ECOLOGY LETTERS

Ecology Letters, (2011) **14**: 1273–1287 doi: 10.1111/j.1461-0248.2011.01685.x

REVIEW AND SYNTHESIS

Individual-scale variation, species-scale differences: inference needed to understand diversity

James S. Clark,[1,2,3*] David M. Bell,[1] Michelle H. Hersh,[2] Matthew C. Kwit,[1] Emily Moran,[1,2] Carl Salk,[1,2] Anne Stine,[1] Denis Valle[1] and Kai Zhu[1]

[1]Nicholas School of the Environment, Duke University, Durham, NC 27708, USA

[2]Department of Biology, Duke University, Durham, NC 27708, USA

[3]Department of Statistical Science, Duke University, Durham, NC 27708, USA

**Correspondence: E-mail: jimclark@duke.edu*

Abstract

As ecological data are usually analysed at a scale different from the one at which the process of interest operates, interpretations can be confusing and controversial. For example, hypothesised differences between species do not operate at the species level, but concern individuals responding to environmental variation, including competition with neighbours. Aggregated data from many individuals subject to spatio-temporal variation are used to produce species-level averages, which marginalise away the relevant (process-level) scale. Paradoxically, the higher the dimensionality, the more ways there are to differ, yet the more species appear the same. The aggregate becomes increasingly irrelevant and misleading. Standard analyses can make species look the same, reverse species rankings along niche axes, make the surprising prediction that a species decreases in abundance when a competitor is removed from a model, or simply preclude parameter estimation. Aggregation explains why niche differences hidden at the species level become apparent upon disaggregation to the individual level, why models suggest that individual-level variation has a minor impact on diversity when disaggregation shows it to be important, and why literature-based synthesis can be unfruitful. We show how to identify when aggregation is the problem, where it has caused controversy, and propose three ways to address it.

Keywords

Aggregated data, functional traits, individual variation, marginal and joint distributions, matrix models, Simpson's paradox, species coexistence, species interactions.

Ecology Letters (2011) **14**: 1273–1287

INTRODUCTION

The processes that control biodiversity operate at a different scale from most of the models and data used to study them. Species do not compete, individuals do. Species do not respond to climate, individuals respond to weather. The fact that important mechanisms operate on individuals does not mean that only individual-level data provide insight. Some processes operate and thus should be studied at coarse scales (e.g. atmospheric circulation). Highly aggregated variables at coarse spatio-temporal scales contribute perspectives that could not have been obtained from experiments on individuals (MacArthur 1972; Brown & Maurer 1989; the volume edited by Ricklefs & Jenkins 2011 is an important recent review). But the critical scales for a process can be overlooked for decades in scientific debates, as efforts to understand and predict biodiversity still rely heavily on highly aggregated data and models, often without consideration of how aggregation itself can preclude further progress. For example: 'How do many late successional species coexist?' Decades ago, ecologists could explain changes in aggregate functional types over succession (e.g. shade-intolerant to shade-tolerant species) and accumulation of aggregate biomass. Despite proliferation of data sets and efficient algorithms, confident predictions still do not go much beyond aggregate biomass and a few functional types, i.e. aggregates of individuals responding to spatially and temporally aggregated variables. Or: 'How will longer growing seasons affect species distributions given that competition for reduced moisture depends on moisture availability?' Climate change predictions are still dominated by spatial calibration of highly aggregated variables – species distributions and regional climate. And: 'Where are all the niches in communities that appear to be dominated by only a few limiting resources?' Still, no models generate high diversity of competing species, unless each is explicitly guaranteed its own niche. Quantifying the strength of species interactions is a research priority (Agrawal *et al.* 2007; Novak *et al.* 2011), and it is studied with models that assume that species rather than the individuals interact. Aggregate species-abundance and species-area distributions do not discriminate between coexistence mechanisms (Nee *et al.* 1991; McGill 2003; Clark 2011; Warren *et al.* 2011), but they remain a favorite for testing theory. Parameter estimates for the most popular models for aggregate population growth (matrix models and integral projection models – IPMs) are usually not constrained by data on population growth (S. Ghosh, A. Gelfand, J.S. Clark and K. Zhu, unpublished data).

Studying the scale of interest, rather than the process that controls it, introduces aggregation problems, like the 'ecological fallacy' recognised in statistics and the social sciences (Bickel *et al.* 1975; Scheiner *et al.* 2000; Clark 2003; Ibanez *et al.* 2006). This term refers to the fact that group-level data can hide and misrepresent individual behaviour. These disciplines recognise that inference about a relationship or process has to derive from information at the scale where it operates or risk aggregation problems (Wakefield & Salway 2001). Typically, individual organisms are not of interest, but that is the scale where competition and important responses to weather occur. Global warming renewed interest in how statistics on weather (climate) relate to statistics on organisms (species abundance). The aggregate statistics shed light on global patterns of climate, adaptation and biogeography. However, the numbers and identities of species

vulnerable to climate change depend on individuals responding to weather in a competitive setting. Important attributes of weather and its effects on competing individuals do not survive the data and model aggregation used to study them. They can be fundamentally misrepresented in aggregate data, as when a species appears negatively correlated with moisture due to the effects of moisture on a natural enemy, host plant, or competitor, or the variation that controls responses and interactions are lost in the averaging. Factors that determine vulnerability and the aggregate outcome may not be interpretable from or predicted by the aggregate.

In studies of biodiversity, aggregation affects interpretations when (A) observations obtained from individual organisms are used to obtain species- or community-level summaries, which then become the basis for interpretation, and (B) when data on a species are abstracted from the full community, analysed independently, and then used to predict properties of ecological communities. Approach A entails aggregating experiments and measurements on individuals to produce estimates of parameters and traits for a species (Clark *et al.* 1999; Reich *et al.* 1999; Bolnick *et al.* 2003; Rozendaal *et al.* 2006; Shipley *et al.* 2006; Westoby & Wright 2006; Ackerly & Cornwell 2007; Messier *et al.* 2010). Approach B entails fitting models or conducting experiments independently for one or a few species, then using the independently fitted models, for example, to predict diversity, i.e. the aggregate behaviour. Niche models (Peterson *et al.* 2002; Guisan & Thuiller 2005; Thuiller *et al.* 2005; Levinsky *et al.* 2007; Buckley *et al.* 2010), invasion experiments (Pathikonda *et al.* 2008; Pyšek *et al.* 2008) and dynamic simulations based on parameter values culled from the literature are examples where analyses of individual species are the basis for predicting community response. In B, the relationship between a species distribution and climate or competition with an invader is conditional, depending on the context of observations, a context that is not carried forward when results from individual species are extrapolated to community-level predictions. For example, there is no reason to expect that tree abundance predictions based on niche models should even predict a closed canopy (fully occupied) forest (see Diversity prediction based on independently modelled species). Aggregation and abstraction change relationships in ways that can make species look the same, reverse species rankings along niche axes, lead to the surprising prediction that a species decreases in abundance when a competitor is removed (see Diversity prediction based on independently modelled species), or simply preclude parameter estimation. Niche differences that are hidden at the species level (Condit *et al.* 2006; Wiegand *et al.* 2007) become apparent upon disaggregation to the individual level (Clark 2010). The inevitable aggregation that comes with synthesis of published literature often is not directly relevant to the scale of the process of interest.

Where possible, the ideal solution is often to 'analyse, then aggregate', rather than 'analyse the aggregate'; this may not be possible, but more often, the advantages can be simply unrecognised. Ecologists have studied the demographic responses of individuals for a long time, but the species-level parameters estimated in these studies aggregate over the variation in individual responses. The individual variation is available and contains critical information, but only the aggregate is quantified. This is a natural tendency, given that we care about species, not individuals. However, the demonstrations that competition is in fact concentrated within, rather than between, species (Clark 2010), and the differential vulnerability of species to climate change (Clark *et al.* 2011) came from analysis of the individual-scale variation. In these cases, analysis of individual level variation was followed by aggregation to the species level of interest. The approach is complementary to, but not the same as, individual-based modelling (IBM), the tracking of individuals in forward simulation models (DeAngelis & Mooij 2005). Individuals in IBMs usually have identical parameter values; our focus is on the missed opportunity that comes from ignoring the joint distribution of individuals when learning about their responses. We demonstrate how aggregation causes confusion and why it underlies debates in ecology, and we provide options for addressing problems caused by aggregation.

The terms 'aggregate' and 'marginalise' are related, sometimes used interchangeably, but they have distinct meanings. 'Aggregation' describes when observations (e.g. a point pattern in Fig. 1a) are summarised by attributes of the group or by models that apply to group characteristics. 'Marginalisation' typically refers to distributions or models of distributions. As an example, consider observations or a model of them in dimensions x and y. The joint distribution is 'disaggregated', consisting of a point pattern (x_i, y_i) for individuals $i = 1, \ldots, n$ (Fig. 1a) or a density function $p(x, y)$ (Fig. 1b). Aggregation occurs when objects that occupy a high dimensional space are

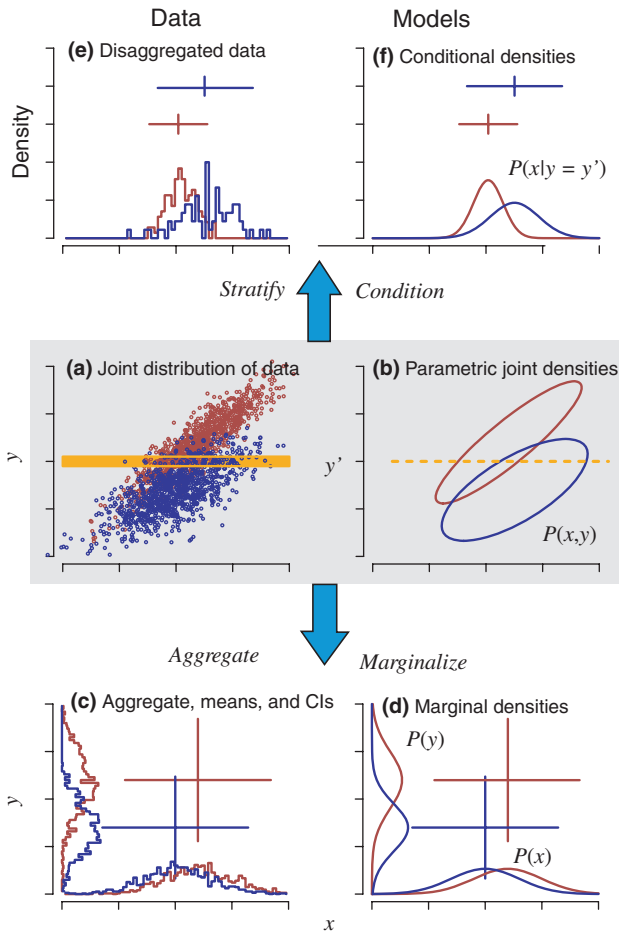

Figure 1 Point pattern (left) and density (right) for two species plotted on axes x (e.g. light) and y (e.g. soil moisture). The joint distribution of observations in (a) and parametric distributions in (b) show species differences. Aggregate and marginal densities (c, d) show the tendency for brown to occur at higher levels of both x and y. The standard technique of projecting marginal densities as crosshairs in (c) and (d) suggests broad species overlap and promotes the interpretation that the brown species occurs at higher levels of x and y. Conditional densities (e, f) show separation and the blue species to have the larger mean response.

projected on fewer dimensions or they are summarised by a mean value. Data aggregation occurs when dimensions are ignored to focus on x_i or y_i, rather than on (x_i, y_i) (Fig. 1c). Marginalising a model involves integration or summation (Fig. 1d). 'Stratification' (Fig. 1e) and 'conditioning' (Fig. 1f) refer respectively to data and models when interest concerns a response in one dimension, given the response in another. Once data have been aggregated, disaggregation may not be possible.

