problem” (Beals 1984), and the double zero or joint absence problem (Legendre and Legendre 1998). None of these, however, directly address the processes that generate dust bunny distributions. It is important to evaluate potential causes of the dust bunny distribution, so that we can better anticipate the effect of study design on the intensity of dust bunnies in our data, promoting more effective analyses. For example, we can select sample unit sizes that are appropriate at the scale of the organisms and environmental factors of interest, such that they have enough in common and low enough dust bunny intensity to analyze them usefully. When dust bunnies are very severe, it becomes difficult to extract gradients of interest even with the most effective analytical techniques. Furthermore, because the dust bunny distribution is intimately related to beta diversity, understanding what factors affect the strength of dust bunny distributions also contributes to our understanding of what controls beta diversity in community samples.

Here we first define measures of the strength of the dust bunny distribution. These are simple, model-free descriptors of the tendency of species abundance data to lie along the high-dimensional edges of the underlying space. We then demonstrate that this pattern could result from any of four well-known ecological processes and examine the impact of each on the severity of the pattern. We do this with a simple population process model simultaneously applied to many species. Because we expect that a given statistical distribution can arise from very different processes, we anticipate that processes other than those that we examine could also create dust bunny distributions.

Our simulation model starts with unoccupied space, then builds populations for 30 species taking into account stochastic immigration, community-limited population growth, species-specific competitive ability, and variable species performance along multiple environmental gradients. We aim not for hyper-realism, but rather to expose simple mechanisms that may produce dust bunnies.

We evaluate possible causes of dust bunny distributions by generating a large number of non-equilibrial communities under varying conditions, then analyzing the relationship between the intensity of the dust bunny distribution in these data sets and the population and environmental drivers. We use our simulations of ecological community development to answer the following questions:

1. How does dust bunny intensity vary with time since disturbance? We hypothesize that stochastic initial colonization renders communities with strong dust bunny distributions, while competitive and environmental effects gradually sort communities into relatively consistent species composition, weakening the dust bunny distribution. The longer the time between disturbances, the more a community approaches equilibrium, assuming a stable environment. The less stable the environment, the less opportunity a community has to approach an equilibrial state.
2. How does dust bunny intensity vary with average niche width (and thus beta diversity) along environmental gradients? We anticipate that narrower niches create stronger dust bunnies, because a larger proportion of each environmental gradient will lie outside a species tolerances, producing more zeros and small values in the matrix.
3. How does dust bunny intensity vary with the degree of community-wide dispersal limitation? We hypothesize that dust bunny distributions can be created by dispersal limitations, with or without the influence of one or more environmental gradients. We anticipate that the stronger the dispersal limitation, the stronger the dust bunnies, because dispersal limitations heighten the relative importance of the stochasticity of immigration, producing many zeros in the data matrix, even in optimum habitats for a given species.
4. How does dust bunny intensity vary with number of influential environmental gradients? We anticipate that increasing the number of environmental controls on species increases the proportion of uninhabitable space for a given species, and thus increases the number of zeros in a data set.

Methods

Community Development Model

We model multiple population dynamics simultaneously and as simply as possible to produce dust bunny distributions (Online Resource 1; briefly summarized here). We develop communities in unoccupied space, simulating response to a community-replacing disturbance. Populations are described by the density or counts of a given species. 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).

Species have Gaussian responses on one or more environmental gradients. 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. This "competitive ability" 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).

Immigration is treated as a species-specific stochastic Poisson process, with immigration pressure varying linearly and negatively with competitive ability. To control the system-wide balance between growth rates and dispersal limitations, we introduce a dispersal limitation factor. This parameter is held constant for a given simulation, but can be varied to increase or decrease the dispersal limitation built into the whole community.

Community matrices.— Each community matrix (**A**, *n* sample units × *p* species), was assembled by running the model once for each sample unit, choosing the following parameters (Online Resource 1): 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. Abundances need not apply to species per se; they can also be higher taxa (genera, families) or frequencies of various genetic markers that are presumed to be shared by closely related organisms. The sample unit is normally a fixed area, fixed volume, or some other standardization of effort; the sample is taken from a variety of locations and/or dates.

Data adjustments.—Most analysts faced with real data sets similar to those analyzed here would transform the data before analysis. Because we visualize the data as counts, and the counts span several orders of magnitude, we transformed by $\log_{10}(a + 1)$ before calculating statistics that describe dust bunny strength. This transformation also preceded our beta diversity calculation that is based on the average Sørensen distance among sample units. Because transformations affect distributional properties, we expect that choosing other transformations could affect measures of dust bunny intensity.

