Title

Macroecological Patterns Out Of Steady State

Author

Brush, Micah

Publication Date

2021

Peer reviewed|Thesis/dissertation

Macroecological Patterns Out Of Steady State

by

Micah Brush

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Physics

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor John Harte, Co-chair
Professor Oskar Hallatschek, Co-chair
Professor Rosemary Gillespie
Professor Hernan Garcia

Summer 2021

Macroecological Patterns Out Of Steady State

Abstract

Macroecological Patterns Out Of Steady State

by

Micah Brush

Doctor of Philosophy in Physics

University of California, Berkeley

Professor John Harte, Co-chair

Professor Oskar Hallatschek, Co-chair

Prevalent macroecological patterns have been identified across a wide range of ecosystems, and these patterns have proven effective for understanding ecosystems at scales relevant for conservation and management. However, empirical studies and macroecological theory to this point have largely focussed on static patterns for ecosystems in steady state, and there is increasing interest in understanding how these metrics change over time and in respond to disturbance.

In my dissertation, I use the Maximum Entropy Theory of Ecology (METE) as a starting point to predict how ecosystems will respond to disturbance and analyze the corresponding shifts in macroecological patterns. METE uses the principal of maximum entropy to predict various macroecological patterns and has proven effective for ecosystems at steady state, though its predictions appear to fail for disturbed ecosystems. The first chapter of my dissertation studies how deviations from METE predictions can inform us about underlying biology by studying macroecological patterns across land uses of different intensities for arthropods in the Azores. I then look at how we can modify METE to improve its predictions for ecosystems out of steady state. In my second chapter I present a new model that extends the spatial predictions of the theory to include intraspecific negative density dependence. Finally, in my third chapter I discuss my work developing DynaMETE: a new hybrid theory of macroecology that combines the maximum entropy methods of METE with explicit mechanisms to predict how patterns change in time. I present a method for iterating this theory in time, and code that implements this iteration scheme.

Ecosystems are faced with increasing levels of human disturbance from habitat fragmentation, to land management, to climate change. This makes it important to study macroecological patterns out of steady state as we work toward understanding how ecosystems will respond to disturbances at the large scales relevant for conservation and management.

To my parents, whose support throughout my life made this possible.

Contents

List of Figures

List of Tables

Acknowledgments

I want to thank all of my collaborators throughout that helped me learn and grow as a scientist. I could not have asked for a better advisor than John Harte, whose empathy and understanding enormously improved my graduate experience. I am inspired by his love of the natural world, and hope to bring the same curiousity and wonder to my future work. I would also like to thank the rest of my research group, in particular Kaito Umemura, who I have had many interesting and useful conversations with throughout my degree, including while hiking in Colorado together. I have also really enjoyed working with and getting to know the undergraduates who have been involved in the Harte lab: Juliette Franzman, Roya Safaeinili, Nina Groleger, and Julia Nicholson.

Thank you to the other research groups at Berkeley where I have been able to collaborate, particularly the labs of Rosemary Gillespie and Benjamin Blonder. Outside of Berkeley, I would like to thank Justin Kitzes for his support and collaboration, as well as collaborators Erica Newman, Paulo A.V. Borges, and Tom Matthews. Also, thank you to the Rocky Mountain Biological Laboratory and everyone I met there for a wonderful and memorable summer. Finally, while not represented in my dissertation, I spent nearly a year working in cosmology, which I very much enjoyed. I want to thank in particular Miguel Zumalacárregui for his friendship and helpful advice, as well as my advisor Zarija Lukić, and my collaborator Eric Linder.

I also want to acknowledge the various ways I was funded throughout my degree. In my first few years I worked as a graduate student instructor, and I want to thank the teaching teams that made that a very enjoyable experience. Working with the Sense & Sensibility & Science team in particular was an incredible experience, and I made many friends through that community. I also really enjoyed working with the Origins team, in another very fun Big Ideas course. I was also fortunate to be selected as a Berkeley Connect Fellow, and am very grateful for that program and the opportunity I had to develop my mentorship abilities. Finally, I would like to acknowledge the NSF, through my advisor, and NSERC, through my PGS D fellowship, which together allowed me to focus on research for the second half of my degree.

Within my department, I want to thank all of the physics staff and student support team, who made it easier to navigate the difficult Berkeley bureaucracy. I want to thank particularly Joelle Miles for her emotional and administrative support. I felt that I could always stop by her office to just say hi, to discuss department policies and ideas for improving them, or to figure out complicated paperwork.

I also want to acknowledge everyone who I interacted with through my professional service, including the Committee on Teaching, and the Graduate Assembly. This work enriched my graduate student experience and gave me a strong and important sense of community at Berkeley. Most notably, I want to thank everyone involved in Respect is Part of Research, and especially Wren Suess. It was inspiring to be around so many folks who care about making the department a better place to be. The workshop that Wren and I organized to share the RPR model with graduate students outside of Berkeley was one of

my favourite experiences in all of graduate school, and was only possible because working together was both enjoyable and productive.

On that note, I have made many close friends during my time at Berkeley. I unfortunately cannot list everyone I shared a coffee with, but I really do appreciate everyone that I have spent time with in the past 5 years. Whether commiserating over the challenges of graduate school or celebrating each others' achievements, I believe that the people I met and spent time with really made my experience here.

I will miss meeting Northside and grabbing lunch with Andreas Biekert, chatting politics over coffee with Vetri Velan, and weekly Bachelor nights with Shoshana Jarvis. I am also very grateful for the support and friendship of Emily Liquin, after we became close friends teaching SSS together. Our year as roommates was really impactful, and I really appreciated the support while I was finding an advisor and switching research fields.

A special mention to the Berkeley Chamber Chorus, which I sang with from my first semester until the pandemic made singing together impossible. The rehearsals gave structure to my week and were always something to look forward to, and concerts helped to demarcate the year and mark progress when research felt slow. I will not forget the experience of singing in the world premiere performance of Dreamers, and the audience response in Zellerbach. I will miss our weekly after rehearsal dinners.

I will enormously miss dinners at King Dong with Jonathan Liu, Matthew Quenneville, and Sam Badman. We have been friends since the first weeks of our time at Berkeley: Matthew and I as we are both SFU alum, Sam and I after discussing Oxbridge while waiting for onboarding, and Jonathan and I because we just kept running into each other at every department and campus event, including choir. Our online board game nights during the pandemic were also a bright spot in a tough year, even if we weren't able to make it to the next season of Pandemic Legacy. I will also miss Tuesday morning coffee at Victory Point with Sam and Matthew, which helped structure my weeks. And I will miss exploring every lunch spot possible near campus with Jonathan while discussing things from physics department policies, to the larger pros and cons of academia, to Berkeley city planning, to broader questions about what we want in life after graduating. I also could not have asked for a better roommate than Matthew, and I will miss having a beer and playing Smash while baking a late night cake.

I also want to thank my support system outside of Berkeley. Leon Senft has been one of my closest friends since middle school, and having him in the area for the last five years has been incredible. I feel very fortunate that our careers took us to the same area, since that was definitely not a given, and it has been extremely nice to have him nearby, especially as someone who knows me so well outside of academics.

I cannot thank my family enough. I have felt constant support during my degree, and I feel that we have managed to maintain a close relationship despite the distance apart. I want to specifically thank my parents, who have always believed in me and have made it possible to be where I am today. I also want to thank my brother Kyle and my sister-in-law Alyssa for their support. When I am feeling overwhelmed with existential dread, Kyle always has something insightful to say that helps me move past it. It has also been incredible to

celebrate Kyle's many life events in the past few years, between getting his wings and getting married, and I am inspired by his accomplishments. I have really cherished the time I have been able to spend with my family during my degree, and it has helped me reset my thinking throughout my degree. One small silver lining of the pandemic was being able to be closer to my family for a few months in the last year.

Finally, I want to thank my partner, Sarah Gord. I can't imagine what the last 3+ years would have been like without her, and I am so grateful that we have been able to share our time together. I truly feel that we are a partnership and complement each other perfectly. I have felt extremely well supported throughout the challenges of graduate school, and I know I can always count on her to be there for me when I don't feel good enough, am feeling uncertain about what to do next, or am generally feeling down. I am looking forward to starting a new chapter of our life together.

Introduction

Ecosystems are complex. They are highly dynamical and governed by many different interactions between a large number of organisms and their environment. Any individual in an ecosystem will interact with other individuals in its species, will have different relationships with individuals of other species in its surroundings, and will interact in complicated ways with its habitat. Describing any one of these interactions in depth could be a full dissertation on its own.

Despite this small scale complexity, there are patterns in ecology that appear consistently across ecosystems at large spatial or temporal scales (Brown 1995; Rosenzweig 1995; Gaston and Blackburn 2000; Magurran and McGill 2011). Many of these patterns have been identified and studied for a long time. This includes the relationship between the number of species and the area of the ecosystem (the species-area relationship or SAR) (Arrhenius 1921), and the distribution of the abundances of each species (the species-abundance distribution or SAD) (Fisher et al. 1943; Preston 1948). Other patterns include distributions of metabolic rates or body sizes (the metabolic distribution of individuals or MRDI), or the spatial distribution of individuals (the species-level spatial-abundance distribution or SSAD). Typical forms of these distributions are shown schematically in Fig. 0.1. These ubiquitous patterns can provide hints as to the structure of unifying theory in ecology (Lawton 1999; Hubbell 2001; McGill 2010; Harte 2011; McGill et al. 2019).