A closer look at Fig. 1 reveals an aggregation problem. Brown and blue can represent two species, both of which respond to light and moisture, labelled x and y. Individuals constitute a point pattern in Fig. 1a, summarised by a model in Fig. 1b. On average, the brown species responds at higher levels of both (Fig. 1c,d). However, at any given moisture level, the blue species responds at higher average light levels (conditionals in Fig. 1e,f). This is an example of Simpson's Paradox, where the species aggregate view leads to a conclusion different from one based the individuals (Bickel *et al.* 1975). In Fig. 1, the conditional relationship is available at the disaggregated scale. Aggregation caused the loss of information (distributions hardly overlap in Fig. 1a,b, but hardly differ in Fig. 1c,d), and qualitatively changed their relationship (reversal from Fig. 1c–e, d–f). Said another way, we could not infer the individual-scale process (e.g. competition) from the aggregate response. The aggregate pattern is generated by processes operating at the individual scale, but appears to misrepresent them. Aggregation problems are not restricted to particular types of dimensions, including geographical location, time, trait space, resource abundance or utilisation, and physiological or demographic responses. The loss of information (Wakefield & Shaddick 2005) results from the fact that it is difficult or impossible to recover a joint distribution from marginals (Gelman & Speed 1993; King 1997; Nelsen 1999), unless there is the possibility of disaggregation.

HOW AGGREGATION PROBLEMS ARISE

There are not always solutions to the aggregation problem, but the pervasiveness of the problem, common pitfalls and options that could be used to address it can benefit from a broader recognition. We begin with a summary of key elements, followed by ways to identify when aggregation has occurred and how to accommodate it. We discuss why it becomes increasingly important as dimensionality increases. We discuss examples of three options, including (1) disaggregating when you can, (2) disaggregation how you can or (3) marginalising the model to accommodate the aggregation in data. We then discuss why independent analysis of species represents a conditional model that can lead to misleading predictions at the community scale. Finally, we discuss advantages of data collection and analysis at the disaggregated scale where critical processes occur, followed by aggregation to the scale of interest.

Aggregating data, marginalising a model

Aggregating data or marginalising a model can cause information loss and change conclusions. For clarity, we illustrate the problem with just two dimensions, which might describe attributes of species (traits, demographic rates), the environment to which they respond (climate, resources, natural enemies), or even more complex data such as the distribution of these quantities in space, time, frequency and so forth. To introduce theory, we use a simple example. We examine whether beech occupies wetter or drier sites than red oak. The Forest Inventory and Analysis data of USDA Forest Service provide an opportunity to analyse relationships between species distributions and climate (Prasad *et al.* 2007; Canham & Thomas 2010). Seedlings of both species plotted against winter temperature T and annual precipitation P for plots $i = 1, \ldots, n$ constitute a point pattern $p(P_i, T_i)$. This is an empirical distribution of observations. A model could be fitted to this point pattern having joint density

$$p(P, T) = p(P|T)p(T) = p(T|P)p(P) \quad (1)$$

This joint density is factored on the right-hand side into a conditional and a marginal distribution. A conditional distribution of P is taken at a slice through the joint distribution at a specific value T' (Fig. 2c,d). These are shown for the fitted distribution, but could also be constructed for the point pattern of observations, by stratifying as in Fig. 1a,e.

The point pattern could be aggregated by ignoring T to produce a point pattern in one dimension $p(P_i)$. This data aggregation could be accommodated by marginalising the joint density over dimension T (Fig. 2e). Rather than slice through $p(P, T)$ at a value T', we now integrate away the variable T,

$$p(P) = \int p(P, T)dT = \int p(P|T)p(T)dT \quad (3a)$$

If T takes discrete values, this marginal is obtained by summation,

$$p(P) = \sum_T p(P|T)p(T) \quad (3b)$$

This is a mixture of two variables, obtained by integrating over the variation in P that occurs across the full range of temperatures. The marginal distribution of P is at least as variable as any conditional distribution of P, because it is not restricted to the variation observed at a specific temperature T'. Not only does the variation increase from conditional to marginal, but the rank of species mean values reverses from marginal in Fig. 2e to conditional in Fig. 2c. The answer to the original question (which species occupies the wetter sites?) is, in aggregate, beech. This is the answer we obtain if we ignore temperature. Conditional on warm winters, the answer is red oak.

One could object to the example in Fig. 2 on the ground that species were selected specifically to show a paradox that might rarely be observed. These species were not selected arbitrarily, but we did not have to look hard for examples. There is nothing strange about the distributions (the fitted models are Gaussian), and the relationships apply to both fitted and the empirical distributions. The information loss that comes with aggregation is general. Conditioning and aggregating are a part of all field studies and most modelling studies. All observations are conditional, depending on the setting in which they were obtained. On the other hand, all observations marginalise over variation during the study (see Options for addressing the aggregation problem).

It is worth mentioning that the large literature on 'scaling' (Weins 1989; Levin 1992; Underwood *et al.* 2005) represents part of a broader challenge we consider. Aggregation can involve Jensen's inequality, perhaps the most commonly discussed 'scaling problem' in ecology (Melbourne and Chesson 2005; Ruel & Ayres 1999), but it is more general. Jensen's inequality concerns error introduced when nonlinear functions of stochastic variables are summarised by mean behaviour (Flyvbjerg *et al.* 1993; Berec 2002). Aggregation is problematic whether or not relationships are nonlinear. Relationships can be linear, empirical point patterns (there is no function in Figs 1 and 2), contingency tables (including only two classes as in Simpson 1951),

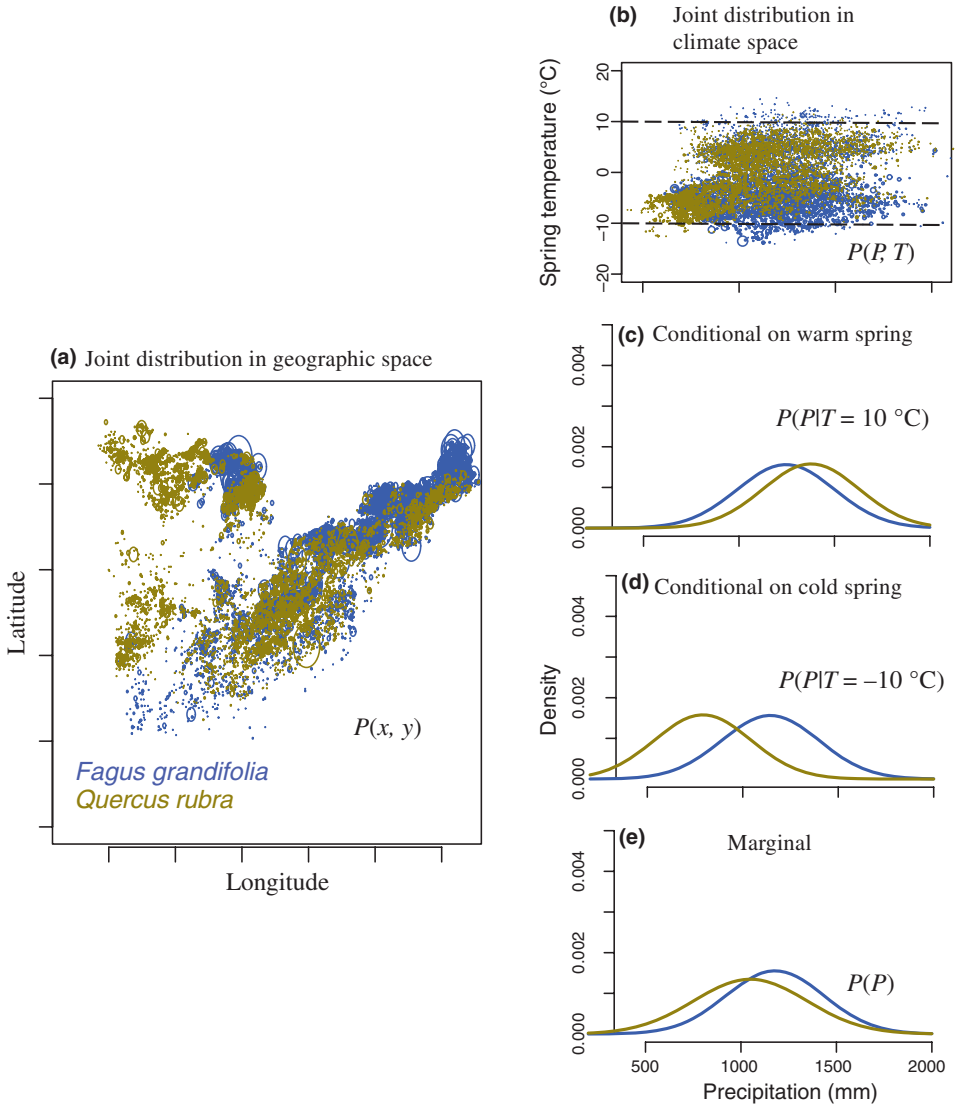

Figure 2 Joint distributions for recruitment of two species in geographical space (a) and climate space for temperature T and precipitation P (b), with symbol size scaled as density the Forest Inventory and Analysis data (Zhu *et al.* 2011). Rank reversals in climate space occur when moving from conditional distributions for warm (c) and cool (d) climates [dashed lines in (b)], and marginally (e). Empirical distributions are shown in (a) and (b). Parametric (bivariate normal) distributions from (c) through (e).

and classes can be nominal (e.g. political party, religion, race). However, they all entail changing dimensionality and produce misleading results whenever there are interactions with unobserved variables (Wagner 1982).

Increasing dimensionality increases aggregation problems

Failure to find correlations between a few environmental variables and species abundance often leads to the conclusion that species are 'neutral'. The apparent sameness is an aggregation problem, transparent in two dimensions (Figs 1 and 2), but progressively obscured with increasing dimensionality. In just two dimensions, there are few ways in which species that are similar marginally can differ jointly. When there are many dimensions, the opposite is true. The probability, that species overlapping in a few dimensions overlap jointly in, say, 10 dimensions, is vanishingly small. There are just two expressions for a binary trait in one dimension, but there are more than 1000 in 10 dimensions. Each new dimension brings combinatorial complexity in the number of ways objects differ. Marginalisation promotes the illusion of overlap and distorts the relationships of interest. Many ecological processes inherently operate at the individual scale (competition, disease, response to weather). High dimensionality comes from the fact that individuals are subjected to combinations of inputs (Clark 2010). Aggregation is the reason why failure to find differences among species is not evidence for sameness.

OPTIONS FOR ADDRESSING THE AGGREGATION PROBLEM

The most powerful solution to the aggregation problem is to analyse at the disaggregated scale, then 'aggregate' the results to the scale of interest. Often, disaggregation can be difficult, but just as often the opportunity is simply overlooked. Examples of options for addressing the aggregation problem that follow depend on data and the scale at which the process operates.