Dust bunnies in High-Dimensional Spaces.— Assume an *n*-dimensional (*n*D) Cartesian coordinate system (i.e. of mutually perpendicular axes, with each axis intersecting the origin). Each pair of axes defines a plane. We define a corner as the intersection of two or more of these planes. For species data we need to consider only the non-negative part of this space (points with all coordinates ≥ 0).

A 2D corner is the intersection of two planes. The points (0,0,*z*) lie on a 2D corner in a 3D space. Similarly, the points (0,0,*y*,z) form a plane that is a 2D corner in a 4D space. Similarly, a 3D corner is the intersection of three planes. The point (0,0,0) lies on a 3D corner in a 3D space and the points (0,0,0,*z*) lie on a 3D corner in a 4D space.

Generalizing, a point in an *n*-dimensional space that has *k* coordinates with a value of zero lies on a *k*D corner of the *n*D space. Applying this to community data, a sample unit with *s* of *p* species lie on a (*p*-*s*)-dimensional corner of the *p*-dimensional species space. For example, a sample unit with 10 species in a data set with 50 species lies on a 40D corner of the 50D species space. In a typical data set, most sample units will be missing many of the species, most sample units lie on high-dimensional corners of the space. Like dust bunnies in a 3D room, dust tends to accumulate not just in the 3D corners of the space, but also in the lower dimensional (2D) corners.

Evaluation of Dust Bunny Intensity.— The intensity of multivariate dust bunny distributions in community data sets can be expressed by various statistics. Dust bunny distributions have high positive skew and kurtosis, but those statistics can also be high for non-dust bunnies. We used the following two simple measures of degree of match with a multivariate dust bunny, as defined qualitatively above.

1. Percentage of the community matrix that is zero. All species with no occurrences have been removed, such that the matrix has at least one nonzero value for each species. The maximum percentage of zeros is obtained with only one nonzero value for each of *p* species, which yields

100(1 – $p/(np)$) = 100-100/n% for a perfect dust bunny and 0% for the strongest anti-dust bunny, where n is the number of sample units. This measure ignores quantitative values in the matrix, in that the proportion of zeros is the same whether species are represented by presence-absence (1 or 0) or quantitative values. The expected value for the percentage of zeros approaches zero for a multivariate normal distribution.

2. We define a quantitative dust bunny intensity (DBI) as one minus the matrix mean when each species is relativized to (0-1) by the maximum ($amax_j$) value observed for that species, assuming that the smallest possible $a_{ij} = 0$ for each species:

$$DBI = 1 - \frac{\sum_{i=1}^{n} \sum_{j=1}^{p} a_{ij}/amax_j}{n \cdot p} \qquad (1)$$

We calculated DBI based on raw numbers as well as their logarithms, applied before relativization.

Empty species (all zeros) are removed before these calculations, because analysts would typically not include species that did not occur in the data set. Empty sample units are, however, retained, because they could be encountered in sampling and might be considered informative (e.g. recent disturbance or harsh environment). Because the standard proportional distance measures cannot be applied to this kind of problem (e.g. Bray-Curtis and chi-square distance require division by sample unit totals), one approach is to remove empty sample units. But analysts may wish to retain empty sample units as carrying part of the signal of interest, choosing a distance measure that accommodates them; therefore, we retain empty sample units here.

Empty sample units tend not to remain empty (McCune and Grace 2002, p. 38) because nature abhors a vacuum ("there does not exist a vacuum in nature"; Spinoza 1677). The multivariate origin is commonly vacant in community data sets while the extreme corner of a physical space usually has the highest concentration of dust. The dust bunny analogy is thus imperfect (McCune and Grace 2002). Nevertheless, after a sample unit is cleared of all species, it will begin to reaccumulate species, migrating out along the corners of the high dimensional space. For example, colonization of an empty sample unit by 3 of 50 species would push that point from the origin to a 47D corner of the 50D space.

The minimum DBI is zero for a perfect anti-dust bunny, where every species occurs at its maximum abundance in every sample unit. The maximum DBI is obtained with only one nonzero value for each species, which yields $1 - (p*1)/(n*p) = 1 - (1/n) \approx 1$ for large n. The expected value of DBI is 0.5 for a multivariate normal distribution. Because the normal distribution is symmetric about the mean, if each species is relativized from 0 to 1, the expected value of the mean is 0.5 for each species. Similarly, the expected value of DBI based on the log abundances is 0.5 for a multivariate lognormal distribution (log transformed, then relativized by $amax_j$)

An advantage of the percentage of zeros as a measure of dust bunny intensity is that it does not vary with most of the usual data transformations applied to species data, including $\log(a + 1)$, $\sqrt{a}$, relativization by species maximum, and relativization by sample unit totals. Conversely, the DBI responds to abundance patterns, which can contribute greatly to the apparent heterogeneity of a data set and influence the performance of ordination, clustering, and group comparison methods.