In physics, simplicity can often arise from complicated underlying behaviour, and there are many mathematical and analytical tools to describe this emergence. Statistical physics as a field studies and describes systems with large population sizes, using probability theory and statistic to understand aggregate properties of complicated systems. Methods from these types of analyses can be useful in studying these large scale patterns in ecology, where we are looking for statistical patterns emerging from complicated underlying mechanism (Banavar et al. 2010; Harte 2011; O'Dwyer and Chisholm 2014; Bertram and Dewar 2015; Azaele et al. 2016; Pigolotti et al. 2018; Gouveia et al. 2020).

One such tool is the method of maximum entropy, or MaxEnt. By maximizing Shannon information entropy (Shannon and Weaver 1949), MaxEnt selects the least informative probability distribution that is compatible with prior constraints (Jaynes 1957; Jaynes 1982). In physics, MaxEnt can be used to derive the Boltzmann distribution for the energy distribution of particles in an idealized gas (Jaynes 1957). In ecology, the Maximum Entropy Theory of Ecology, or METE, uses MaxEnt to simultaneously predict many macroecological patterns,

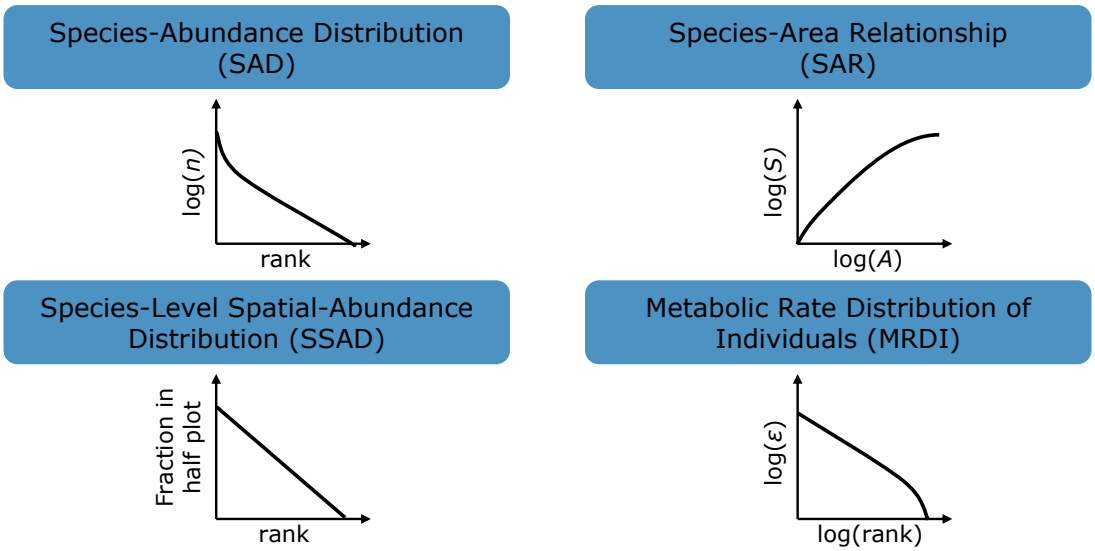


Figure 0.1: A schematic of some common macroecological patterns. All lines shown here are the corresponding METE predictions. The SAD shows rank ordered species abundances n, the SAR shows how the number of species S changes with the area of observation A, the SSAD shows the rank ordered fraction of individuals in one half of a plot, and the MRDI shows the rank ordered metabolic rates of individuals ε.

including those in Fig. 0.1 (Harte et al. 2008; Harte 2011; Harte and Newman 2014).

METE imposes prior constraints in the form of state variables: macroscopic observables analogous to pressure, volume or temperature in thermodynamics. These state variables capture enough of static ecosystems to describe their large scale patterns without having to describe underlying mechanism. In METE, the relevant state variables are the number of species S, the total number of individuals N, the total metabolic rate of all individuals E, and the area of observation A. For ecosystems in steady state, these state variables are sufficient for METE to accurately predict many macroecological metrics (Harte 2011; White et al. 2012; McGlinn et al. 2013; Xiao et al. 2015). Steady state, in this case, is characterized by relatively constant state variables. This is distinct from true equilibrium, as the underlying dynamical biological processes are still occurring.

Outside of steady state, when the state variables are changing relatively rapidly, METE predictions are generally less successful (Newman et al. 2020; Franzman et al. 2021). This is analogous to the ideal gas law, which is valid for an idealized gas at steady state when derived using the MaxEnt predicted Boltzmann distribution for the particle energies, but which may no longer hold when the gas is out of steady state, such as in a gas with a rapidly changing temperature. In this case, the pressure, volume, and temperature no longer determine the energy distribution of particles and the ideal gas law may not hold. In the case of METE,

it seems that different types of disturbance affect patterns in different ways (Kempton and Taylor 1974; Carey et al. 2006; Supp et al. 2012; Rominger et al. 2016), and therefore deviations from METE can provide useful insights about the ecosystem. This additionally makes it natural to modify METE by incorporating mechanistic disturbance to predict how patterns change over time.

In this dissertation, I use METE as a baseline to study how changes in macroecological distributions out of steady state can inform us about the underlying biology, and to build theory to understand and predict how these patterns will change over time in ecosystems under disturbance. This is especially important given the degree of anthropogenic disturbance and global change in the Anthropocene (Turner 2010; Pereira et al. 2012; Díaz et al. 2019; Newman 2019).

The first chapter focuses specifically on the effects of land use, given that land management is a primary driver of ecological disturbance around the world (Foley et al. 2005; Newbold et al. 2015; Newbold et al. 2018). I compare METE predictions to macroecological patterns of arthropods in the Azores Islands, an isolated archipelago in the Atlantic Ocean. Human settlement has drastically changed the island from largely undisturbed natural forest to mixed land uses, including intensively managed pasture land (Cardoso et al. 2009; Norder et al. 2020). Comparing the SAD, MRDI, and SAR predictions to data provides information about how land use disturbance is affecting these arthropod communities.

The second chapter focuses on a specific type of disturbance, density dependence, and its effects on the spatial predictions of METE. I present a new model of spatial aggregation that uses METE as a starting point and includes a parameter that characterizes the strength of negative density dependence (Brush and Harte 2021). This parameter can be interpreted as a density dependent correction away from the MaxEnt predicted spatial distribution, and by analysing spatially explicit data we can determine to what degree ecosystems are consistent with different levels of underlying density dependence.

Finally, in the third chapter I present DynaMETE: a broad theoretical framework for predicting how macroecological patterns change in time in response to disturbance (Harte et al. 2021). This theory combines explicit mechanism governing disturbance and MaxEnt inference, and reduces to METE in steady state. I provide code for iterating the theory forward in time given the initial iteration scheme and discuss possible alternative formulations with the eventual goal of connecting the theory results directly to data.

Chapter 1

The influence of land use on arthropod macroecology in the Azores

Abstract

Human activity and land management practices have resulted in global loss of biodiversity. These types of disturbances affect the shape of macroecological patterns, and analysing these patterns can provide insights into how ecosystems are affected by land use. The Maximum Entropy Theory of Ecology (METE) is a theory that simultaneously predicts many of these patterns. Its predictions are successful across habitats and taxa in undisturbed natural ecosystems, though they perform less well in disturbed ecosystems. Deviations from METE therefore contain information about the effects of disturbance. We here compare predictions from METE to arthropod data from Terceira Island in the Azores archipelago across four different land uses. Ranked in order of management intensity, these land uses are: 1. Native forest, 2. Exotic forest, 3. Semi-natural pasture, and 4. Intensive pasture. We simultaneously predict the species-abundance distribution (SAD), the metabolic rate distribution of individuals (MRDI), and the species-area relationship (SAR) and compare to observations at 96 sites across the four land uses. Across these patterns, we find that the forest sites are the best fit by METE and the semi-natural pasture is consistently the worst fit by METE. The intensive pasture is intermediately well fit for the SAD and MRDI, and comparatively well fit for the SAR, though the residuals are not normally distributed. The direction of failure of the METE predictions at the pasture sites is likely due to the highly abundant introduced spider species present there. We hypothesize that the particularly poor fit at the semi-natural pasture is due to the mix of arthropod communities out of equilibrium and the changing management practices throughout the year, whereas the comparatively better fit at the intensive pasture results from arthropod communities that are well adapted to intensive management.

1.1 Introduction

Human management of land is a primary driver of ecological disturbance around the world (Foley et al. 2005; Pereira et al. 2012; Klein Goldewijk et al. 2017). Land use has large effects on landscape heterogeneity, and can fragment the habitat of endemic or native species leaving a mosaic of habitat types (Fahrig 2003; Fischer and Lindenmayer 2007; Cardoso et al. 2009; Fahrig 2019). This type of human driven disturbance is leading to global biodiversity loss (Martins et al. 2014; Pimm et al. 2014; Newbold et al. 2015; Maxwell et al. 2016; Newbold et al. 2018). On oceanic islands, this is especially pronounced as conversion of native vegetation to agricultural and pasture land may put native and endemic species at risk through the spread of exotic species (Gillespie and Roderick 2002; Borges et al. 2006; Gillespie et al. 2008).