Option 1: Model at the disaggregated scale

The most direct solution to aggregation problems is to analyse at the disaggregated scale, fully exploiting information contained in the joint distribution. Consider the study of tradeoffs in traits or demographic responses that could explain coexistence. Each species is treated as an observation in trait space (e.g. Turnbull 1991; Kitajima 1994; Wright 2002; Baraloto *et al.* 2005; Seiwa 2007; Clark *et al.* 2010; Poorter *et al.* 2010) summarised as marginals in one dimension or as crosshairs in two dimensions (Fig. 3). This approach is used to compare responses to resources, such as growth in high light vs. survival in low light. Some studies find evidence of this tradeoff, but others do not (Welden *et al.* 1991; Walters & Reich 1996; Wright 2002; Baraloto *et al.* 2005; Valladares & Niinemets 2008; Clark *et al.* 2010). Plots like Fig. 3 can be misleading, suggesting joint distributions where there are only marginals – the crosshairs contain no more information than densities plotted along the margins. The joint relationship is available at the individual level, but its value is rarely recognised or exploited.

Aggregation not only obscures mean differences, but also masks responses to the environment in fitted models. For example, physiological responses to light and CO_2 determine succession, species coexistence and responses to global change (Tilman 1988; DeLucia & Thomas 2000; LaDeau & Clark 2006). Despite obvious species differences in these responses (Fig. 4a), models fitted to long-term species-level data provide a misleading view that these resources are inconsequential and that different species respond to them in essentially the same, weak fashion (Fig. 4b), certainly not with the differences required for coexistence in competition models (Fig. 4c) (Bazzaz 1979; Tilman 1988). Yet, these are just two of many differences that are obvious at the physiological level, but hidden when aggregated over individuals and over time (Fig. 4b).

The interpretation that species are the same is based on the overlap in aggregate (Figs 3 and 4b). Responses g (photosynthetic rate or seedling growth) in Fig. 4 depend on light (L) and CO_2 (C). Figure 4b marginalises over the variables x that were not measured,

$$p(g|L,C,\Omega) = \int_{\Omega} p(g|L,C,x)p(x)dx \qquad (4a)$$

$p(x)$ is the density of inputs x, which varied over a range of values Ω during the study. Not only is x unknown, but so too is Ω – we do not know what we are marginalising over, nor what we are conditioning on. We do know that the higher the dimension of x (number of variables affecting g) and the wider the range during the study Ω, the broader the marginal distribution and, thus, the more overlap between species. Increasing dimensionality and range degrade the fit as unknown x increasingly smears over contributions from known L and C.

By contrast, short-term and tightly controlled experiments limit the range of variation Ω to the conditions prevailing during a short time and small area, call it Ω' (Fig. 4a comes from carefully selected conditions). This conditional distribution

$$\lim_{\Omega \to \Omega'} \left(\int_{\Omega} p(g|L,C,x)p(x)dx \right) \to p(g|L,C,\Omega') \qquad (4b)$$

shows species differences. The range of integration Ω (left side of eqn 4b) becomes so small that the important conditional differences between species (right side of eqn 4b) become evident. However, there are many possible conditional distributions, one for each Ω'.

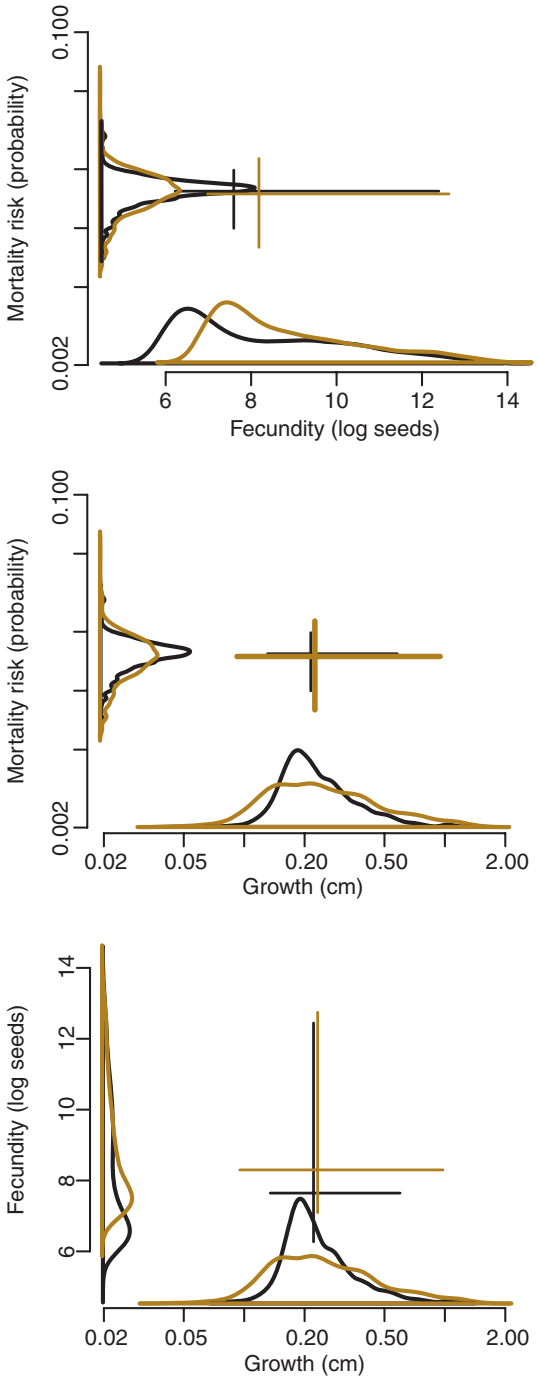

Figure 3 Density plots for demographic rates of two tree species from the analysis of Clark *et al.* (2010) broadly overlap, taken over individuals and years. Trait data are typically analysed based on such marginal relationships. Crosshairs for combinations locate means and span 95% of the estimates.

Here, again is the dimensionality paradox: the more ways there are to differ, the more they appear the same (the greater the overlap in aggregate).

The wide 95% predictive intervals in Fig. 4b result primarily from variation among individuals (Mohan *et al.* 2007). Despite the broad overlap at the species level, the prediction intervals for individuals are narrow. In other words, there is substantial information about the response, once inference shifts from the species-level aggregate to

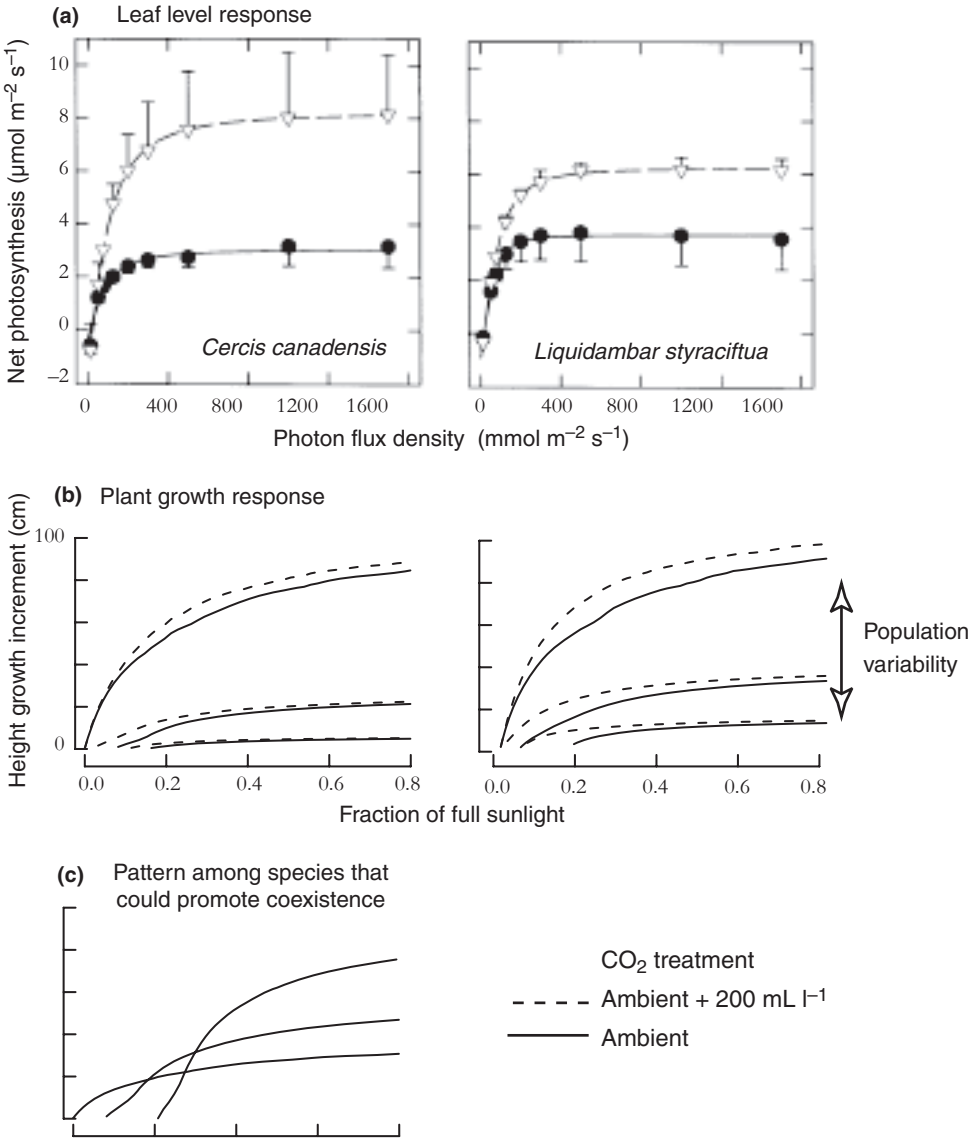

Figure 4 Short-term leaf-level responses (a) show clear effects of and different species responses to light and CO_2 (from DeLucia & Thomas 2000). Long-term growth rates of trees from the same sites (b) show little effect of either variable and imperceptible species differences (solid lines include mean and 95% of individual responses for ambient (solid lines) and elevated (dashed lines) CO_2 (from Mohan *et al.* 2007). These species coexist throughout eastern North American forests, but aggregated relationships in (b) do not show the relationships required if coexistence depends on partitioning light (c).

the individuals that make up that species. An example of how to quantify species differences at the individual scale is available in J.S. Clark, B. Soltoff, A. Powell & Q. Read (unpublished data), who examined whether or not there is evidence for the negative correlations between understory and gap responses of species required for coexistence in colonisation–competition tradeoff models. They found that all species grow faster in gaps than in the understory on average, but that they differ in their joint distributions of responses to understory and gap. We summarise how the understory/gap inputs can be disaggregated to a joint distribution of responses for individuals of each species and how those distributions relate to unobserved variables.

Consider a bivariate Gaussian model for observations of $[u, g]_{is,t}$, the growth response vector for individual i of species s at time t,

$$N_2([u, g]_{is,t} | \alpha_s, \Sigma_s.) \quad (5)$$

Examples of bivariate Gaussian distributions are shown in Figs 1b and 5b. The bivariate response in eqn 5 consists of $u_{i,t}$, the understory growth rate and $g_{i,t}$, the change in annual growth rate from gap formation (positive or negative). The first parameter to the right of the vertical bar is the vector of two mean values α_s. The second parameter is the 2×2 covariance matrix of residuals Σ_s. Of course, many variables affect the response that are unmeasured. We represent them here with a vector x. The question is: How can we learn about the contribution of x?