Real data sets for comparison.—To evaluate the similarity of distributional properties of simulated data sets to real data sets, we selected ten real data sets for comparison (Online Resource 2). These data sets were chosen arbitrarily, subject to the following constraints: (1) include abundances that are quantitative, rather than binary or with abundance classes, (2) represent a variety of taxa, (3) represent a variety of abundance measures, including counts, densities, areal cover, peak heights for molecular markers, and frequencies of DNA sequences detected.

Model Applications

Overall sensitivity to population processes.—We generated 19,200 data sets of 200 sample units × 30 species with 3 replicates per combination of level of dispersal limitation, niche widths, number of environmental gradients, and maximum number of time steps (Table 1; details of model inputs, outputs, and parameters in Online Resource 1). The model was implemented in the program DustBunny.dll (free add-in to PC-ORD 6; McCune and Mefford 2011; Fortran 90 source code available from McCune).

Community-level descriptors were calculated for each data set both before and after $\log_{10}(x+1)$ transformation. Those descriptors included our two dust bunny indices (percent of zeros and DBI), as well as Whittaker's beta diversity (β_w) and beta diversity in half changes (β_d), calculated by exponential transformation of the average Sørensen distances (D, Bray and Curtis 1957, Legendre and Legendre 1998, eqn. 7.57) among sample units within a data set:

$$\beta_d = \log(1 - D)\ /\log(0.5) \qquad (2)$$

Sensitivity analysis measured the importance of factors that we varied as model inputs (method in Online Resource 1). Sensitivities were analyzed by first fitting a multidimensional response surface for the DBI = f(inputs), 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 modeled response surfaces with nonparametric multiplicative regression (NPMR; McCune 2006) using a multiplicative Gaussian kernel with a local linear model.

Results and Discussion

Dust bunny intensity in real and simulated data

All community data sets, both simulated and real, departed strongly from multivariate normal and multivariate lognormal distributions. The two measures of dust bunny intensity, DBI and percent zeros in the community matrix, had a strong positive relationship (Fig. 2). That relationship was stronger in the case of log transformed data, because log transformation diminishes the effect of large abundance values, shifting the data toward presence-absence, the basis of the percent zeros metric.

Simulated data sets had dust bunny statistics similar to those in real data sets, whether analyzed as raw data or their logarithms (Fig. 2). For log-transformed data, four of ten real data sets fell below the cloud of simulated data sets, having lower DBI for a given percentage of zeros in the community matrix.

In all four cases, sample units in the community matrix were averages of a large number of spatial or temporal subsamples rather than individual sample units. While averaging tends to reduce both the percentage of zeros and DBI, it apparently has a larger effect on the DBI based on log-transformed data, such that DBI is lower than expected for a given percentage of zeros. Three of these four data sets had the lowest average species maximum, when expressed as standard deviations from the species mean. Thus relativizations by species maximum tend to yield relatively high matrix means when the original data consist of averages of many sample units. Log transformation exaggerates this effect, which, for positively skewed data, brings the species mean much closer to the maximum than in the untransformed data.

For 19,200 simulated data sets, the percentage of zeros and DBI ranged from 20.3−99.5% and 0.775−0.995 respectively. Average species richness ranged from near zero to 24. Beta diversity (β_d) ranged from 0.0−12.5 half changes.

Effect of log transformation on DBI

Log transformation of the abundance data diminished the intensity of the dust bunny distribution, for both real and simulated data (Fig. 2). In both cases, however, log transformed data still departed strongly from multivariate normality (Fig. 2). Although log transformation will often improve the ability of analytical techniques to extract pattern by deemphasizing very large values in the data matrix and providing more sensitivity at low abundances, log transformation is typically insufficient to achieve a distribution approaching multivariate normality.

Controls over intensity of dust bunny distributions

All four factors examined influenced intensity of dust bunnies, but varied greatly in effect size, in order of decreasing effect: number of environmental factors, dispersal limitation, time since community-replacing disturbance, and niche width. Each of these is addressed below.

Number of environmental gradients.— The strongest of the four factors examined was the number of influential environmental gradients. Sensitivity analysis showed that dust bunny intensity was over ten times as sensitive to the number of environmental gradients as to the next most important factor (Table 1). Simulated data sets with three strong environmental gradients always had DBI > 0.9, regardless of the settings of the other factors (Fig. 3).