Disturbance as a result of human activity can often be observed in macroecological patterns, as deviations from the expected shapes can be interpreted as disturbance (Kempton and Taylor 1974; Carey et al. 2006; Dornelas et al. 2009; Supp et al. 2012; Matthews and Whittaker 2015; Franzman et al. 2021). In order to interpret the patterns in this way, we require a theoretical prediction for what these different patterns should look like in an ecosystem that has not been disturbed or managed.

The Maximum Entropy Theory of Ecology (METE) predicts numerous macroecological patterns simultaneously using the principle of maximizing information entropy (Harte et al. 2008; Harte 2011; Harte and Newman 2014; Brummer and Newman 2019). METE is characterized by three so-called state variables that are used to constrain the predicted distributions for a given ecosystem: the number of individuals N_0, the total metabolic rate E_0, and the species richness S_0. To make spatial predictions, METE also requires the total area of the site A_0. It has been found to well describe empirical patterns across diverse taxa and habitats (Harte 2011; White et al. 2012; Xiao et al. 2015). However, there is increasing evidence that these predictions perform less well in disturbed ecosystems (Carey et al. 2006; Rominger et al. 2016; Newman et al. 2020; Franzman et al. 2021; Harte et al. 2021). Most disturbance to this point in relation to METE has been characterized by ecosystems with rapidly changing state variables, and METE seems to well describe ecosystems where the state variables are relatively constant in time.

The effects of land use and related disturbances on macroecological patterns and their deviations from METE have not yet been explored. Given that METE predictions appear to perform better in pristine ecosystems, we expect that land uses that introduce significant disturbance should result in patterns that deviate from them. How well the data fit METE across land use can then tell us something about disturbance at that land use. Analyzing ecosystems in this way can provide insights about how different land uses affect these large scale patterns, and by extension how management can affect biodiversity.

Here, we investigate how land use affects several patterns predicted by METE with arthropod data from Terceira Island in the Azores archipelago. The Azores are an isolated island chain in the Atlantic Ocean that have been populated for about 600 years (Norder et al. 2020). In this time, they have undergone a large change from largely undisturbed natural

forest to mixed land uses, including managed forest plantations and intensive pasture land (Cardoso et al. 2009).

Arthropod species have also been introduced. These exotic species have changed the ecological landscape (Florencio et al. 2013) and have different functional trait composition than indigenous species (Rigal et al. 2018). However, they appear to be integrated in these ecosystems, perhaps by replacing indigenous species or filling empty niche space (Gaston et al. 2006; Rigal et al. 2013).

Species-abundance distributions of arthropods in the Azores have already been shown to be useful for biogeographical purposes (Fattorini et al. 2016; Borda-de-Água et al. 2017). Additionally, Rigal et al. (2018) found that functional trait composition varies strongly with management intensity. This, together with the fact that there are small but important regions of remaining native forest, makes the Azores an ideal system to test how land use affects macroecological patterns and their deviation from METE predictions.

Here we will analyze three predictions of METE simultaneously: the species-abundance distribution (SAD), the metabolic energy rate distribution of individuals (MRDI), and the species-area relationship (SAR). Analyzing multiple patterns simultaneously avoids treating any individual pattern in isolation (McGill et al. 2007), especially since single patterns can often be predicted from many different underlying theories. We compare all three of these patterns with arthropod data across land use types on Terceira Island. We expect to see that METE predictions fit the data better for less intensively managed land uses, and we also analyze deviations from the predicted patterns for information about how disturbance is affecting the ecosystem.

1.2 Methods

Study area and data

The Azores Islands are an isolated island chain in the Atlantic Ocean of volcanic origin. All of the data analyzed here come from Terceira Island, which before human colonization was almost entirely forest and now has a mix of land uses, including agricultural and intensively managed pastures. The four major land uses, ranked in increasing order of management intensity, are 1. Native forest, 2. Exotic forest, 3. Semi-natural pasture, and 4. Intensive pasture (Rigal et al. 2018). These land uses comprise about 87% of the total island area, which is broken down by land use in Table 1.1 (Cardoso et al. 2009). Figure S1.1 in Rigal et al. (2018) shows a land use distribution map of Terceira Island with more specific spatial information.

The native forest is made of perennial trees and shrubs adapted to a hyper-humid Atlantic climate, and is now restricted to elevations above 500 m above sea level (a.s.l.) and dominated by *Juniperus-Ilex* forests and *Juniperus* woodlands (Elias et al. 2016). Exotic plantations of the fast growing tree *Cryptomeria japonica* were planted after the Second World War to reforest large areas of previous native forest that were destroyed in the previous decades for

Land use	% Area	Sites	Total S_0	Total N_0	Median S_0	Median N_0
Native forest	9	44	148 (86, 60)	10 291 (7288, 3001)	24 (16, 8)	195 (129, 50)
Exotic forest	15	12	87 (44, 42)	3385 (1476, 1908)	20 (10, 9)	196 (11, 51)
Semi-natural pasture	15	16	127 (50, 76)	11 421 (2110, 9310)	28 (10, 17)	766 (101, 623)
Intensive pasture	48	24	136 (40, 94)	21 153 (4076, 17 070)	36 (10, 27)	878 (161, 684)

Table 1.1: The total number of species and individuals observed for each land use, and the median number across sites within one land use. The number in parentheses is the number of indigenous species or individuals, followed by the number of exotic species or individuals. Additionally, the number of sites where data was collected for each land use, and the percent of the total island area occupied by that land use. Across all land uses, there are a total of 271 species and 46 250 individuals, with four species constituting 11 individuals that are not identified as indigenous or exotic.

fuel. These plantations are dense and almost no understory is present. Semi-natural pastures are located around 400-600 m a.s.l., have a mixture of native and exotic herbs and grasses, and are mostly grazed in the spring and summer with low cattle density. Intensive pastures are located between 100-500 m a.s.l. and are grazed every three weeks (and sometimes up to every 12 days in the summer) with high cattle density.

The data set analyzed here was collected from pitfall traps. Each of the 96 sites has a single 150 m transect with 30 pitfall traps spaced out at 5 m intervals. All data was collected over summers on Terceira Island (see Cardoso et al. 2009; Rigal et al. 2018). Measurements of individuals were performed as described in Rigal et al. (2018) and Macías-Hernández et al. (2020). For spiders we have measurements of both males and females and at least four individuals per sex. For some beetle species we have also several individuals measured.

Table 1.1 shows the number of sites analysed for each land use, along with the total and median number of species S_0 and individuals N_0 across all transects of that land use. The number of indigenous and exotic species or individuals are shown in parentheses. Indigenous species are defined as those that are endemic (occur only in the Azores) or native (appear in the Azores Islands and other nearby islands). Exotic species are those believed to have been introduced by human settlement. Unidentified species that share a genus, subfamily, or family with other species present in the archipelago are put into the same colonization category as those species (Borges et al. 2010; Florencio et al. 2013). There are four remaining species that are not identified as indigenous or exotic that constitute 11 individuals. Across all land uses, there are a total of 271 species and 46 250 individuals, with 126 indigenous species and 14 950 indigenous individuals and 141 exotic species and 31 289 exotic individuals.

METE Review

METE as a theory predicts many different macroecological patterns simultaneously by maximizing Shannon information entropy given a set of constraints (Harte et al. 2008; Harte 2011; Harte and Newman 2014; Brummer and Newman 2019). Its core distribution is the ecological structure function $R(n, \varepsilon|S_0, N_0, E_0)$, which is a joint conditional distribution over abundance n and metabolic rate ε given the number of species S_0, the number of individuals N_0, and the total metabolic rate E_0. Thus, $Rd\varepsilon$ is the probability that a species is picked from the species pool has abundance n, and an individual picked at random from that species has metabolic rate between ε and $\varepsilon + d\varepsilon$. Note that n is discrete, and ε is continuous. In practice, we scale the metabolic rate such that the smallest metabolic rate in the ecosystem has $\varepsilon = 1$.

We then use the method of Lagrange multipliers to maximize the information entropy $\sum_n \int_\varepsilon d\varepsilon R \log(R)$ given the following constraints:

$$\begin{aligned} \frac{N_0}{S_0} &= \sum_{n=1}^{N_0} \int_{\varepsilon=1}^{E_0} d\varepsilon \; nR(n, \varepsilon) \\ \frac{E_0}{S_0} &= \sum_{n=1}^{N_0} \int_{\varepsilon=1}^{E_0} d\varepsilon \; n\varepsilon R(n, \varepsilon). \end{aligned} \tag{1.1}$$

We additionally require the distribution R to be normalized such that $\sum_{n=1}^{N_0} \int_{\varepsilon=1}^{E_0} d\varepsilon \; R(n, \varepsilon) = 1$. The solution for the ecological structure function is

$$R(n, \varepsilon|S_0, N_0, E_0) = \frac{\exp\left(-\lambda_1 n - \lambda_2 n\varepsilon\right)}{Z}, \tag{1.2}$$

where the Lagrange multipliers λ_1 and λ_2 are solved from the constraints, and the normalization Z is calculated as $\sum_{n=1}^{N_0} \int_{\varepsilon=1}^{E_0} d\varepsilon \; \exp\left(-\lambda_1 n - \lambda_2 n\varepsilon\right)$. To a very good approximation given typical empirical values for S_0, N_0, and E_0 (Harte 2011; Brummer and Newman 2019), $\lambda_2 = S_0/(E_0 - N_0)$, and λ_1 can then be solved from

$$\frac{N_0}{S_0} = \frac{\sum_{n=1}^{N_0} e^{-\beta n}}{\sum_{n=1}^{N_0} e^{-\beta n}/n}. \tag{1.3}$$

where $\beta = \lambda_1 + \lambda_2$.