Conditioning on individuals can provide some insight into the role of hidden variables. At minimum, the covariance in eqn 5 contains contributions from observation error E_s and individual differences V_s,

$$\Sigma_s = E_s + V_s$$

For heuristic purposes, the understory/gap mean response vector for an individual α_{is} in eqn 5 can be expressed in terms of the variables

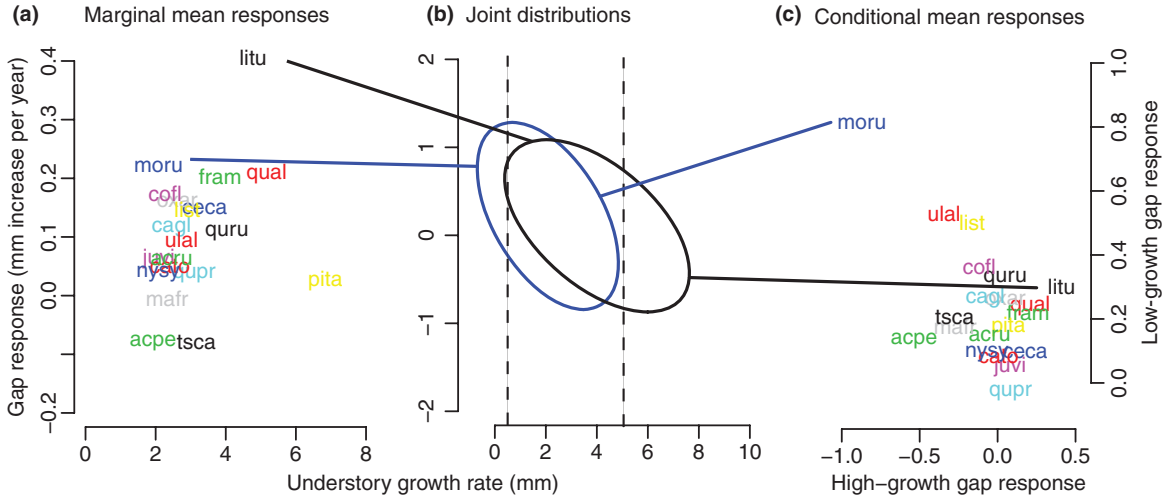

Figure 5 (a) The marginal mean growth responses to understory and gap do not show the negative correlation required for coexistence, but instead are positive ($r = 0.65$, $P = 0.003$). (b) Joint distribution of individuals from eqn 8 for two species. (c) Conditional means for gap response for low and high understory growth response [vertical lines in (b)] show negative correlation ($r = -0.51$, $P = 0.023$).

that affect it, say $N_2(\alpha_{is}|x_i\beta_s, H_s)$. The mean vector from eqn 5 now has a mean vector of its own, determined by a length-k vector x_i, and parameters in the $k \times 2$ matrix β_s. For example, if the light and moisture levels vary within gaps and understory, and these variables are not measured, then one of the x's could be light and another could be moisture. Covariance matrix H_s takes up variation in α_{is} not accounted for by the mean due to the model simplicity and missing sources of variation (e.g. genetic). The distribution of responses depends not only on what these inputs are but also on their distribution within the study $p(x) = N_k(x|\mu, V_x)$. This distribution for the inputs x has a length-k mean vector μ and covariance matrix V_x. By marginalising the model as in eqn 4a, eqn 5 can equivalently be expressed as

$$[u,g]_{is,t} \sim N_2(m_s, E_s + V_s) \tag{6}$$

where

$$m_s = \mu\beta_s$$

$$V_s = H_s + \beta_s^T V_x \beta_s$$

This alternative way of writing eqn 5 interprets the response in terms of the mean vector μ and covariance V_x of the hidden variables x, but it does not help us identify those effects or determine if they are meaningful. However, by conditioning on the individual, we can reduce the overall variation while learning about the relationships between responses across individuals. For individual i, we have the density

$$p([u,g]_{s,t}|i) = N_2([u,g]_{s,t}|m_s + b_{is}, \Sigma_s.) \tag{7}$$

where m_s has the definition from eqn 6, and the individual coefficient vector is $b_{is} = (x_i - \mu)\beta_s$. At the population scale, the distribution of individual effects is now $N_2(b_{is}|0, V_s)$, with overall population covariance reduced to $\Sigma_s = E_s$. The individual effect b_{is} includes departure of x_i from the mean vector μ, as translated by the species effect β_s. This conditioning on individual i can result in a large reduction in the residual variance in the error covariance Σ_s, because much of the variation is associated with individuals (Clark *et al.* 2003; Clark *et al.* 2010). Variance is reduced to E_s because x varies at the individual scale. This can be interpreted as a random effects model, commonly applied to longitudinal studies of individuals (Clark *et al.* 2003).

The gain in understanding that comes from moving from eqns 5 to 7 depends on whether or not there is individual variation and how long individuals are studied. Contained within eqn 7 is a joint distribution of individuals

$$N_2(\alpha_{is}|m_s, V_s) \tag{8}$$

where V_s describes tendency for responses within the vector α_{is} to covary among individuals, within the species. Equation 5 is a model for observations where no attention is paid to which individual they came from, whereas eqn 7 has structure. If there is no variation among individuals, or they are not tracked over time, then little is gained by conditioning on individuals. From eqn 6, it is clear that covariances can be positive or negative, depending not only on the covariance structure of H_s, but also on the environmental variables V_x and on how environmental variation is translated by each species, β_s. Equation 8 is useful when the process of interest operates at the individual scale, as it does here (individuals competing in the understory and gaps).

A standard comparison of species is a plot of mean values in α_s from eqn 5 for understory and gap responses for species in aggregate (Fig. 5a), fitted as a Bayesian hierarchical model (Clark *et al.*, unpublished data). The positive correlation is not the pattern that would promote coexistence in models – species that grow fast on average in the understory also respond most on average to gaps. However, the marginal mean values in Fig. 5a represent an aggregate value for the entire species. They are not the important summary because species do not compete. The conditional distributions (Fig. 5c) are the relevant perspective, and they tell a different story – species having individuals that respond most to gaps when growing slowly in the understory are not the same as those that respond most to gaps when growing more rapidly in the understory. The example distributions in Fig. 5b show that the fastest growing individuals in the

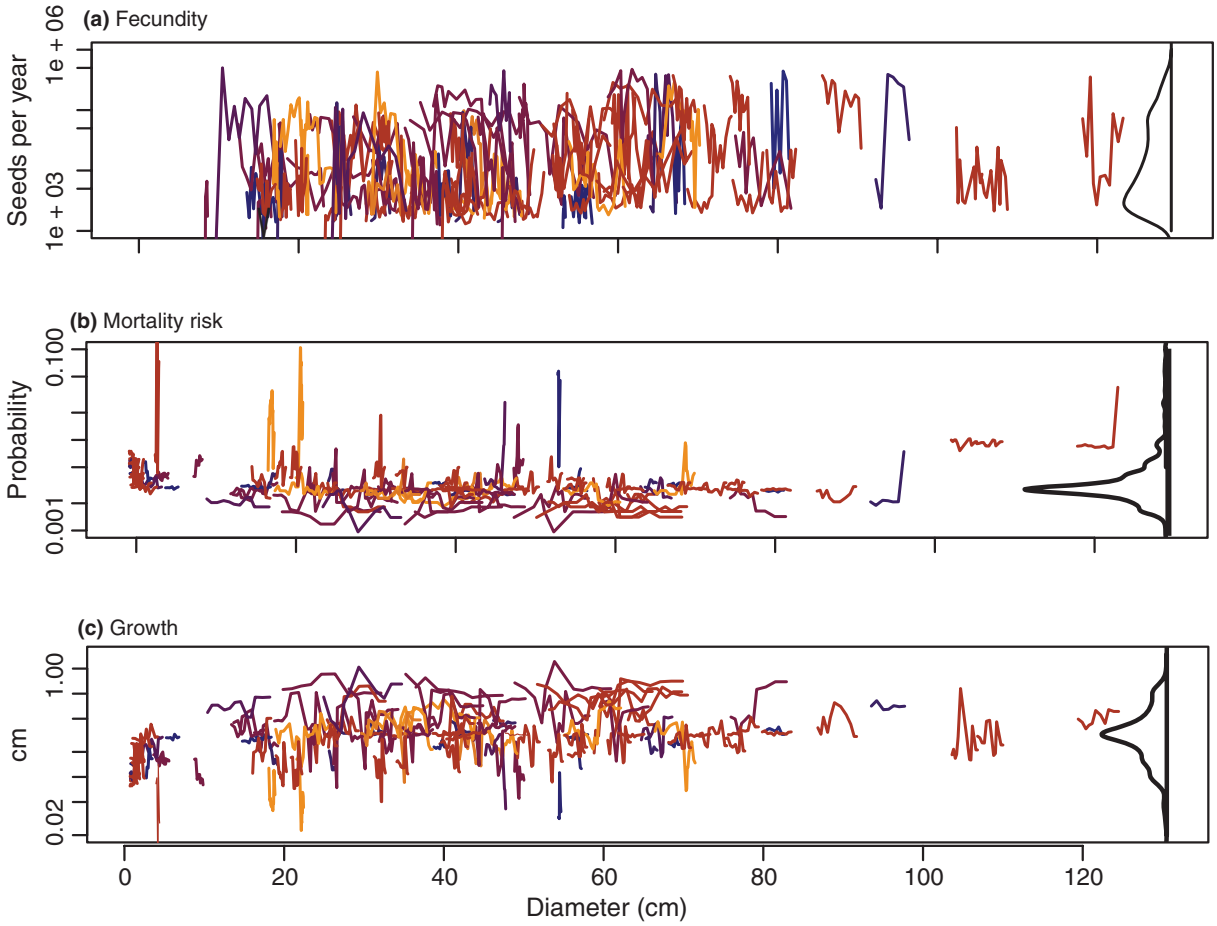

Figure 6 Posterior mean values for demographic rates of randomly sampled *Liriodendron* trees, each line tracking an individual over 7–19 years, with marginal densities at right. Colours show low elevation Piedmont (blue) to high elevation northern hardwoods (orange) plots. The full posterior distribution shown along the right margins includes these states for all individuals over all years (from the analysis of Clark *et al.* 2010).

understory do not show the largest growth responses for either species. Although *Morus* grows more slowly on average than *Liriodendron* in both settings, it shows a stronger response to gaps when previously growing slowly in the understory (Fig. 5b). The negative relationship in Fig. 5c does not in itself explain coexistence, but demonstrates that there are ways to partition gaps that cannot be inferred from the aggregate mean values. The joint density from eqn 8 shows the covariance structure contributed by V_s, much of which comes from unobserved x and is different for each individual.

When a joint distribution of x is unavailable, there is often the option to disaggregate by individual and time, i.e. not the joint distribution we would prefer, but one that still provides insight. In the foregoing example, inputs (gap vs. understory) were part of the experimental design. Even where no inputs are observed, variation between and within individuals (over time) can provide insight, despite not being of primary interest, as discussed in the next section.

Option 2: Disaggregate how you can

Individuals can reveal species differences when variables responsible for those differences are unobserved, exposing patterns in one dimension (individual organisms) that assist inference in another (competition for resources). Growth responses in Fig. 4b are conditioned on mean annual light and CO_2 but marginalise over variation in moisture, temperature, nutrients, and infection statuses, to name a few. When there are unobserved variables affecting dynamics of individuals, disaggregation by individual and over time recovers information. The joint distribution is high dimensional, including three demographic states for every individual, every year (Fig. 6). Despite its complexity, this joint distribution does not substitute for knowledge of all relevant variables $x_{i,t}$. For example, Fig. 5b tells us about the joint distributions of individuals with respect to understory and gap, but it does not tell us how variables such as light and moisture regulate responses in understory and gap. We prefer to disaggregate in terms of x, i.e. 'define the niche' in terms of such things as resources, but most of the important variables will not be known. When that option is not available, disaggregating instead by individual, state and time provides insight. Individuals are not the objects of interest, but they have the right scale – growth and reproduction of individuals as the environment varies in time.