Determining how many environmental gradients have affected the species in a real data set is difficult, if not impossible. Data reduction methods such as ordination commonly find 2-3 statistically supported dimensions, but these methods are designed to filter out the influence of weak gradients as noise (Gauch 1982). So although our multivariate analytical methods can reliably detect only a few underlying dimensions, surely there are many, in a declining series of importance, tapering down to obscure historical factors that left a slowly fading mark.

On the other hand, species distribution or habitat models provide good evidence of the underlying complexity of community data. One can slice a community data set into its component species, then individually fit models that relate those species to a common pool of predictors. Typically, many species will express a dominant environmental or disturbance gradient, but individual species will express habitat relationships that are ignored by other species, such that the list of predictors related to one or more species is long. Because the best models for different species models incorporate different predictors, even when they are made orthogonal as in principal components of climate variables, the communities must be influenced by numerous environmental gradients.

Dispersal limitation.— Dust bunny distributions were readily produced by dispersal limitations, even in the absence of environmental gradients. The intensity of dust bunny distribution rapidly increased with increasing dispersal limitation (Fig. 3). The relationship was nonlinear with a diminishing rate of increase as dispersal limitation increased.

Sensitivity analysis of the simulation model revealed that, apart from the number of influential environmental gradients, the degree of dispersal limitations most strongly controlled the departure from multivariate normality toward the dust bunny distribution (Table 1). This was true whether the intensity of the dust bunny distribution was measured as the proportion of zeros in the community matrix or as the dust bunny intensity metric (DBI).

Dispersal limitation, as defined here, tunes the balance between immigration pressure on the one hand, and competitive effects and intrinsic growth rates on the other. A high dispersal limitation means that the stochastic effects of immigration are large, relative to the intrinsic rates of increase in populations and the rate at which competition is expressed. If dispersal limitations are pervasive in natural populations, as suggested by many authors (e.g. Freestone and Inouye 2006; Ricklefs 1987), then this alone is sufficient to produce the dust bunny distributions typically seen in community data sets.

Time since disturbance.— Dust bunny intensity diminished with number of generations elapsed since disturbance (Fig. 3). This effect was most apparent with lower dispersal limitations and fewer influential environmental gradients. This means that strong dust bunny distributions are more likely when sampling communities soon after disturbance, relative to the generation time of the organisms. In other words, sampling early successional communities is more likely to yield strong dust bunny distributions than sampling old communities, where competition and immigration has have had a longer time to be expressed, yielding relatively deterministic and stable communities.

Note that "stability" and "time since disturbance" are both relative terms, an appropriate standard being the turnover rates of the organisms that comprise the communities. For example, a forest 300 years after stand-replacing fire may still be dominated by long-lived trees that first colonized after the fire (McCune and Allen 1985).

Number of generations elapsed was third in relative importance of the four factors examined, as shown by sensitivity analysis (Table 1). Dust bunny intensity, as measured by proportion of zeros in the community matrix, was about one-third as responsive to number of generations, as compared to dispersal limitation. Similarly, dust bunny intensity as measured by DBI was between one third and one half as responsive to number of generations as to dispersal limitation.

Niche width.— Narrower niches tended to result in stronger dust bunnies. This effect was similar in intensity for any number of environmental gradients, other than zero where niche width had no effect on dust bunny intensity (Fig. 4). Although niche width was the weakest of the four factors examined, its effect could be observed by holding both dispersal limitation and number of generations constant at moderate values, while allowing number of environmental gradients and niche width to vary (Fig. 4). Sensitivity analysis showed dust bunny intensity as measured by DBI to be about half as sensitive to niche width as to number of generations since disturbance, and only about a fourth as sensitive to niche width as to dispersal limitation (Table 1).

Relationships among dust bunny measures and beta diversity

The term beta diversity has taken on various meanings since Whittaker (1972); see reviews by Anderson et al. (2011) and Tuomisto (2010). In its most general sense, beta diversity measures heterogeneity of a community sample. Naturally, measures of beta diversity will tend to be correlated with the proportion of zeros in a community matrix. In fact, Whittaker's simplest beta diversity measure (β_{w}) that is applicable to any community sample can be considered a hyperbolic rescaling of the proportion of zeros in a matrix (PctZeros, Online Resource 3, Fig. 3-1). Specifically, start with Whittaker's $\beta_{\mathrm{w}} = (S_c/S) - 1$, where S_c is the total number of species in the sample and S is the average number of species per sample unit. If q is the number of nonzero elements in the matrix of sample units × species, then $q = \Sigma s_i$ across the i sample units and $S = q/n$. Then by algebra $\beta_{\mathrm{w}} = n\cdot S_c /q – 1$, PctZeros = 1 - $q/(n\cdot S_c)$, and β_{w} = 100/(100-PctZeros) – 1. Similarly, $S = p$(1 - PctZeros/100), where p is the number of species in the data set.