The distribution R can then be used to derive other macroecological distributions that can be compared to data. We here show derivations of the relevant distributions.

Species-abundance distribution

We obtain the METE SAD prediction $\Phi(n)$ by integrating the structure function over ε. The METE prediction is equivalent to the max likelihood prediction for the log series (White et al. 2012, Appendix A). This prediction assumes that the number of species is large enough

that we can ignore certain terms, which eliminates any dependence on E_0. The resulting prediction is the log series distribution

$$\Phi(n|S_0, N_0) = \frac{e^{-\beta n}}{n \log\left(1/\left(1 - e^{-\beta}\right)\right)} \tag{1.4}$$

Metabolic rate distribution of individuals

The METE MRDI prediction $\Psi(\varepsilon)$ is obtained by summing the structure function multiplied by n over n and correcting the normalization,

$$\Psi(\varepsilon) = \frac{S_0}{N_0} \sum_{n=1}^{N_0} nR(n, \varepsilon). \tag{1.5}$$

This sum gives

$$\Psi(\varepsilon|S_0, N_0, E_0) = \lambda_2 \left(e^{\beta} - 1\right) \frac{e^{-\gamma}}{(1 - e^{-\gamma})^2} \tag{1.6}$$

where $\gamma = \lambda_0 + \lambda_1 \varepsilon$. Note that we use a slightly different form for Ψ compared to Eq. 7.33 in Harte (2011), where β has been replaced with $e^{\beta} - 1$. The normalization of $\Psi(\varepsilon)$ in Eq. 1.6 is significantly better, as $\int_{\varepsilon} d\varepsilon\ \Psi(\varepsilon)$ is much closer to 1. This form still allows the cumulative distribution function and the rank ordered distribution to be solved analytically, and is numerically very similar to the full expression without any approximations, even down to the individual transect level.

Species-area relationship

The SAR can be predicted by combining the SAD with the species-level spatial abundance distribution (SSAD) $\Pi(n|A, A_0, n_0)$, which predicts the number of individuals present in an area A given n_0 individuals of that species in a larger area A_0. The number of species at a given scale A can be predicted by multiplying the SAD by the probability that a species with abundance n_0 is present at that scale and then summing over n_0,

$$S(A) = \sum_{n_0=1}^{N_0} \Phi(n_0) \left(1 - \Pi(0|A, A_0, n_0)\right). \tag{1.7}$$

The SSAD can also be predicted by maximizing information entropy given the constraint $\sum_{n=0}^{n_0} n\Pi(n) = n_0 A/A_0$. The solution corresponds to the finite negative binomial distribution (Conlisk et al. 2007; Zillio and He 2010)

$$\Pi(n|A, A_0, n_0) = \frac{\binom{n+k-1}{n}\binom{n_0-n+kA_0/A-k-1}{n_0-n}}{\binom{n_0+kA_0/A-1}{n_0}}. \tag{1.8}$$

with aggregation parameter $k = 1$ (Harte 2011; Wilber et al. 2015).

METE also predicts that all nested SARs will collapse onto a single universal curve when plotted as the slope of the SAR z versus $D = \log(N_0/S_0)$, a scale parameter. (Harte 2011; Wilber et al. 2015)

Comparing METE predictions with data

We divide the data by land use and compared the observations to the predictions of METE. Because we have multiple transects for each land use, we can either compare our results by aggregating across all transects in one land use category, or at the individual transect level by treating each transect as a replicate. We primarily analyze our results at the individual transect level. This is because METE makes predictions within a single community, and aggregating data over several locations may create a mismatch between the theory predictions and the data. Additionally, the number of species observed will not be the same across many small patches compared to a single large patch of the same area. Despite these issues, we do find mostly similar results aggregating the data by land use (Appendix A.1).

Our data are separated into juvenile and adult data. Where possible, we treat these together as a single dataset that accurately captures all ground dwelling arthropods. Note that some species will have adult forms that will not be captured by pitfall traps. However, in some sense the arthropods that are not captured by the pitfall traps are not part of the same ecosystem of ground dwelling arthropods. These individuals likely interact with different species and habitats, and thus do not necessarily need to be included for METE to make accurate predictions about the composition of the study ecosystem. Unfortunately, for the predictions for metabolic rate, we were only able to use adult arthropods as we use scaling relationships to calculate the empirical metabolic rates that are only available for adults.

Finally, our data categorizes species as indigenous (native or endemic) or exotic (likely introduced by humans). We analyze these groups independently for the species-abundance distributions, though we believe these species are integrated into the community, at least in macroecological patterns (Gaston et al. 2006; Rigal et al. 2013).

Species-abundance distribution

There are many existing methods for comparing empirical SADs to data, however there is not single metric that can well capture goodness of fit (Connolly and Dornelas 2011; Matthews and Whittaker 2014). In general, binning should be avoided as the shape of the distribution depends on the binning interval, and given the sparse data each observed bin is unlikely to have a sufficiently large expectation to have a meaningful χ^2 test (Williamson and Gaston 2005; Gray et al. 2006; Ulrich et al. 2010). Log-likelihood based methods are also inappropriate in this case as we are not trying to determine a preferred model and instead are looking for a goodness of fit test.

Given this, we use the mean least squares of the rank ordered natural log of abundance as our primary goodness of fit metric to compare the METE predictions to data. Although

Matthews and Whittaker (2014) notes that mean least squares of rank ordered data violates some underlying statistical assumptions, namely that the data points are not statistically independent (Connolly and Dornelas 2011), we still find it preferable to their recommended method of using a parametric bootstrap. For many test statistics this method scales with the number of points, making comparisons between land uses challenging. As that is our primary objective here, we prefer to use the mean least squares.

To ensure our results are robust to our choice of goodness of fit metric, we additionally performed Kolmogorov-Smirnov (KS) tests for each transect and obtained at the two-sided test statistic D_{KS} when the empirical CDF is compared to the METE predicted CDF. These results can be found in Appendix A.2, and are very similar to the results obtained by mean least squares. For the KS test for the SAD, we used the R package DGOF which implements the KS test for discrete distributions (Arnold and Emerson 2011).

Metabolic rate distribution of individuals

The raw data includes information on average body lengths for each species of arthropod. For adult arthropods, we can use empirical scaling relationships to convert these to body mass data, which can be related to metabolic rate using metabolic scaling. Since we only have average body lengths, we do not capture the variation in body lengths within a single species. To reintroduce this variation, we look at several beetle and spider species where we have multiple measurements for body length and convert these into distributions for intraspecific variation in body mass. We find in this case that these distributions can be roughly approximated by normal distributions. Gouws et al. (2011) also found that intraspecific body size distributions are usually normal across many species of insects, including beetle species. Note that for spiders we find that this distribution is much more likely to be bimodal because of the sexual dimorphism present in many of the species, and in this data information about the sex of the individual is available for species with multiple individuals. However, the full dataset provides only one average body length regardless of sex, and so we approximate the spiders body mass distribution as a single normal distribution, as with the beetles. This approximation should be closer to representing the female spiders as we are more likely to sample a larger number of them due to their longer life span.

For both the beetles and the spiders, we find that the relationship between the $\log_{10}$ of the mean body mass and the $\log_{10}$ of the variance is roughly linear, without an obvious trend in the residuals. The slopes and intercepts are similar for the beetles and spiders, although the slope for the spiders is slightly steeper (2.22 ± 0.13 versus 1.99 ± 0.12). This means the variance will be higher for spiders for most species with comparable mass, which makes sense given that we expect the true distribution for the spiders to be bimodal. Plots of individual beetle and spider species and the log(body mass) versus log(variance) relationship can be found in Appendix A.3.

We then use these relationships for our analysis. For each species, we draw n_0 samples from the corresponding normal distribution for body mass before converting to metabolic rate and rank ordering. Because the samples are drawn randomly, some samples for low

mass species will be below zero, even though mass must be strictly positive. We fix this by setting these draws to the smallest positive mass plus a small amount of noise (1% of the mass multiplied by a normal distribution with mean and variance equal to one) to avoid duplicates. We used the relationship for beetles for all orders except for Araneae, where we used the relationship we obtained for spiders. Some of the orders may be more similar to spiders, or be quite different overall, but since Coleoptera and Araneae are the two most common orders in the dataset, difference among other orders should not overly impact the analysis.