The joint distribution of individuals, states and time preserves dependence structure that results from unobserved $x_{i,t}$ (eqns 4b, 6 and 7). Let $y_{i,t} = (g_{i,t}, f_{i,t}, s_{i,t})$ be a vector of states (growth, fecundity, mortality risk) and $\{y_{i,t}\}$ the set of response vectors. Marginal distributions describe aggregate variation in x encountered by all n individuals over all T years,

$$p(\{y_{i,t}\}|L, C, \boldsymbol{\Omega}) = \frac{1}{n}\int_{t}^{t+T} p(\{y_{i,t}\}|\{L_{i,t'}, C_{i,t'}, x_{i,t'}\}) \; p_{\boldsymbol{\Omega}}(\{x_{i,t'}\})d\{x_{i,t'}\} \quad (9)$$

(right margins of Fig. 6). The interpretation of overlap in Fig. 3 should not be that species are the same, but rather that there is essentially no information. We cannot disaggregate by x (it is unknown), but we can vastly reduce its range $\boldsymbol{\Omega}$, disaggregating instead by individual and year.

The joint distribution of individuals, states and years shows how each individual responded to the restricted variation it experienced each year

$$p\left(\{y_{i,t}\}|\{L_{i,t}, C_{i,t}, x_{i,t}\}\right)$$

We are now marginalising only the variation in x that one individual experienced in 1 year, rather than all variations experienced by all individuals over all years. Marginalising over all individuals and years (right side of Fig. 6) leads to the conclusion that species do not differ (Mouillot *et al.* 2005; Condit *et al.* 2006). The standard interpretation of a regression through the aggregate observations used to produce Fig. 4b would be that growth is unresponsive to light and CO_2 and essentially the same for all species – both conclusions known to be wrong (Fig. 4a). As the environment varies in space and time, disaggregated observations can recover species differences that depend on the joint response in many unmeasured dimensions. They can be examined for correlation structure (Clark 2010). The differences evident in correlation structure can be fitted to variables in x, such as climate, light availability and soil moisture (Clark *et al.* 2011).

Option 3: Model the Aggregate

When data are more aggregated than the processes of interest, there can be a third option, to marginalise the model itself. This means that we integrate out (eqn 3a) or sum away (eqn 3b) sources of variation that have contributed variation to observations. Models can be constructed to provide inference on conditional relationships, despite aggregated data. This option is not a substitute for disaggregated data, but an acknowledgement that aggregation has occurred. In the first example of this section, we point out how the marginalisation that is needed to model aggregated data can be overlooked. In the second example, we show how marginalising the model provides detailed inference, because it properly represents the aggregated data, and information may enter in multiple ways.

Aggregate data, disaggregated model

Ecologists use stage distributions to evaluate demography, where classes are discrete (matrix models, Fig. 7a) or continuous (integrated projection models or IPMs). The models are written at the population scale, in terms of density for size classes, but they are unconstrained by population size, fitted instead at a different scale, individual survival and growth (Ghosh *et al.*, unpublished data). Where fitted at the individual scale, they are models of individuals, not population growth. At the individual scale, parameters can be fitted independently of one another, sacrificing knowledge of their relationships. Recent studies suggest that parameters can be fitted together without individual level data, using Bayesian approaches (Gross *et al.* 2002) or optimisation (Wielgus *et al.* 2008). Problems identifying parameters are attributed to the ways in

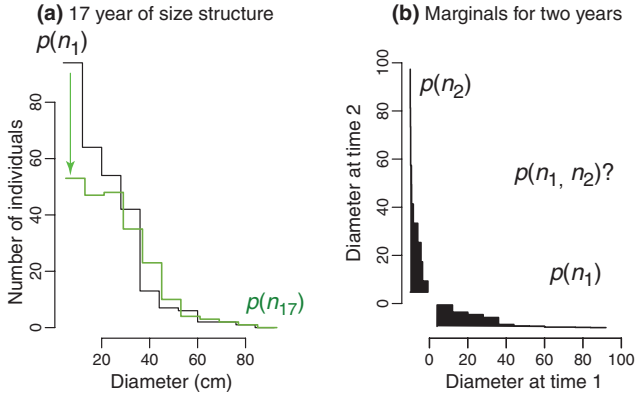

Figure 7 (a) Diameter classes of *Betula alleghaniensis* from 1992 [$p(n_1)$ in black] and 2008 [$p(n_{17})$ in green], showing the shift to larger diameter trees over time. (b) Two marginal distributions for years 1 and 2; the joint distribution $p(n_1, n_2)$ cannot be reconstructed from these aggregate data.

which parameters combine in models that project a distribution of stages at time n_t to time n_{t+1}. However, the more important issue has not been recognised, the fact that the joint distribution of (n_t, n_{t+1}) is unknown (Fig. 7b), having been lost when individuals were aggregated at time t and $t + 1$.

Consider a matrix model of m stages,

$$n_{t+1} = An_t$$
$$n_{1,t+1} = \sum_{i=1}^{m} n_{i,t}f_i \quad i = 1, \ldots, m, \quad t = 1, \ldots, T$$

with $m \times m$ matrix A containing coefficients a_{ij}, which describe transition from stage j to stage i, a vector of abundances at time t, $n_t = (n_{1,t}, \ldots, n_{m,t})$, and f_i, the birth rate for stage i. Elements of A are typically estimated independently.

A synthetic model could be used to estimate uncertainty in growth (e.g. Gross 2002; Clark 2003), but there is a problem when observations do not come from individuals. Consider the sequence of marginal stage distributions over time n_1, n_2, … . In this example, survivors from stage 1 at time t move to stage 2 at time $t + 1$, and those in stage 2 at time $t + 1$ come either from stage 1 at time t (grow and survive) or stage 2 at time t (already in stage 2, survive). If u individuals come from stage 1 and v from stage 2, then $n_{2,t+1} = u + v$. Based on observations at t and $t + 1$, we do not know how many came from stage 1 or stage 2, only their sum, the marginal

$$p\left(n_{2,t+1} = S_{uv}\right) = p(u + v) = \sum_{u} p_v(v|u)p_u(u) = \sum_{u} p_v(S_{uv} - u)p_u(u) \quad (10)$$

This is an example of eqn 3b. The summation is a convolution over the ways in which u and v could combine to produce the number of individuals in stage 2 at $t + 1$.

Complexity increases with the number of stages. For three stages that could transition to stage 2, we have

$$p(n_{2,t+1} = S_{uvw}) = p(u + v + w) = \sum_{v} p_w(S_{uvw} - v) \sum_{u} p_v(v - u)p_u(u)$$

So the problem is exploding, but it is still worse than this twofold convolution, because the u individuals that move from stage 1 to 2 could not remain in stage 1. In other words, the stages are interdependent in more complex ways. Even where applied properly,

the convolution does not substitute for the missing joint distribution, it only acknowledges the aggregation. In other words, even when the model is properly marginalised to accommodate data aggregation, there is less information than would be available from the joint distribution of individuals (Lavine *et al.* 2002).

Gross *et al.* (2002) make the important point that parameters in matrix A are hard to identify. For example, transitions from stage 1 could involve two parameters,

$$a_{11} = (1 - g_1)s_1$$
$$a_{21} = g_1 s_1$$

where g_1 is the probability of growing out of stage 1, and s_1 is probability of survival. Parameters appear as products and thus can be difficult to identify when fitted to aggregate data. Bayesian methods will not save the situation, unless the prior is allowed to control everything.

The more important problem for parameter estimation is the fact that the aggregation in observations is overlooked. Disaggregated data from individuals that are followed over time allow simultaneous inference on parameters because there is no convolution over the possible ways in which many individuals could combine to yield an aggregate distribution of stages. This longitudinal treatment of individuals is standard practice in public health and increasingly in ecological studies (e.g. capture–recapture methods). Knowledge of which individuals moved between stages provides the joint distribution $p(n_1, n_2, \ldots)$ (Fig. 1a). The problem with fitting matrix models to marginal distributions $p(n_1), p(n_2), \ldots$ comes from the fact that a model with proper aggregation is complex, and information is lost when moving from a joint distribution of individuals to the marginal distributions of classes. Ghosh *et al.* (unpublished data) address the more general aggregation issue for IPMs, the fact that they are fitted as models for individuals, but they are projected forward as though they were fitted to population growth data.

A marginalised model to accommodate aggregated data

Models based on conditional relationships often can be marginalised to match the aggregation in data and used for productive inference. For example, hypothesised negative density dependence occurs if competitors suffer when locally abundant, benefit when rare, or both. This self-regulation can be caused by natural enemies. Studies often look for a decline in survival with increasing abundance of conspecifics (e.g. Stoll & Newbery 2005; Petermann *et al.* 2008; Comita & Hubbell 2009; Gonzalez *et al.* 2010). As field studies typically do not benefit from knowledge of the factors responsible, observations of survival marginalise over unmeasured variables. For example, tree seedlings of many species benefit from soil moisture, as do the damping-off pathogens that attack them (Martin & Loper 1999). Understanding the role of potential pathogens requires conditional relationships, such as pathogen incidence and infection impact on survival, given soil moisture. Densities of survival probability aggregated over wet and dry sites broadly overlap (Fig. 8a), hiding the conditional relationships where self-regulation could be evident. A model can begin with the conditional relationships of interest and then marginalise to match the aggregation in data, thus providing inference on those conditional relationships.

To illustrate how a model can specify and fit conditional relationships where data represent marginal quantities, consider a model for pathogen incidence P (typically known only if hosts are

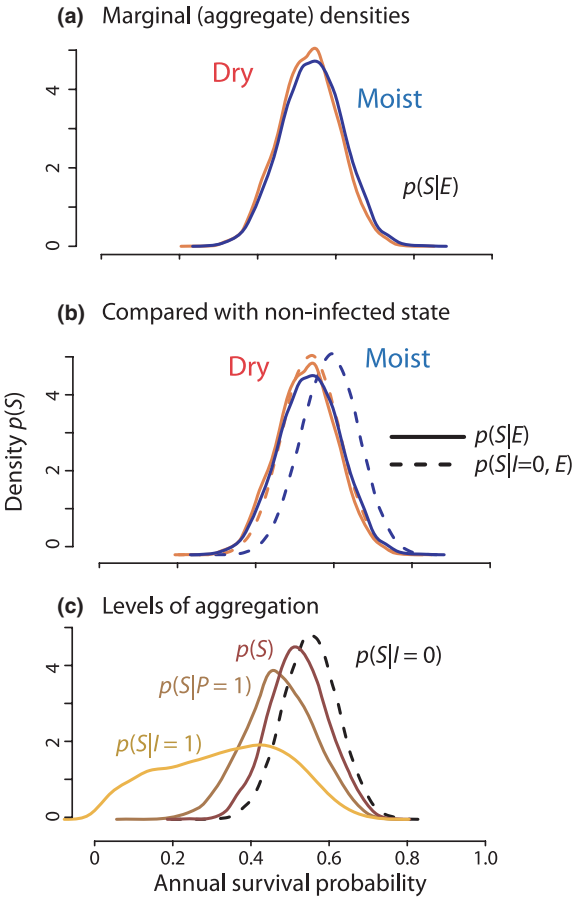

Figure 8 (a) Survival probability $p(S)$ of *Diospyros virginiana* seedlings planted in sites with high and low soil moisture. Densities marginalise over infection and incidence of potential fungal pathogens. (b) The importance of infection in moist sites is apparent from densities conditional on infection status (dashed lines), but hidden in marginals. (c) The large impact of infection status becomes apparent in conditionals [$p(S|I = 1)$ vs. $p(S|I = 0)$], and further depends on knowledge of incidence, as when there is no information on the fungus, $p(S)$, or information is limited to sites where the fungus is present, $P = 1$, $p(S|P = 1)$ (from analysis of Clark & Hersh 2009).