Where to find dust bunnies

Sousa (1984) wrote "The differential expression of life history attributes under different regimes of disturbance produces much of the spatial and temporal heterogeneity one observes in natural assemblages." Our models demonstrate that point, and suggest that the differential expression of life history attributes along environmental gradients provides a fundamental mechanism for producing the statistical properties of ecological community data. In other words, dust bunnies can be produced simply by interspecific variation in optima on environmental gradients combined with tradeoffs in life history characters (dispersal ability, competitive ability, and intrinsic population growth rates). It is likely, however, that dust bunnies can be observed even without tradeoffs in life history characters. For example, even under the assumption of ecological equivalence of all species, neutral theory holds that stronger dispersal limitation increases species turnover (Hubbell 2001) and thus results in more zeros in the data.

Dust bunny distributions were found in all real and simulated data sets examined, but these distributions varied greatly in their degree of departure from multivariate normality. Our model results lead to predictions about what kinds of community samples are likely to have weak or strong dust bunny distributions, and thus how seriously they will depart from multivariate normality. We list several examples below.

Sample units measured soon after community-replacing disturbance but within a narrow range of environments lead to strong dust bunnies. Even with little environmental variation, stochastic colonization by pioneer species and slow colonization of better competitors lead to a zero-rich data matrix.

Sampling near-equilibrial communities many generations after disturbance and from a narrow range in environments should lead to only weak dust bunnies. This would be the closest approximation to multivariate normal distributions that community ecologists are likely to find with data sets on natural (as opposed to experimental) communities. Experimentally-constructed communities are likely to have unnaturally few zeros, because experimenters typically introduce fewer species than encountered in nature.

Intensity of dust bunny distributions is driven both by population processes and properties of a sample. For example, when sampling within a given ecological domain, dust bunny intensity will depend on sample unit size (e.g. area or volume). Consider two extremes: infinitesimally small (point) sample units, and very large sample units that span multiple environments. If sample units are literally points in space, then only one species can occupy a point at a given time, and each sample unit will contain at most one species. Geometrically, this means that in a *p*-dimensional species space, the point lies on a *p*-1 dimensional edge of that space. After removing empty sample units, the DBI, percentage of zeros and beta diversity will be maximal.

On the contrary, if sample units are large, spanning substantial portions of major environmental gradients, then the length and steepness of the gradients are effectively diminished, beta diversity will decrease, and apparent niches will be broader. We have seen that reducing the environmental gradients in a data set will decrease the intensity of dust bunnies.

Community samples taken across multiple simultaneous strong environmental gradients are likely to show strong dust bunnies, regardless of the successional state, niche width of the component species, and degree of dispersal limitation. In our experience, most community data sets have multiple strong abiotic or biotic environmental influences, therefore strong dust bunnies are the norm in community ecology.

In conclusion, departure from multivariate normality toward the dust bunny distribution, characterized by community sample points lying along high-dimensional edges of multidimensional species space, can be predicted by population processes. Mechanisms that create this distribution include species' differential responses along multiple environmental gradients, stochasticity of dispersal limitation, time since disturbance, and ecological niche width of species. As usual it is difficult to infer process from pattern because often a pattern can be produced by more than one process. Because of this, observing a strong dust bunny does not allow us to infer which processes are acting on the community. We can, however, infer in the opposite direction: knowledge of population processes and sample characteristics allow us to anticipate the intensity of dust bunnies in our data. Furthermore, broader recognition of the dust bunny distribution should help us to choose analytical methods.

Acknowledgments

We dedicate this paper to David Goodall and his foundational work in ecology. We thank contributors of data; students and colleagues for helpful discussion; Dave Roberts for sharing a draft chapter on schools of community ecology, Amy Charron for dust bunny drawings, and Stéphane Dray, Andy Jones, Lisa Madsen, Patricia Muir and anonymous reviewers for critiquing the manuscript.

Literature Cited

Anderson MJ (2001). A new method for non-parametric multivariate analysis of variance. Austral Ecol 26:32-46.

Anderson M J, Crist TO, Chase JM, Vellend M, Inouye BD, Freestone AL, Sanders NJ, Cornell HV, Comita LS, Davies KF, Harrison SP, Kraft NJB, Stegen JC, Swenson NG (2011) Navigating the multiple meanings of β diversity: a roadmap for the practicing ecologist. Ecol Lett 14:19-28.

Beals EW (1973) Ordination: mathematical elegance and ecological naivete. J Ecol 61:23-35.