As with the SAD, we primarily use the mean least squares of the rank ordered natural log of metabolic rates to as a goodness of fit metric at each land use. Alternative options for this comparison would be R^2 as defined in eg. Xiao et al. (2015), or to bin the data and use a χ^2 test. Again here though, binning relies on a large number of points per bin that we will not have at the individual transect level, and the results can depend on the bin width. In order to test the robustness of our goodness of fit metric we additionally performed Kolmogorov-Smirnov goodness of fit tests for the empirical CDF compared to the METE predicted CDF. Again, we obtain similar results to the mean least squared analysis (Appendix A.2).

Species-area relationship

At each transect, there are 30 individual pitfall traps arranged linearly. We compare the resulting empirical SAR to the METE prediction by first averaging the number of species at different scales. We chose scales of 1, 2, 3, 5, 6, 10, 15, and 30 traps. These relative scales were chosen as they use all of the data available at every scale, or put another way, these numbers are all factors of 30. This is slightly different than has been done in other comparisons, which use a number of cells that is a factor of two and divide repeatedly in half (eg. Franzman et al. 2021).

To then compare these average numbers of species to the METE prediction, we use the functions for the log series SAD $\phi(n)$ and the finite negative binomial SSAD $\Pi(n)$ from the `macroeco` software package (Kitzes et al. 2015; Kitzes and Wilber 2016), which we have adapted for use with Python 3.0.

There are many different methods for comparing the predicted SAR to data. We could fix the number of species and individuals at the largest scale and predict the number of species at every smaller scale from this anchor scale, or we could compare the predicted slope z of the SAR against the scale parameter $D = \log(N_0/S_0)$, which as noted above collapses SARs onto a single, universal curve (Harte 2011; Wilber et al. 2015). Harte (2011) provides an equation for z given the empirical number of species and individuals at a given scale, assuming that we bisect the plot in two. We use a similar approach, in that we use the number of species and individuals at a given scale to predict the number of species at the next smallest scale in consideration and then use that to predict a slope. Mathematically, we write

$$z_i = \frac{\log(S_i/S_{i-1})}{\log(A_i/A_{i-1})}, \tag{1.9}$$

where i indexes the scale. In the list above then, $i = 1$ corresponds to the average number of species at the scale of 1 cell, and $i = 8$ corresponds to the number of species in all 30 cells.

Given we are predicting the number of species directly using this approach, we can also compare the SAR directly (see Appendix A.4). However, given the collapse onto a single curve in the $z - D$ plots we prefer that comparison. We obtain the empirical slope by comparing the number of species at the scale being considered to the number of species at the next smallest scale. This is similar to the theoretical prediction, but now the number of species at the smaller scale is also empirical. This method allows us to compare slopes at all except the smallest scale, as we do not have a smaller scale with which to make the empirical slope prediction.

We therefore have 7 data points to make the comparison and again use the mean least squares between the prediction and the empirical data to judge goodness of fit to METE. Note that for many transects we will have fewer than 7 data points as we additionally only use scales where the empirical average for the number of species is greater than four ($S_0 > 4$). This is because several METE simplifications break down for small S_0, including the fact that we can ignore E_0 and derive β from only N_0 and S_0. This means that transects with lower abundance we will have fewer than 7 points of comparison.

1.3 Results

Species-abundance distribution

The closed circles in Fig. 1.1 show the mean of the mean least squared error over transects with its standard error at each land use for the SAD. We find that the semi-natural pasture is particularly poorly described by METE, and the native forest is the most well fit. The exotic forest and intensive pasture are fairly similar and intermediate between the native forest and the semi-intensive pasture. The standard deviation of the mean is the lowest for the native forest sites, and the highest for the semi-natural pasture. Similar results for the KS test statistic D_{KS} across transects are shown in Appendix A.2.

Plots of the SAD at each transect are shown in Appendix A.5. To combine all of the SADs onto a single plot, we plot the residuals of $\log_{10}$(abundance) in Fig. 1.2, where the x-axis has been scaled by the number of species to facilitate comparison. Each line in this plot represents the SAD at a single site. In the semi-natural pasture in particular, we see that METE consistently under predicts the abundance of the most abundant species and over predicts the abundance of the species and intermediate rank. We can see a similar pattern across land uses where the residual of the most abundant species tends to be positive, though it is most prevalent at the pasture sites and least common at the native forest sites. Across all sites, METE generally under predicts the number of singletons, though this is again least common at the native forest sites.

Results aggregating over transects rather than treating them as replicates can be found in Appendix A.1. In that case, the forest sites are again better fit than the pasture sites,

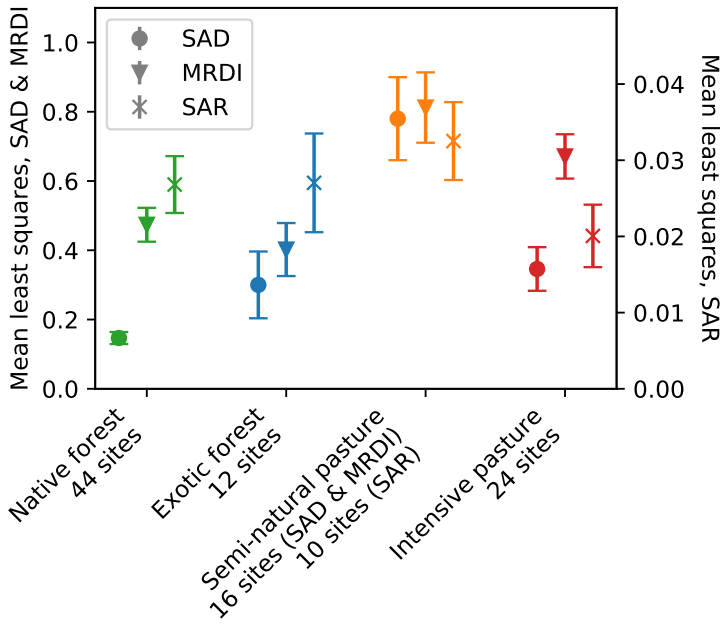


Figure 1.1: A comparison of means of the mean least squares and their standard errors across all three patterns and land uses. Note the difference in y-scale for the SAR, where the mean least squared error was much smaller. The shape of the marker indicates the pattern and the color indicates the land use.

but the ranking is slightly different. The overall goodness of fit is also worse, which is in line with our expectation that METE predictions are more accurate for individual transects.

Indigenous and exotic species

We additionally analyzed the SADs independently for species classified as indigenous and exotic at each land use. Figure 1.3 shows the mean and associated standard error of mean least squares across transects, separated by species that are indigenous and introduced, at all four different land uses. We see that by far the biggest difference between indigenous and introduced species is at the semi-natural pasture sites, where the introduced species fit quite poorly compared to the indigenous species. Across other land uses indigenous and introduced species are comparably well fit, though the introduced species fit slightly better at the exotic forest sites. Again, these results are similar when analyzed at the community level (Appendix A.1).

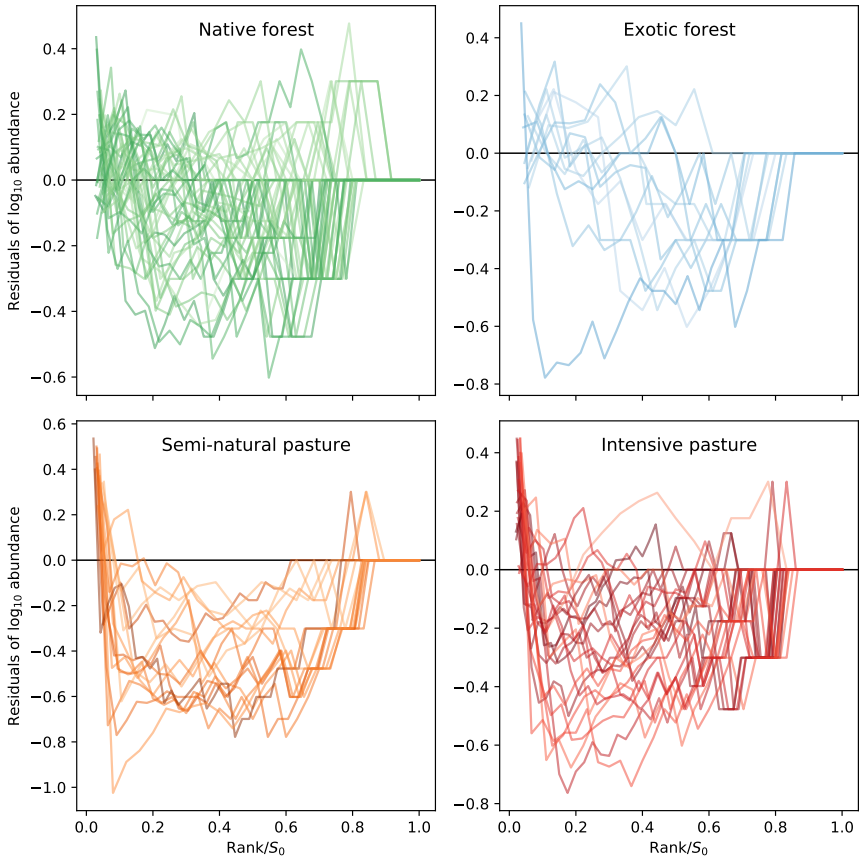


Figure 1.2: The observed $\log_{10}$ of abundance minus the predicted $\log_{10}$ of abundance from METE for each transect across land uses. The darker lines are sites with a higher number of species, and lighter lines represent sites with fewer species. The colors correspond to the different land uses.