obviously infected), infection of seedling hosts I (often unobserved) and the survival S of infected ($S|I = 1$) and uninfected ($S|I = 0$) hosts (Clark & Hersh 2009; Hersh *et al.* 2011). The environment E regulates incidence and the effects of infection on host survival. The joint distribution of events is factored,

$$p(S, I, P|E) = p(S|I, E)p(I|P)p(P|E) \quad (11)$$

including how infection and environment affect host survival, infection if the pathogen is present, and how pathogen incidence depends on the environment respectively. Unfortunately, data represent aggregate quantities. If infection is detected, we have the most informative conditional density for survival given infection

$$p(S|I = 1, E) \quad (12)$$

If infection is not detected, but the pathogen is present (known from infection of neighbours), we have the aggregate quantity

$$p(S|P = 1, E) = \sum_{I=0,1} p(S|I, E)p(I|P = 1) \quad (13)$$

Equation 13 is less informative than 12, because infection status is unknown. If pathogen incidence is also unknown (e.g. there are no neighbours or their infection status is unknown), information is degraded further

$$P(S|E) = \sum_{P=0,1} \sum_{I=0,1} p(S|I,E)p(I|P)p(P|E) \quad (14)$$

Not shown here is a detection model for infection. Even with this knowledge, we still could not identify infection probability and incidence using eqn 14 alone. Equation 14 represents a model that properly marginalises over the unknown incidence and infection status. Despite being 'correct' (with respect to the degree of aggregation in data), it does not help us infer the important conditional relationships. However, each individual confers a different level of information, represented by eqns 12, 13, or 14.

Together with informative priors on detection, the models for different levels of aggregation (eqns 12–14) provide the conditional relationships of interest (Clark & Hersh 2009). Host survival is reduced by fungal infection under moist conditions, apparent from comparison with the conditional distribution $p(S|I = 0, E)$ (dashed lines in Fig. 8b), but hidden in the marginalisation that gives $p(S|E)$ (solid lines in Fig. 8b). Both host and fungus benefit from moisture. For hosts on moist sites, the negative effect of infection may outweigh the direct moisture benefit. Dashed lines in Fig. 8b show the survival difference for infection-free hosts. Solid lines show no survival difference when we aggregate over unknown infection status. The effects of different levels of aggregation corresponding to eqns 12–14 in Fig. 8c show that information in field studies can vary dramatically, depending on the degree of aggregation recognised.

These examples show how models are marginalised to match the aggregation in data and demonstrate that they can provide valuable inference. Even some of the most familiar types of studies (e.g. matrix models) can overlook the underlying aggregation. Proper marginalisation adds complexity to models, and there is a limit to how much can be learnt. On the other hand, information may enter at many stages of a hierarchical model, thus allowing for inference when information is sufficient.

There may be no options

The foregoing approaches cannot resolve all aggregation problems. In many cases, ecological data contain little or no information at the critical scale. For example, extensive plot arrays established for periodic inventory (e.g. Fig. 2) contain limited information on the effects of climate on annual demographic rates of trees, the scale assumed in many forest stand simulators. It generally will not be possible to disaggregate the multiyear demographic measurements to annual values or to disaggregate the coarse-resolution climate data to the scale where weather affects demography (individuals within local competitive environments). Models properly marginalised to match the aggregation in data will not provide inference at the required scale if there is no information in aggregated data. Regional climate data are relevant to climate experienced by an individual in the study only to the extent that regional climate is correlated with the weather that the individual experienced. The average demographic response over a 5-year interval may not reflect the extreme years during that interval that had the important effects on demography. There is a large and expanding literature on the need to properly match spatio-temporal reference in data and models (e.g. Banerjee *et al.* 2004; Gelfand *et al.* 2006), but the mismatches are easy to overlook in continental scale studies of climate change. This is an aggregation problem, involving data at one scale and processes at another.

DIVERSITY PREDICTION BASED ON INDEPENDENTLY MODELLED SPECIES

Niche models are used to predict diversity based on calibration with environmental variables (Kirilenko & Solomon 1998; Thomas *et al.* 2004; Thuiller *et al.* 2005; Prasad *et al.* 2007). Species are abstracted from communities, fitted to climate variables, and then reassembled as a biodiversity prediction. Niche modelling treats conditional distributions for individual species as though they were the joint distribution for a community. Despite widespread recognition that species depend on one another, there does not appear to be an articulation of why modelling species independently could be problematic and to what degree modelling species jointly could help. Here we are referring to the species that are modelled as response variables. In some cases, some species are treated as predictors of other species (e.g. Araújo & Luoto 2007; Barbet-Massin & Jiguet 2011). The discussion that follows concerns models for multiple species responding to physical and biotic variables.

There are at least two considerations with abstracting data for individual species followed by calibrating niche variables and biodiversity prediction. The more obvious issue concerns the unknown contribution of competition to current distributions (Hutchinson 1961; Pearson & Dawson 2003; Ibanez *et al.* 2006; Suttle *et al.* 2007; Clark *et al.* 2011). Data marginalise over competition and all environmental variables not included in the model. In the absence of information on competition, there is no obvious way to gauge its impact on predictions. For example, how would competitors expand following loss of chestnut in eastern North America (e.g. McCormick & Platt 1980; Elliot & Swank 2008)?

A second less obvious issue concerns the fact that in the absence of detailed environmental information, species provide the most information about one another. As species are responding to many of the same hidden variables, the best predictor of abundance can rely on knowledge of other species. Consider a community of species $s = 1, \ldots, S$ at locations $j = 1, \ldots, J$. At a given location j, there is dependence between species, because they compete (negative dependence), and they react similarly to the environment (positive dependence). As the joint distribution of environment/competition effects is unavailable, the effect of competition cannot be disaggregated. A comparison of two models that differ in how they treat species dependence illustrates aggregation problems in both.

For transparency, consider modelling species with a GLM with extra-Poisson (Gaussian) variation in the log link function,

$$\prod_{s=1}^{S}\prod_{j=1}^{J} Pois\left(a_{sj}|A_j\lambda_{sj}\right) \prod_{s=1}^{S}\prod_{j=1}^{J} N\left(\ln\lambda_{sj}|x_j\beta_s, \sigma_s^2\right) \quad (15)$$

a_{sj} is the number of individuals counted in a sample with effort (e.g. plot area) A_j and λ_{sj} is the mean number of individuals per unit area, or intensity, for species s in plot j. Input variables occupy x_j, a length-k vector of predictors, e.g. climate. Standard practice is to fit each

species, construct predictive distributions and combine the maps that result. A predictive distribution entails a scenario for x (e.g. a predicted climate), propagation of error in parameters and integration of the first-stage Gaussian and second-stage Poisson variation (Clark 2007). There will be a parameter set (β, σ^2) for each species. Predictions are independent. Each fitted model is controlled by the hidden relationships (e.g. eqn 6) – the fit for any one species is implicitly conditional on all others.

If instead we analyse the joint distribution of species, the first part of eqn 15 remains the same, but the second part is now

$$\prod_{j=1}^{J} N\left(\ln \lambda_j | x_j \beta, \Sigma_s\right) \quad (16)$$

The length-S vector of log intensities λ_j at location j is taken to be multivariate normal. The vectors β_s are gathered into the k by S matrix β. The only difference between eqns 15 and 16 is the fact that we are allowing that species are related not only through the variables that can be measured (x_j), but also those that cannot (Σ_s). There is nothing in either approach that entails how species interact, just their direct relationships to predictors and remaining covariances.

The two approaches are illustrated where species vary with soil moisture and elevation, summarised by two principle components (Fig. 9a). Priors on β, σ^2, and Σ are weak. The posterior distribution of these parameters and the imputed $\{\lambda_{ij}\}$ were simulated with Metropolis-within-Gibbs sampling. Predictive means for the two approaches look the same (Fig. 9c), but residual errors differ (Fig. 9b, see below). However, a change in the presence of one species causes a large change in predictions from the multivariate model, but no change in independently fitted models (Fig. 9d). And neither is 'correct'. The univariate models are unrealistic because the removal of a competitor should allow expansion of the species that occur with it. Instead, the univariate models simply leave the space empty, having no way to correct for the missing species. The multivariate model errs in a different way – upon removal of a species, it predicts a decline in any species that is positively correlated with it, and vice versa (Fig. 9c). This is the opposite of what should occur: species abundant on the same sites will benefit most, not least, by removal of a competitor.

The advantage of modelling species together comes from the fact that prediction benefits not only from information in x but also from information in Σ_s. There is low residual error in the multivariate model (Fig. 9b) and small prediction error because much of the information comes from other species. However, the case for fitting species together goes no further than this. No additional accommodation for species interactions comes from the multivariate model because there is no opportunity to disaggregate competition and environment. Both models depend on the assumption that the same species are present for calibration and prediction. Both suffer from the fact that the data cannot be disaggregated into environment and competition. Once again, aggregation is the challenge.

DISCUSSION

Debates about niche modelling, contributions of individual variation to biodiversity, identification of species differences and inference for demographic models owe much to a commonly overlooked problem. The problem is not a failure to recognise that processes operate at different scales, the subject of a large 'scaling' literature. The confusion results instead from aggregating over the relevant scale to draw conclusions from models and data informed by a different scale. Species do not interact or respond to climate change, but individuals do. Data collection implicitly conditions on the setting from which observations were obtained; the inevitable marginalisation over sample space and time degrades information. Analysis degrades information further, as observations from many organisms and plots are aggregated to produce species-level averages. Species differences are hidden and qualitatively changed (Fig. 5). The relationships can be recovered if there is opportunity to disaggregate.

Recent emphasis on synthesis has motivated many studies that treat species as observations, including comparisons of trait values from the literature. In some cases, aggregation is desirable. In other cases, treating species as observations is the only option, and it can provide

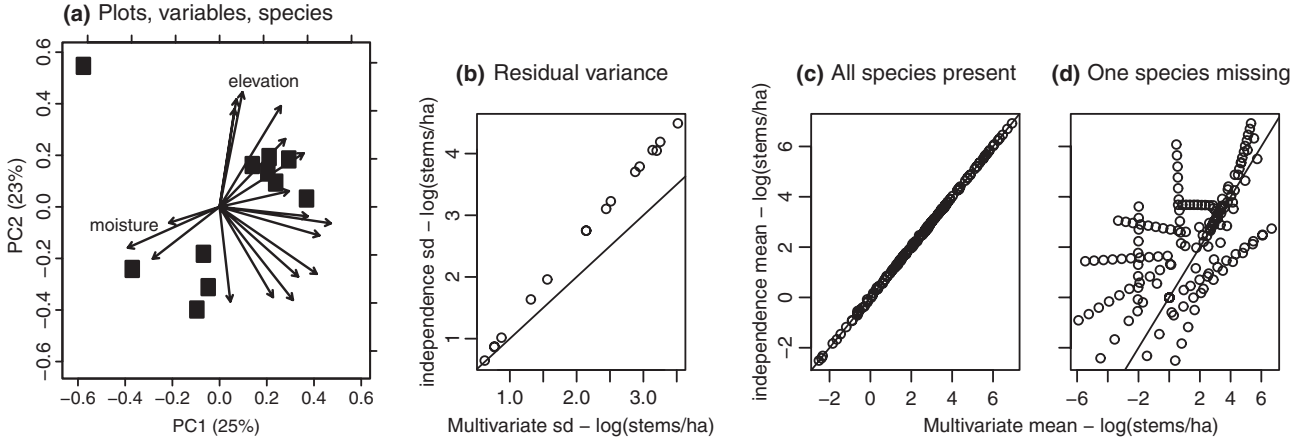

Figure 9 (a) PCA summary for 11 plots (black letters), 16 dominant tree species and two environmental variables (elevation, moisture index) for stands analysed by Clark *et al.* (2010). (b–d) Each dot is one species, plotted as estimates from models fitted independently for each species (vertical axis) and together (horizontal axis). In (b) residual standard deviations for independently fitted models (σ_s in eqn 15) are larger than for a single multivariate model [diag(Σ_s) in eqn 16]. (c) Predictions of mean abundance (λ_{sj}) across an elevation gradient are essentially the same. (d) When a species is removed (*Liriodendron tulipifera*), the multivariate model (eqn 16) predicts large decline in species that tend to occur with *Liriodendron* (left of 1 : 1 line) and large increase in species that tend not to occur with *Liriodendron*. Neither correctly predicts that species that occur with *Liriodendron* would benefit most.

useful insight (Reich *et al.* 1999; Wright *et al.* 2004; Westoby & Wright 2006). Failure to recognise aggregation, not aggregation itself, is the source of problems. The lost of information is critical when the processes of interest operate at the disaggregated level.