Beals EW (1984) Bray-Curtis ordination: an effective strategy for analysis of multivariate ecological data. Adv Ecol Res 14:1-55.

Bliss CI, Fisher RA (1953) Fitting the negative binomial distribution to biological data. Note on the efficient fitting of the negative binomial. Biometrics 9:176-200.

Bray JR, Curtis JT (1957) An ordination of the upland forest communities in southern Wisconsin. Ecol Monogr 27:325-349.

Cassie RM (1962) Frequency Distribution Models in the Ecology of Plankton and Other Organisms. J Anim Ecol 31:65-92.

Clarke KR (1993) Non-parametric multivariate analyses of changes in community structure. Australian J Ecol 18:117-143.

Freestone AL, Inouye BD (2006) Dispersal limitation and environmental heterogeneity shape scale-dependent diversity patterns in plant communities. Ecol 87:2425–2432.

Gaston KJ, McArdle BH (1994) The temporal variability of animal abundances: measures, methods and patterns. Phil Trans Roy Soc London B, Biol Sci 345:335–358.

Gauch HG (1982) Noise reduction by eigenvector ordination. Ecol 63:1643-1649.

Goodall DW (1963) The continuum and the individualistic association. Vegetatio 11:297-316.

Grime JP (1977) Evidence for the existence of three primary strategies in plants and its relevance to ecological and evolutionary theory Am Nat 111:1169–1194.

Kruskal JB (1964) Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. Psychometrika 29:1-27.

Legendre P, Legendre L (1998) Numerical Ecology. 2nd Ed. Elsevier, Amsterdam.

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

McCune B, Allen TFH (1985) Will similar forests develop on similar sites? Can J Bot 63:367-376.

McCune B, Grace JB (2002) Analysis of Ecological Communities. MjM Software, Gleneden Beach, Oregon, USA.

McCune B, Mefford MJ (2011) PC-ORD. Multivariate Analysis of Ecological Data. Version 6.08. MjM Software, Gleneden Beach, Oregon, USA.

Peck JE (2010) Multivariate Analysis for Community Ecologists. Step-by-Step Using PC-ORD. MjM Software Design, Gleneden Beach, Oregon, USA.

Pianka ER (1970) On r and K selection. Am Nat 104:592–597.

Ricklefs, RE (1987) Community diversity: relative roles of local and regional processes. Science 235:167–171.

Sousa WP (1984) The role of disturbance in natural communities. Ann Rev Ecol Syst 15:353-391.

Spinoza B (1677) The Ethics. Part I, Proposition XV, note. Transl. R. H. M. Elwes. Univ. of Adelaide ebooks.

Tuomisto H (2010) A diversity of beta diversities: straightening up a concept gone awry. Part 2. Quantifying beta diversity and related phenomena. Ecography 33:23–45.

Whittaker RH (1972) Evolution and measurement of species diversity. Taxon 21:213-251.

Whittaker RH, Gauch, HG Jr (1973) Evaluation of ordination techniques. Handb Veg Sci 5:287-321.

Table 1 Sensitivity of strength of dust bunny distributions to variation in dispersal limitation, niche width, number of environmental gradients, and number of generations since community-replacing disturbance. All combinations of levels of the model inputs were applied; "Incr" is the increment between levels. Sensitivities, *Q*, of the dust bunny indices were calculated by nudging the model variables for each data point, one variable at a time, then expressing the range in response relative to the amount nudged (Online Resource 1). *Q*=1 means that response variable (strength of dust bunny distribution) changes by an amount equal to the amount that the input variable was nudged. *Q*=0 means no response of DBI to the input variable. Dust bunny intensity (DBI) is based on the matrix mean of log-transformed abundances relativized by species maximum.

						Sensitivity, *Q*	
Model inputs	Units	Symbol	Min	Max	Incr.	% zeros	DBI
Dispersal limitation, slope of Poisson parameter vs. competitive ability	unitless	d	1	20	1	0.324	0.324
Niche width, environmental gradients	standard deviates	s	15	50	5	0.080	0.080
Number of environmental gradients	count	q	0	3	1	6.720	5.872
Number of generations	count	t_{max}	1	10	1	0.183	0.136

FIGURE LEGENDS – MCCUNE AND ROOT

Fig 1. The dust bunny distribution in ecological community data, illustrated with a simple hypothetical data set where abundances of three species form a series of unimodal distributions along a single environmental gradient (upper left). In species space (lower right), where each species' abundance defines an axis, the data form a dust bunny distribution, shown here with three levels of abstraction. Background: a dust bunny in the vernacular, an accumulation of lint and dirt particles in the corner of a room. Middle: sample units (dots) in 3D species space. Abundances of species 1 and 3 peak on the extremes of the gradient. Species 2 peaks in the middle of the gradient. Foreground: The underlying environmental gradient forms a strongly nonlinear shape in species space (adapted from McCune and Grace 2002).