Metabolic rate distribution of individuals

The triangles in Fig. 1.1 show the mean and its standard error for the mean least squared error across transects for the MRDI. This metric shows that the forest sites are better fit by METE than the pasture sites, and the semi-natural pasture is particularly poorly fit by METE.

The rank ordered plots for the MRDIs at each transect are plotted in Appendix A.6, and Fig. 1.4 shows the residuals from the rank ordered metabolic rates at each transect on a single plot. Each line here represents a single site at that land use. Across land uses, we see that there are often far more individuals around a single metabolic rate than predicted by METE at any given site. This appears as long lines of similar slope in the residuals. See particularly the pasture sites, where this pattern in the residuals is especially common.

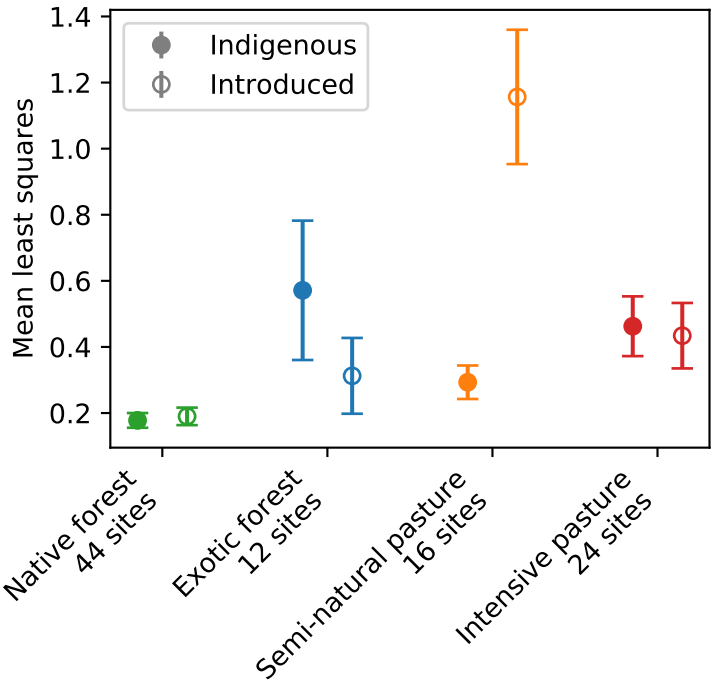


Figure 1.3: Mean and the standard error of the mean of the mean least squares for the SAD across transects for each land use, for both indigenous (closed circles) and introduced species (open circles). Colors represent land use type.

The results are again similar if analyzed using the KS test statistic (Appendix A.2) or at the community level (Appendix A.1, though here the intensive pasture is much worse fit).

Species-area relationship

The X markers in Fig. 1.1 show the mean of the mean least squared error over transects with its standard error at each land use for the slope z. Note the different y-axis scale compared to the SAD and MRDI. Here, the mean least squared error for each transect has been averaged over the number of scales where the empirical $S_0 > 4$.

Figure 1.5 compares the data for each site, organized by land use. Each point here represents a single transect at a single scale $D = \log(N_0/S_0)$, and the lines are the corresponding METE predictions. The scatter in this plot is significant across land use. Looking at the residuals, we see that METE tends to under predict the slope at larger scales, particularly in the pasture sites. We can see this pattern across sites in Fig. 1.6, which uses the scale-collapse of the $z - D$ relationship to display all of the sites together on a single plot.

The results are similar if we analyze the predicted number of species at each scale rather

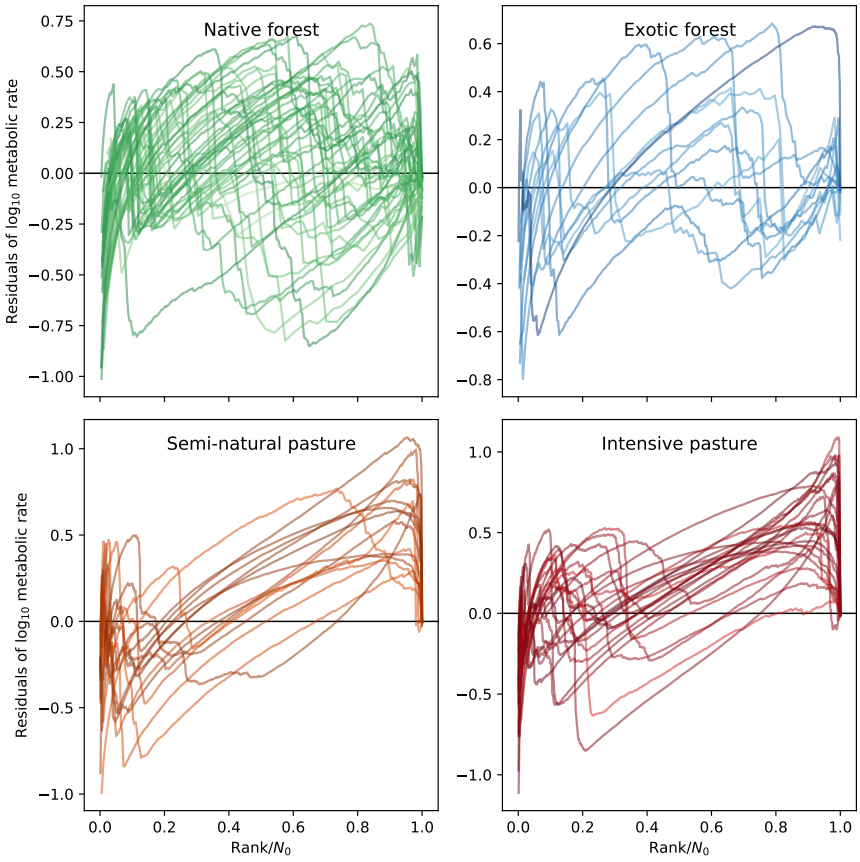


Figure 1.4: The observed $\log_{10}$ of the metabolic rate minus predicted $\log_{10}$ of the metabolic rate for the rank ordered plots. The darker lines are sites with a higher number of individuals, and lighter lines represent sites with fewer individuals. The colors correspond to the different land uses.

than the slope (Appendix A.4).

1.4 Discussion

Species-abundance distribution

The most distinctive pattern in the SAD residuals in Fig. 1.2 is the consistent under prediction of the most abundant species, particularly at the pasture sites. In other words, these sites have a few very abundant species that are much more abundant than predicted by METE. We attribute this to small-bodied, highly dispersive, mostly introduced spider species. Rigal et al. (2018) found that these types of species were very prevalent at sites with high land use intensity.

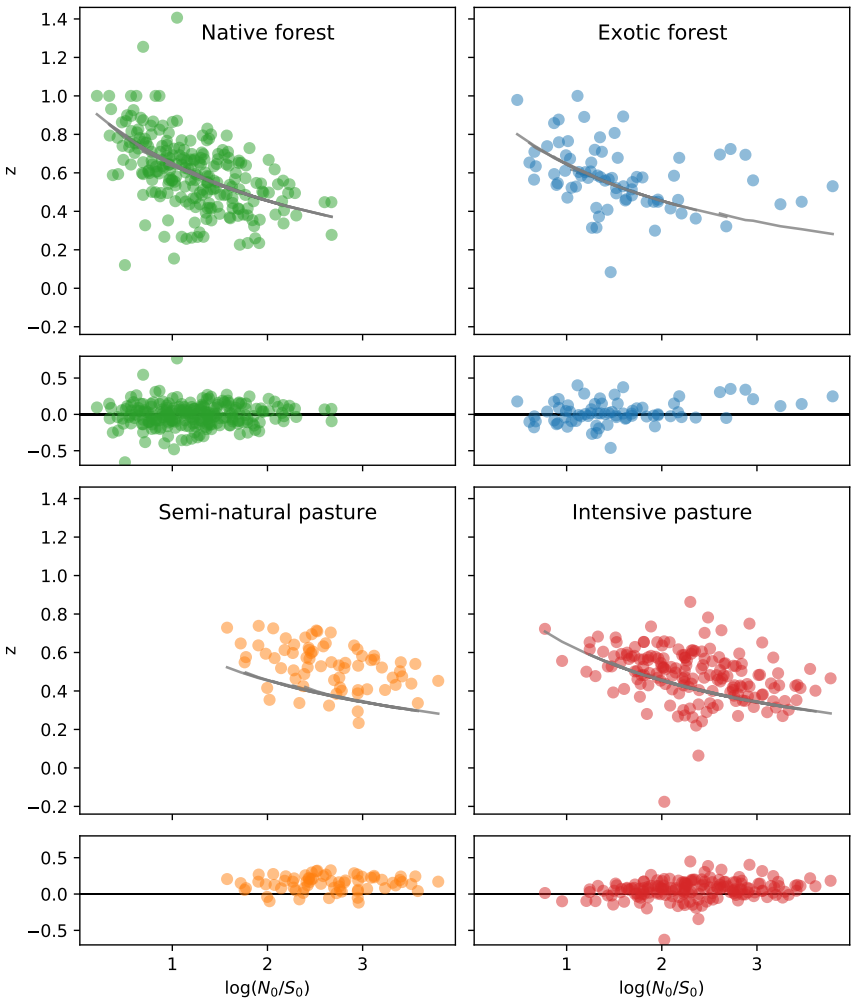


Figure 1.5: The species-area relationship for each transect across land use. Each point represents a single transect at a specific scale, where the scale is determined by $D = \log(N_0/S_0)$, and are colored according to land use type. The gray lines are the METE predictions. Here we have plotted the slope of the relationship on the y-axis so that all points collapse onto one universal curve. The residuals for each land use are shown immediately below the plot for that land use.