The three options we summarise can be valuable even before data collection begins. In many cases, data are collected and 'pre-marginalised' or they are collected in a way that misses opportunity to learn from the joint distribution. For example, a study of specific leaf area and leaf nitrogen along a gradient might not come from the same leaves or even from the same individuals. Samples obtained from random individuals are often aggregated by species and plot. Each species has a value for each variable at each location, but there is no joint distribution. Species projections on two dimensions are misleading (Fig. 3) and inadequate for inference on processes that operate at the individual scale (Figs 5 and 8).

In many cases, the solution can be simple. By obtaining samples from the same individuals, one can construct the joint distribution of traits (Figs 1 and 5). Where disaggregation is not possible, or it cannot be done along the dimensions of interest, we suggest disaggregation in other dimensions that could increase information. Disaggregation from species to individuals and years reveals species differences that are not apparent from species-level data (Clark 2010). A third option of marginalising the model essentially degrades the model to admit aggregated data and sometimes recover conditional relationships (Fig. 8).

Aggregation plays a role in predicting diversity based on independent analysis of individual species. Invasion by exotic species depends on those already present (Stohlgren *et al.* 2001; Drake & Lodge 2006; Pathikonda *et al.* 2008). The aggregation in distribution data frustrates interpretation regardless of whether species are analysed independently or jointly (Fig. 9).

The species-level differences that explain coexistence and environmental response are evident in individual-level variation, but unobservable when using aggregated data or models. Species differences long known from physiological studies affect short-term uptake of moisture and nutrients, responses to sunflecks, phenological differences, capacity to withstand short-term and long-term drought and responses to pathogens and herbivores. Growth, survival and fecundity are a direct consequence of individual health, and fitness is a direct consequence of demography. The fact that individuals respond to variation more like others of the same species concentrates competition within the species, thereby promoting coexistence (Clark 2010). The evidence is widespread in field data. In our studies, record white oak fecundity in 2009 resulted in intense intraspecific seedling competition. Late frost in 1997 led to loss of red maple recruitment at mid elevations, but not other species. Growth and fecundity of several elm species closely track summer drought, whereas coexisting persimmon, sweet gum and red maple do not (Clark *et al.* 2011). These are a few of many examples of responses that amplify self-regulation in ways that are not accommodated by models used to analyse species differences and to explain coexistence. They do not enter resource-competition models or metapopulation models, including those viewed as including individual variation.

Equations 7 and 8 have the heuristic value of showing how individual variation can reveal species differences when information is limited. Unobserved variables contribute to covariance structure, rather than mean structure. Even without genetic variation, individuals differ due to the distribution of unobserved influences that enter through V_s. The translation of this variation by the individual enters through β_s, different for each species. Species differ in terms of their distributions of individuals, and this individual scale is critical for understanding biodiversity.

ACKNOWLEDGEMENTS

This work was supported by grants from NSF (DEB-0955904, CDI 0940671, DDDAS 0540347, LTER), the Department of Energy, and the US Forest Service. For comments on the manuscript, we thank three anonymous referees and Marcel Holyoak.

REFERENCES

Ackerly, D.D. & Cornwell, W.K. (2007). A trait-based approach to community assembly: partitioning of species trait values into within- and among-community components. *Ecol. Lett.*, 10, 135–145.

Agrawal, A.A., Ackerly, D.D., Adler, F., Arnold, A.E., Cáceres, C., Doak, D.F. *et al.* (2007). Filling key gaps in population and community ecology. *Front. Ecol. Environ.*, 5, 145–152.

Araújo, M.B. & Luoto, M. (2007). The importance of biotic interactions for modelling species distributions under climate change. *Glob. Ecol. Biogeogr.*, 16, 743–753.

Banerjee, S., Carlin, B.P. & Gelfand, A.E. (2004). *Hierarchical Modeling and Analysis for Spatial Data*. Chapman and Hall, Boca Raton, FL.

Baraloto, C., Goldberg, D.E. & Bonal, D. (2005). Performance trade-offs among tropical tree seedlings in contrasting microhabitats. *Ecology*, 86, 2461–2472.

Barbet-Massin, M. & Jiguet, F. (2011). Back from a predicted climatic extinction of an Island endemic: a future for the Corsican Nuthatch. *PLoS ONE*, 6, e18228, doi: 10.1371/journal.pone.0018228.

Bazzaz, F. (1979). The physiological ecology of plant succession. *Annu. Rev. Ecol. Syst.*, 10, 351–371.

Berec, L. (2002). Techniques of spatially explicit individual-based models: construction, simulation, and mean-field analysis. *Ecol. Modell.*, 150, 55–81.

Bickel, P.J., Hammel, E.A. & O'Connell, J.W. (1975). Sex bias in graduate admissions: data from Berkeley. *Science*, 187, 398–404.

Bolnick, D.I., Svanbäck, R., Fordyce, J.A., Yang, L.H., Davis, J.M. & Hulsey, C.D. *et al.* (2003). The ecology of individual: incidence and implications of individual specialization. *Am. Nat.*, 161, 1–28.

Brown, J.H. & Maurer, B.A. (1989). Macroecology: the division of food and space among species on continents. *Science*, 243, 1145–1150.

Buckley, L.B., Urban, M.C., Angilletta, M.J., Crozier, L.G., Rissler, L.J. & Sears, M.W. (2010). Can mechanism inform species' distribution models? *Ecol. Lett.*, 13, 1041–1054.

Canham, C.D. & Thomas, R.Q. (2010). Frequency, not relative abundance, of temperate tree species varies along climate gradients in eastern North America. *Ecology*, 91, 3433–3440.

Clark, J.S. (2003). Uncertainty in ecological inference and forecasting. (Special Feature). *Ecology*, 84, 1349–1350.

Clark, J.S. (2007). *Ecological data models with R*. Princeton University Press, Princeton, NJ, USA.

Clark, J.S. (2009). Beyond neutral science. *Trends Ecol. Evol.*, 24, 8–15.

Clark, J.S. (2010). Individuals and the variation needed for high species diversity. *Science*, 327, 1129–1132.

Clark, J.S. (2011). The coherence problem with the Unified Neutral Theory of Biodiversity. *Trends in Ecology and Evolution*, in press.

Clark, J.S. & Hersh, M.H. (2009). Inference when multiple pathogens affect multiple hosts: Bayesian model selection. *Bayesian Anal.*, 4, 337–366, doi: 10.1214/09-BA413.

Clark, J.S., Mohan, J., Dietze, M. & Ibanez, I. (2003). Coexistence: how to identify trophic tradeoffs. *Ecology*, 84, 17–31.

Clark, J.S., Bell, D. *et al.* (2010). High dimensional coexistence based on individual variation: a synthesis of evidence. *Ecol. Monogr.*, 80, 569–608.

Clark, J.S., Bell, D.M., Hersh, M.H. & Nichols, L. (2011). Climate change vulnerability of forest biodiversity: climate and resource tracking of demographic rates. *Glob. Change Biol.*, 17, 1834–1849.

Clark, J.S., LaDeau, S. & Ibanez, I. (2004). Fecundity of trees and the colonization-competition hypothesis. *Ecological Monographs*, 74, 415–442.

Clark, J.S., Silman, M., Kern, R., Macklin, E. & Hille Ris Lambers, J. (1999). Seed dispersal near and far: generalized patterns across temperate and tropical forests. *Ecology*, 80, 1475–1494.

Comita, L.S. & Hubbell, S.P. (2009). Local neighborhood and species' shade tolerance influence survival in a diverse seedling bank. *Ecology*, 90, 328–334.

Condit, R., Ashton, P.S., Bunyavejchewin, S., Dattaraja, H.S., Davies, S., Esufali, S. *et al.* (2006). The importance of demographic niches to tree diversity. *Science*, 313, 98–101.

Courbaud, B., Vieilledent, G. *et al.* (2011). Intra-specific variability and the competition–colonisation trade-off: coexistence, abundance and stability patterns. *Theor. Ecol.*, in press.

DeAngelis, D. & Mooij, W.M. (2005). Individual-based modelling of ecological and evolutionary processes. *Annu. Rev. Ecol. Evol. Syst.*, 3, 147–168.

DeLucia, E.H. & Thomas, R.B. (2000). Photosynthetic responses to CO_2 enrichment of four hardwood species in a forest understory. *Oecologia*, 122, 11–19.

Drake, M.J. & Lodge, D.M. (2006). Allee effects, propagule pressure and the probability of establishment: risk analysis for biological invasions. *Biol. Invasions*, 8, 365–375.

Easterling, M.R., Ellner, S.P. *et al.* (2000). Size-specific sensitivity: applying a new structured population model. *Ecology*, 81, 694–708.

Elliot, K.J. & Swank, W.T. (2008). Long-term changes in forest composition and diversity following early logging (1919-1923) and the decline of American chestnut (*Castanea dentata*). *Plant Ecol.*, 197, 155–172.

Englund, G. & Leonardsson, K. (2008). Scaling up the functional response for spatially heterogeneous systems. *Ecol. Lett.*, 11, 440–449.

Flyvbjerg, H., Sneppen, K. *et al.* (1993). Mean field theory for a simple model of evolution. *Phys. Rev. Lett.*, 71, 4087–4090.

Gelfand, A.E., Silander, J.A. Jr, Wu, S., Latimer, A.M., Lewis, P., Anthony Rebelo, G. *et al.* (2006). Explaining species distribution patterns through hierarchical modeling. *Bayesian Anal.*, 1, 41–92.

Gelman, A. & Speed, T.P. (1993). Characterizing a joint probability distribution by conditionals. *J. R. Stat. Soc. B*, 55, 185–188.

Gonzalez, M.A., Roger, A. *et al.* (2010). Shifts in species and phylogenetic diversity between sapling and tree communities indicate negative density dependence in a lowland rain forest. *J. Ecol.*, 98, 137–146.

Gross, G., Craig, B.A. *et al.* (2002). Bayesian estimation of a demographic matrix model from stage-frequency data. *Ecology*, 83, 3285–3298.

Gross, K., Craig, B.A. & Hutchison, W.D. (2002). Bayesian estimation of a demographic matrix model from stage-frequency data. *Ecology*, 83, 3285–3298.