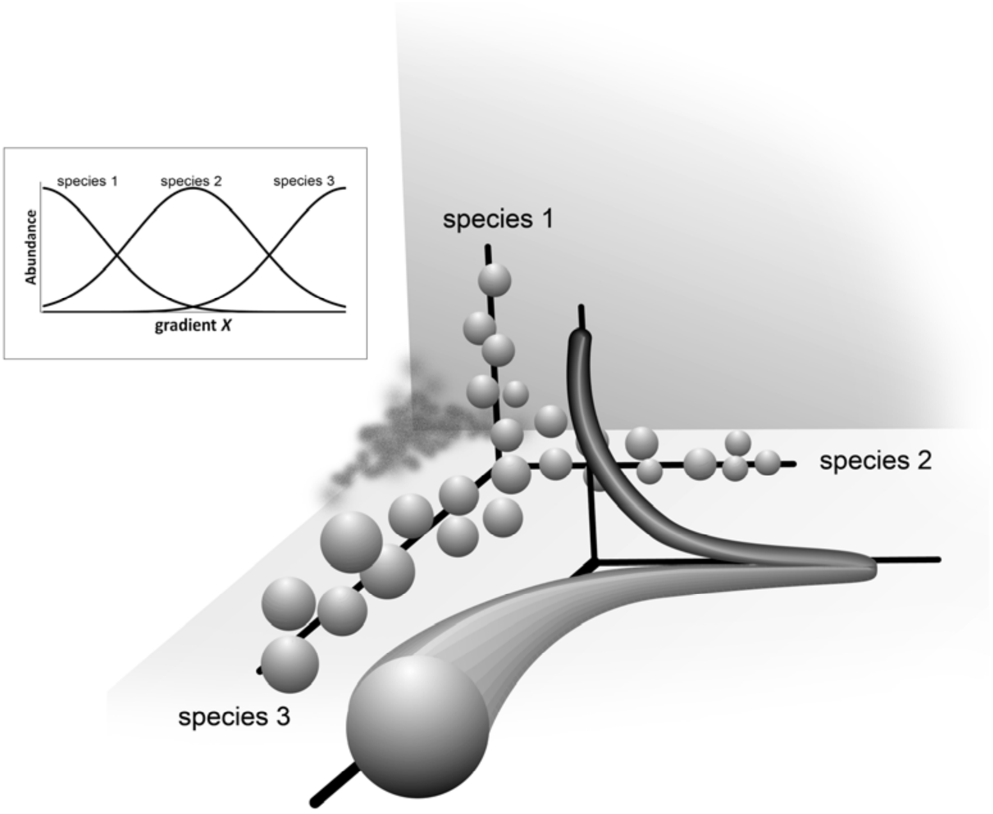


Fig 2 Comparison of ten real data sets (black dots; sources and characteristics: see Online Resource 2) with 19,200 simulated data sets (gray circles) for two measures of the departure of a dust bunny distribution from multivariate normality. Percent zeros in the species matrix vs. dust bunny intensity (DBI) based on raw data values (left panel) and log transformed species abundances (right panel). Expected values for multivariate normal and lognormal distributions are shown by **+**. Real data sets falling below the cloud of simulated data sets had raw data values consisting of averages rather than individual observations.