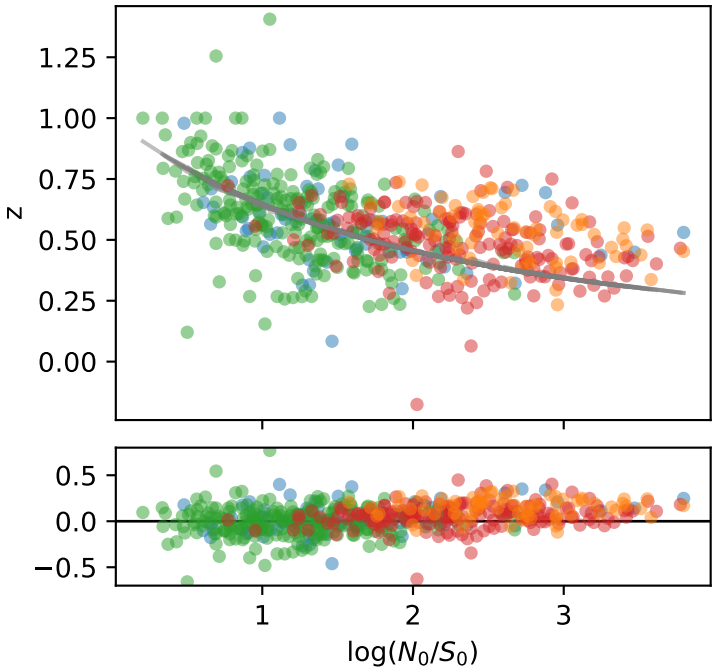


Figure 1.6: The species-area relationship for each transect all on a single plot, color coded by land use. Each point represents a single transect at a specific scale, where the scale is determined by $\log(N_0/S_0)$. The gray lines correspond to the METE predictions. Here we have plotted the slope of the relationship on the y-axis so that all points collapse onto one universal curve. The residuals are shown below.

The dispersal ability of these spiders also appears in the SAD. Borda-de-Água et al. (2017) consider how the shape of the SAD changes according to different dispersal abilities in Azorean arthropods, and they find that an intermediate mode in abundance appears in the Preston plot (number of species versus $\log_2$(abundance)) for lower dispersal species. When rank ordered, this results in plots that are less steep at low rank. For species with high dispersal rates, the rank ordered curve should then be steeper. We see these steep rank ordered SADs in the pasture sites in Fig. 1.2 and in Appendix A.5, which correspond to the highly dispersive introduced spider species. These species have not yet been selected for reduced dispersal as they have not been present on the island for evolutionary time scales.

These spider species additionally have multiple generations per year, which allows them to recolonize the pasture sites from the nearby surrounding forest after disturbance due to land management or cattle grazing (see Rigal et al. 2018, Figure S1.1 for a map). This recolonization is particularly relevant for the semi-natural pasture sites, where cattle grazing and fertilization is seasonal. The particularly high abundance of these species in this data may be related to the fact that this data was collected in the summers, when the semi-natural

pasture sites are likely to be subject to cattle grazing.

Another pattern observed in the residuals is that METE tends to under predict the abundance of the intermediate rank species, again particularly at the pasture sites. However, given that METE is constrained such that the number of individuals must be N_0, this is likely a consequence of these highly abundant spider species. If the METE SAD under predicts the abundance at some rank, it must over predict the abundance at another rank such that this constraint is satisfied.

Similarly, METE tends to under predict the number of singletons at most sites, except in some cases for the native forest sites where we see that METE over predicts the number of species with small abundance. This could be related to sampling, as the traps are less likely to capture multiple individuals of rare species, but it could also be related to the METE prediction for the number of singletons. METE predicts a number of singletons equal to βN_0, and therefore increasing N_0 while holding S_0 constant decreases the expected number of singletons (Harte 2011, Chapter 7.3). From Table 1.1, we see that N_0/S_0 is large for the pasture sites compared to the forest sites, and therefore METE predicts proportionally fewer singletons at these sites. If ecologically we still expect a similar number of singletons, given that the high N_0/S_0 is being driven by a small number of very abundant species, then this could mean that METE under predicts the number of singletons.

Indigenous and exotic species

The results by separating indigenous and exotic species support our hypothesis that the poor fit of the SADs for the semi-natural pasture sites is driven by introduced species. Outside of the semi-natural pasture, there is little difference in goodness of fit for the SADs between indigenous and introduced species. This is in line with previous studies that studied the relationship between occupancy, variance, and abundance (Gaston et al. 2006) and the interspecific abundance-occupancy relationship (Rigal et al. 2013) and found that exotic species were integrated with the indigenous species in the Azorean community. Some evidence for this integration comes from comparing the goodness of fit results in Fig. 1.1 and Fig. 1.3. We find that for all land uses except the semi-natural pasture, the mean least squares of the indigenous and introduced species are each slightly higher than the goodness of fit for the SAD when these species are combined. This means that when the indigenous and introduced species are analyzed together, the METE prediction performs better than when they are analyzed separate, except for the semi-natural pasture.

One reason this could be the case is that these species are filling different niches. Rigal et al. (2018) found different functional trait profiles for indigenous and introduced species across the four different land uses considered, which could mean that introduced species are filling vacant ecological niches left by indigenous species. This should be particularly true at the pasture sites, which rely mostly on species introduced by humans as indigenous species are not well adapted to these novel habitats.

That the semi-natural pasture has such a difference in goodness of fit between the indigenous and exotic species could therefore be pointing to more complex dynamics at these

sites. This could perhaps result from more similar numbers of introduced and exotic species (Table 1.1) when compared to the intensive pasture sites. These sites are also in greater proximity to forest sites, which could impact how species use these sites. For example, perhaps some species, particularly indigenous species, use the semi-natural pasture to connect different fragmented forest habitats.

Metabolic rate distribution of individuals

Across land use, we find that the MRDIs are not particularly well described by the METE prediction. The Data section in Xiao et al. (2015) discusses why we should not necessarily expect animals (rather than plants) to follow the METE predicted MRDI. Namely, body sizes of animals belonging to the same species cluster around an intermediate value, and much larger or smaller species are rare (eg. Gouws et al. 2011). This means we are likely to end up with multimodal MRDIs (Thibault et al. 2011) rather than the monotonically decreasing form predicted by METE, which always predicts that the smallest size class is the most abundant.

A further reason that we may not necessarily expect these distributions to fit well is the number of approximations required to obtain these distributions. We used scaling relationships to covert from body length to body mass, and then approximated the intraspecific body mass distributions to be normal, and then further used a scaling relationship to covert to metabolic rate.

Despite these concerns and the number of approximations, we can still quantify which land use is the most well described by METE, and we see a similar relationship between land use and goodness of fit when compared with the SAD and SAR.

One direction we would expect the distribution to be skewed is that only adults are included in the metabolic rate distribution. This could result in missing the lower end of the unscaled MRDI. However, because METE predicts relative metabolic rate (scaled so that the smallest organism is $\varepsilon = 1$), this will result in over predicting the metabolic rate of the individuals with the greatest metabolic rate. We see this in Fig. 1.4, in that at low rank the METE predictions are consistently too high.

Another possible cause is that the Azorean arthropods are not saturating the ecological niches for large species. There is evidence that the indigenous species pool is unsaturated (Borges and Hortal 2009; Triantis et al. 2012; Rigal et al. 2013), and that specifically in the native forest sites the most successful introduced species were large bodied relative to indigenous species (Rigal et al. 2018). Though we note that if introduced species are filling these niches, then the species pool when combined may not have vacant niche space.

At the pasture sites, we consistently under predict the low rank, high metabolic rate individuals, and over predict the high rank, low metabolic rate individuals, resulting in a pattern where the residuals have positive slope (Fig. 1.4). These patterns are indicative of a large number of individuals with similar metabolic rate (see Appendix A.6 for plots of each site). As discussed in relation to the SAD, these sites have a few highly abundant, small bodied spider species. These species have comparatively low metabolic rate, and the

variation in metabolic rate within a species is smaller than the variation across species. We therefore end up with long lines of positive slope in the residuals as the METE prediction slopes downward over rank but the empirical MRDI remains roughly constant. We see this especially at intermediate and low ranks as these species have low metabolic rate. Thus this pattern is also likely driven by a few highly abundant species.