Guisan, A. & Thuiller, W. (2005). Predicting species distribution: offering more than simple habitat models. *Ecol. Lett.*, 8, 993–1009.

Hersh, M., Clark, J.S. & Vilgalys, R. (2011). Evaluating the impacts of fungal seedling pathogens on temperate forest seedling survival. *Ecology*, in press.

Hutchinson, G.E. (1961). The paradox of the plankton. *Am. Nat.*, 95, 137–145.

Ibanez, I., Clark, J.S. *et al.* (2006). Predicting biodiversity change: outside the climate envelope, beyond the species-area curve. *Ecology*, 87, 1896–1906.

Infante, J.M., Rambal, S. *et al.* (1997). Modeling transpiration in holm-oak savannah: scaling up from the leaf to the tree scale. *Agric. For. Meteorol.*, 87, 273–289.

King, G. (1997). *A Solution to the Ecological Inference Problem*. Princeton University Press, Princeton, NJ.

Kirilenko, A.P. & Solomon, A.M. (1998). Modeling dynamic vegetation response to rapid climate change using bioclimatic classification. *Clim. Change*, 38, 15–49.

Kitajima, K. (1994). Relative importance of photosynthetic traits and allocation patterns as correlates of seedling shade tolerance of 13 tropical trees. *Oecologia*, 98, 419–428.

LaDeau, S.L. & Clark, J.S. (2006). Elevated CO_2 and tree fecundity: the role of tree size, interannual variability, and population heterogeneity. *Glob. Change Biol.*, 12, 822–833.

Lavine, M., Beckage, B. *et al.* (2002). Statistical modelling of seedling mortality. *J. Agric. Biol. Environ. Stat.*, 7, 21–41.

Levin, S.A. (1992). The problem of pattern and scale in Ecology. *Ecology*, 73, 1943–1967.

Levins, R. (1979). Coexistence in a variable environment. *Am. Nat.*, 114, 765–783.

Levinsky, I., Skov, F. *et al.* (2007). Potential impacts of climate change on the distributions and diversity patterns of European mammals. *Biodivers. Conserv.*, 16, 3803–3816.

Lichstein, J.W., Dushoff, J. *et al.* (2007). Intraspecific variation and species coexistence. *Am. Nat.*, 170, 807–818.

MacArthur, R.H. (1972). *Geographical Ecology*. Princeton University Press, Princeton, NJ.

Martin, F.N. & Loper, J.E. (1999). Soilborne plant diseases caused by *Pythium* spp: ecology, epidemiology, and prospects for biological control. *Crit. Rev. Plant Sci.*, 18, 111–181.

McCormick, J.F. & Platt, R.B. (1980). Recovery of an Appalachian forest following the chestnut blight or Catherine Keever-you were right! *Am. Midl. Nat.*, 104, 264–273.

McGill, B.J. (2003). Does Mother Nature really prefer rare species or are log-left skewed SADs a sampling artifact? *Ecol. Lett.*, 6, 766–773.

Melbourne, B.A. & Chesson, P. (2005). Scaling up population dynamics: integrating theory and data. *Oecologia*, 145, 179–187.

Messier, J., McGill, B.J. *et al.* (2010). How do traits vary across ecological scales? A case for trait-based ecology. *Ecol. Lett.*, 13, 1–11.

Mohan, J.E., Clark, J.S. & Schlesinger, W.H. (2007). Long-term CO_2 enrichment of an intact forest ecosystem: implications for temperate forest regeneration and succession. *Ecol. Appl.*, 17, 1198–1212.

Mouillot, D., Stubbs, W. *et al.* (2005). Niche overlap estimates based on quantitative functional traits: a new family of non-parametric indices. *Oecologia*, 145, 345–353.

Nakashizuka, T. (2001). Species coexistence in temperate, mixed deciduous forests. *Trends Ecol. Evol.*, 16, 205–210.

Nee, S., Harvey, P.H. & May, R.M. (1991). Lifting the veil on abundance patterns. *Proc. R. Soc. B Biol. Sci.*, 243, 161–163.

Nelsen, R.B. (1999). *An Introduction to Copulas*. Springer, New York.

Novak, M., Wootton, J.T., Doak, D.F., Emmerson, M., Estes, J.A. & Tinker, M.T. (2011). Predicting community responses to perturbations in the face of imperfect knowledge and network complexity. *Ecology*, 92, 836–846.

Oishi, A.C., Oren, R. *et al.* (2010). Interannual invariability of forest evapotranspiration and its consequence to water flow downstream. *Ecosystems*, 13, 421–436.

Pathikonda, S.A.S., Ackleh, K.H. *et al.* (2008). Invasion, disturbance, and competition: modeling the fate of coastal plant populations. *Conserv. Biol.*, 23, 164–173.

Pearson, R.G. & Dawson, T.P. (2003). Predicting the impacts of climate change on the distribution of species: are bioclimate envelope models useful? *Glob. Ecol. Biogeogr.*, 12, 361–371.

Petermann, J.S., Fergus, A.J.F. *et al.* (2008). Janzen-Connell effects are widespread and strong enough to maintain diversity in grasslands. *Ecology*, 89, 2399–2406.

Peterson, A., Ortega-Huerta, M. *et al.* (2002). Future projections for Mexican faunas under global climate change scenarios. *Nature*, 416, 626–629.

Poorter, L., Kitajima, K. *et al.* (2010). Resprouting as a persistence strategy of tropical forest trees: relations with carbohydrate storage and shade tolerance. *Ecology*, 91, 2613–2627.

Prasad, A.M., Iverson, L.R., Matthews, S. & Peters, M. (2007). *A Climate Change Atlas for 134 Forest Tree Species of the Eastern United States [database]*. Northern Research Station, USDA Forest Service, Delaware, OH. Available at: http://www.nrs.fs.fed.us/atlas/tree Last accessed 15 September 2011.

Pyšek, P., Richardson, D.M. *et al.* (2008). Geographical and taxonomic biases in invasion ecology. *Trends Ecol. Evol.*, 23, 237–244.

Reich, P.B., Ellsworth, D.S. *et al.* (1999). Generality of leaf trait relationships: a test across six biomes. *Ecology*, 80, 1955–1969.

Ricklefs, R.E. & Jenkins, D.G. (2011). Biogeography and ecology: towards the integration of two disciplines. *Philos. Trans. R. Soc. Lond. Ser. B*, 366, 2438–2448.

Rozendaal, D.M.A., Hurtado, V.H. *et al.* (2006). Plasticity in leaf traits of 38 tropical tree species in response to light; relationships with light demand and adult stature. *Funct. Ecol.*, 20, 207–216.

Ruel, J.J. & Ayres, M.P. (1999). Jensen's inequality predicts effects of environmental variation. *Trends Ecol. Evol.*, 14, 361–366.

Running, S., Nemani, R. *et al.* (2004). A continuous satellite-derived measure of global terrestrial primary production. *Bioscience*, 54, 547–560.

Scheiner, S.M., Cox, S.B. *et al.* (2000). Species richness, species–area curves, and Simpson's paradox. *Evol. Ecol. Res.*, 2, 791–802.

Seiwa, K. (2007). Trade-offs between seedling growth and survival in deciduous broadleaved trees in a temperate forest. *Ann. Bot.*, 99, 537–544.

Shipley, B., Lechowicz, M.J., Wright, I. & Reich, P.B. (2006). Fundamental tradeoffs generating the worldwide leaf economics spectrum. *Ecology*, 87, 535–541.
Simpson, E.H. (1951). The interpretation of interaction in contingency tables. *J. R. Stat. Soc. Ser. B*, 13, 238–241.
Stohlgren, T.J., Otsuki, Y. *et al.* (2001). Patterns of plant invasions: a case example in native species. *Biol. Invasions*, 3, 37–50.
Stoll, P. & Newbery, D.M. (2005). Evidence of species-specific neighborhood effects in the dipterocarpaceae of a Bornean rain forest. *Ecology*, 86, 3048–3062.
Streng, D.R., Glitzenstein, J.S. *et al.* (1989). Woody seedling dynamics in an East Texas floodplain forest. *Ecol. Monogr.*, 59, 177–204.
Suttle, K.B. *et al.* (2007). Species interactions reverse grassland responses to changing climate. *Science*, 315, 640–642.
Thomas, C.D. *et al.* (2004). Extinction risk from climate change. *Nature*, 427, 145–148.
Thuiller, W., Lavorel, S. *et al.* (2005). Climate change threats to plant diversity in Europe. *Proc. Natl Acad. Sci. USA*, 102, 8245–8250.
Tilman, D. (1988). *Dynamics and Structure of Plant Communities. Monographs in Population Biology 26*. Princeton University Press, Princeton, NJ, USA.
Turnbull, M.H. (1991). The effect of light quantity and quality during development on the photosynthetic characteristics of six Australian rainforest tree species. *Oecologia*, 87, 110–117.
Underwood, N., Hambäck, P. *et al.* (2005). Large-scale questions and small-scale data: empirical and theoretical methods for scaling up in ecology. *Oecologia*, 145, 176–177.
Valladares, F. & Niinemets, Ü. (2008). Shade tolerance, a key plant feature of complex nature and consequences. *Ann. Rev. Ecol. Evol. Syst.*, 39, 237–257.
Wagner, C.H. (1982). Simpson's Paradox in real life. *Am. Stat.*, 36, 46–47.
Wakefield, J. & Salway, R. (2001). Analysis and interpretation of disease clusters and ecological studies. *J. R. Stat. Soc. Ser. A*, 164, 119–137.
Wakefield, J. & Shaddick, G. (2005). Health-exposure modeling and the ecological fallacy. *Biostatistics*, 1, 1–19.
Walters, M.B. & Reich, P.B. (1996). Are shade tolerance, survival, and growth linked? Low light and, nitrogen effects on hardwood seedlings. *Ecology*, 77, 841–853.
Warren, R.J. II, Skelly, D.K., Schmitz, O.J. & Bradford, M.A. (2011). Universal ecological patterns in college basketball communities. *PLoS ONE*, 6, e17342.
Weins, J.A. (1989). Spatial scaling in ecology. *Funct. Ecol.*, 3, 385–397.
Welden, C.W., Hewett, S.W. *et al.* (1991). Sapling survival, growth, and recruitment: relationship to canopy height in a neotropical forest. *Ecology*, 72, 35–50.
Westoby, M. & Wright, I.J. (2006). Land-plant ecology on the basis of functional traits. *Trends Ecol. Evol.*, 21, 261–268.
Wiegand, T., Gunatilleke, C.V.S. *et al.* (2007). How individual species structure diversity in tropical forests. *Proc. Natl Acad. Sci. USA*, 104, 19029.
Wielgus, J., M. Gonzalez-Suarez, D. *et al.* (2008). A noninvasive demographic assessment of sea lions based on stage-specific abundances. *Ecol. Appl.*, 18, 1287–1296.
Wright, S.J. (2002). Plant diversity in tropical forests: a review of mechanisms of species coexistence. *Oecologia*, 130, 1–14.
Wright, I.J., Reich, P.B. *et al.* (2004). The worldwide leaf economics spectrum. *Nature*, 428, 821–827.
Zhu, K., Woodall, C. & Clark, J.S. (2011). The climate migration lag in forest trees. *Global Change Biology*, in press.

Editor, Jerome Chave
Manuscript received 3 February 2011
First decision made 17 March 2011
Second decision made 19 June 2011
Third decision made 21 July 2011
Manuscript accepted 19 August 2011