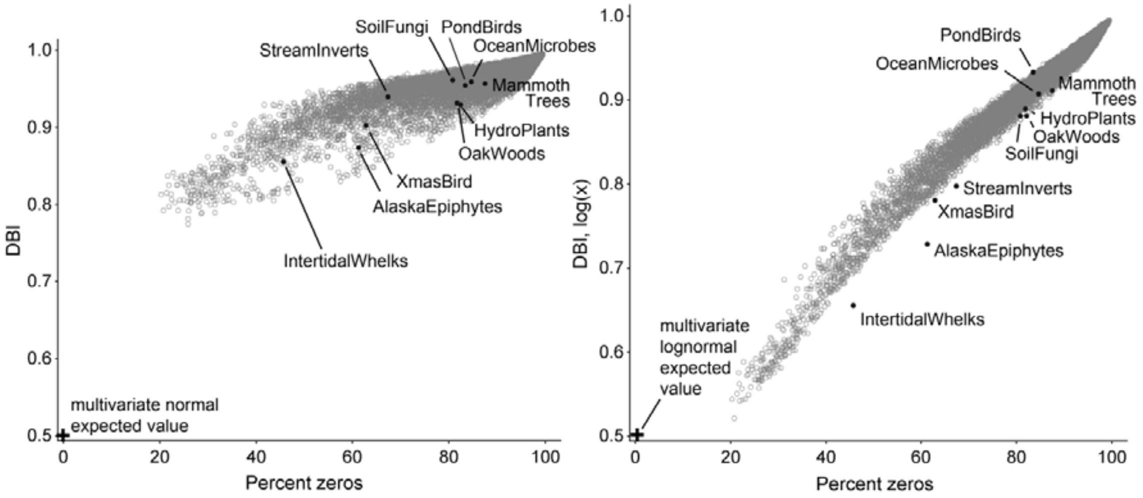


Fig. **3**. Response surfaces from simulated data sets showing the dependence of dust bunny strength (DBI) on the three strongest controlling factors, number of environmental gradients dispersal limitation, *d*, and number of generations (or time steps). A. zero environmental gradients. B. one gradient. C. Two gradients. D Three gradients. These factors push community samples from multivariate normal distribution toward a dust bunny distribution. All response surfaces are for a constant niche width, *s* = 25.

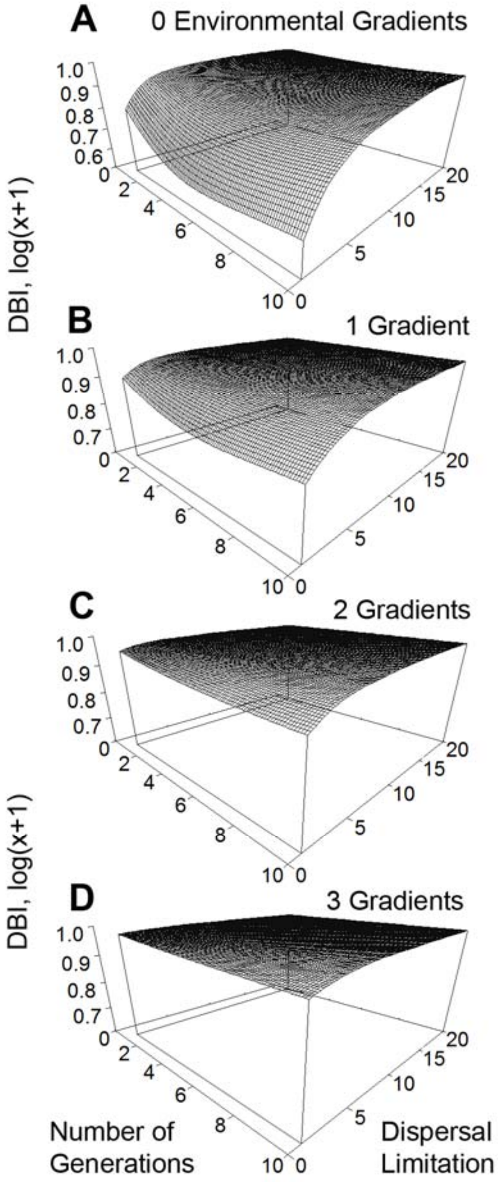
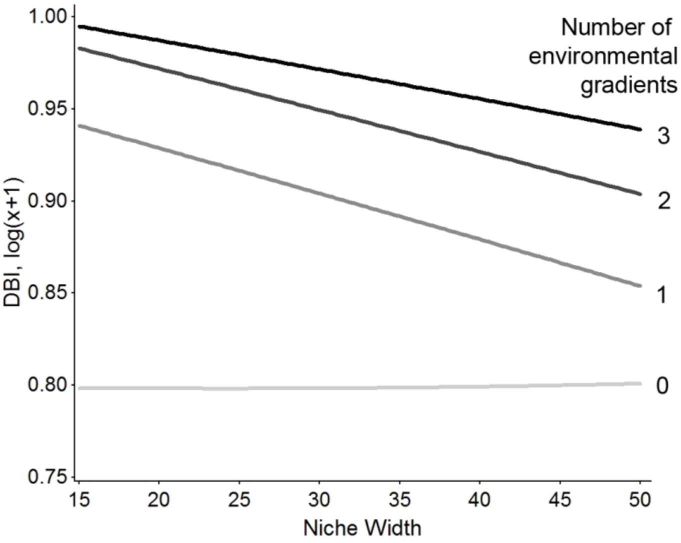


Fig 4 Dependence of dust bunny intensity (DBI) on the weakest controlling factor, niche width, for each number of environmental gradients. The other important variables were held constant at moderate values (number of generations = 3, dispersal limitation = 3).


DBI, log(x+1)
Niche Width
Number of environmental gradients