Species-area relationship

The mean least squares comparisons for the SAR in Figs. 1.1 and A.9 are noticeably different from those for the SAD and MRDI. Again here we find that the semi-natural pasture is the worst fit by METE, but it is not as dramatic as in other cases and is much closer to both forest sites. Additionally, we find that the intensive pasture is the most well fit by METE. However, the mean least squares is not the only goodness of fit metric. Particularly in the case of the pasture sites here, we see clear patterns in the residuals in Figs. 1.5 and A.10 that METE under predicts z and correspondingly over predicts the number of species at small scales. This is in line with our analysis of the SAD at the pasture sites, in that these sites have more high abundance species compared to the METE prediction (ie. the SAD is more hollow than the METE predicted log series). Overall, even though the mean least squares is smaller at the intensive pasture site, the direction of the difference is more biased.

In general, the pasture sites correspond to larger N_0/S_0 than the forest sites (see Fig. 1.5 and Table 1.1). When using D as a scale variable, this means that the pastures are testing a different scale compared to the forest sites. We see this in the residuals in Fig. 1.6, where the METE prediction for z is noticeably lower than the data points starting around $\log(N_0/S_0) \approx 2$, which is also where most of the pasture data points are clustered. This could indicate that the failure of METE to accurately predict the SAR is coming more from the underlying prediction of a log series abundance distribution, rather than from the spatial SSAD prediction, as the log series prediction for the pasture sites under predicts the abundance of the most abundant species.

Summary

Across METE predictions, the forest habitats are better predicted by METE than the semi-natural pasture habitat. The intensive pasture is intermediately well fit for the SAD and MRDI, and better fit for the SAR, though the residuals are not normally distributed.

For the forest sites, the native forest has very little human management but may be subject to the spread of invasive plants, and the exotic forest is subject to some human management and is in close proximity to the pastures. The deviations from METE in these sites are comparatively small and there are less noticeable trends in the residuals, though as with all sites METE over predicts the metabolic rate of the highest metabolic rate individuals.

The pasture sites are characterized by a few very abundant species, which is consistent with the abundance of several small bodied introduced spiders. In the SAD, METE consistently under predicts the abundance of the most abundant species, as well as the number of

singletons. In the MRDI, these very abundant species appear in the residuals as long lines of positive slope, as the variation in metabolic rate within a species is relatively small and the METE prediction falls off more rapidly than the empirical distribution.

The semi-natural pasture is particularly poorly described by METE across metrics. This is also the only land use with a large difference in goodness of fit between indigenous and introduced species, and the poor fit for the SAD appears to be driven by introduced species. Interpreting the deviation from METE as disturbance, this means that the semi-natural pasture is in some sense more disturbed than the intensive pasture. This could be due to complex interactions between indigenous and introduced species, particularly because of the proximity to other land uses, or because of the varying levels of management over the course of a year. It could also mean that the arthropod communities at the intensive pasture sites are more well adapted to the high level of management intensity, perhaps resulting in the large number of introduced species present at these sites.

In other studies of METE, disturbance is often linked to rapid change in state variables (Newman et al. 2020; Franzman et al. 2021; Harte et al. 2021). The dynamics are then out of steady state, and the state variables alone are not sufficient to describe the macroecological patterns. For example, Franzman et al. (2021) analyzed the change in state variables over time in a declining alpine meadow and found that macroecological patterns moved away from METE predictions over a six year period of observation. Here, we instead analyze how land use affects deviation from METE predictions assuming that this deviation is relatively static in time at any given land use. Assuming that disturbance is connected to changing state variables, we could interpret the poor fit of the semi-natural pasture as indicating that N_0, E_0 and/or S_0 are not constant on ecological time scales. We could test this hypothesis with time resolved data of arthropod composition. For example, connecting to our previous hypothesis, we might expect the state variables to change with management intensity over the year. It also may be the case that disturbance is more general and cannot always be characterized by changing state variables, and may depend on additional factors such as the rate of migration in and out of the ecosystem rather than just the net difference.

Analyzing the deviation from METE predictions across land use has provided us with useful information about how land use and related disturbance is affecting macroecological patterns in Azorean arthropods. While we initially expected the intensive pasture sites to be the most poorly fit by METE, this analysis points to the semi-natural pasture as the land use where arthropod communities are the most out of steady state. We were additionally able to interpret the deviations from METE predictions ecologically. We expect this type of comparison between METE predictions and ecosystems under land management disturbance to be helpful in identifying how land use affects macroecological patterns across other habitats and taxa.

Acknowledgments

I would like to thank Paulo A.V. Borges for his expertise on the Azorean system and for sharing the data, and Tom Matthews for his collaboration and work calculating metabolic rates from body mass data. I would also like to thank John Harte for his guidance, and Rosemary Gillespie and Kaito Umemura for their feedback.

Chapter 2

Relating the strength of density dependence and the spatial distribution of individuals

Published in *Frontiers in Ecology and Evolution* as Brush and Harte (2021).

Abstract

Spatial patterns in ecology contain useful information about underlying mechanisms and processes. Although there are many summary statistics used to quantify these spatial patterns, there are far fewer models that directly link explicit ecological mechanisms to observed patterns easily derived from available data. We present a model of intraspecific spatial aggregation that quantitatively relates static spatial patterning to negative density dependence. Individuals are placed according to the colonization rule consistent with the Maximum Entropy Theory of Ecology (METE), and die with probability proportional to their abundance raised to a power α, a parameter indicating the degree of density dependence. This model can therefore be interpreted as a hybridization of MaxEnt and mechanism. Our model shows quantitatively and generally that increasing density dependence randomizes spatial patterning. $\alpha = 1$ recovers the strongly aggregated METE distribution that is consistent with many ecosystems empirically, and as $\alpha \to 2$ our prediction approaches the binomial distribution consistent with random placement. For $1 < \alpha < 2$, our model predicts more aggregation than random placement but less than METE. We additionally relate our mechanistic parameter α to the statistical aggregation parameter k in the negative binomial distribution, giving it an ecological interpretation in the context of density dependence. We use our model to analyze two contrasting datasets, a 50 ha tropical forest and a 64 m^2 serpentine grassland plot. For each dataset, we infer α for individual species as well as a community α parameter. We find that α is generally larger in the tightly packed forest than the sparse grassland, and the degree of density dependence increases at smaller scales. These results are consistent with

current understanding in both ecosystems, and we infer this underlying density dependence using only empirical spatial patterns. Our model can easily be applied to other datasets where spatially explicit data are available.

2.1 Introduction

Spatial patterns in ecology have been studied extensively (e.g. Wiegand and Moloney 2013; Diggle 2014), and contain useful information about what processes shape ecosystems (Law et al. 2009; Brown et al. 2011; Münkemüller et al. 2020). Quantitative understanding of these patterns can therefore be used to infer the importance of various mechanisms, and illuminate underlying processes (Levin 1992; Rosenzweig 1995; Brown et al. 2016). Additionally, models of spatial patterns allow us to better predict ecosystem response to natural and anthropogenic disturbances (Thomas et al. 2004; Newman et al. 2020), are critical in understanding the well studied species-area relationship (Arrhenius 1921; Plotkin et al. 2000; Drakare et al. 2006; Harte and Kitzes 2015), and have applications in reserve designs and conservation (Kitzes and Shirley 2016).

A common approach to quantifying these patterns is the use of various summary statistics (Wiegand et al. 2013), which have been shown to be able to distinguish different ecological mechanisms (Brown et al. 2016). Here we take a slightly different approach and directly model the impact of an important mechanism in population dynamics: intraspecific negative density dependence. We focus on the effects of this ecological mechanism on spatial patterning.

More specifically, we consider the effects of intraspecific negative density dependence on the spatially explicit species-level abundance distribution. This distribution, $\Pi(n|A, A_0, n_0)$, is defined as the probability that if a species has n_0 individuals in a plot of area A_0, then it has n individuals in a randomly selected subplot of area A. In this analysis, we will focus on this distribution in bisected plots where $A = A_0/2$. Studying bisections is well motivated theoretically as it often leads to simpler expressions which can be easily compared across models. Here it keeps our model analytically tractable and facilitates comparison to empirical data. We note limitations to this approach in the Discussion.

One prediction of the function $\Pi(n|n_0)$ comes from the Maximum Entropy Theory of Ecology (METE), which successfully and simultaneously predicts many macroecological patterns (Harte 2011; Harte and Newman 2014) across a wide range of spatial scales, taxa, and habitats (White et al. 2012; Xiao et al. 2015). METE predicts very strong spatial aggregation, which is consistent with many observed ecosystems, and obtains the functional form of $\Pi(n)$ by maximizing entropy while constraining the mean number of individuals in a subplot. However, the same functional form can be obtained using a colonization rule, which is the approach we will use in our model.

Colonization rules assign spatial locations to new individuals based on the location of existing individuals. Chapter 4.1.2 in Harte (2011) shows that using the Laplace rule of succession as a colonization rule results in the same geometric distribution for $\Pi(n)$ that

METE predicts. Because METE agrees well with empirical data in many cases, we will use this colonization rule in our model. Occasionally, however, we see that the empirical degree of aggregation is less than the METE prediction (Conlisk et al. 2012; McGlinn et al. 2015). To study this, Conlisk et al. (2007) added an extra parameter to the relevant colonization rule that allows $\Pi(n)$ to vary, but it has no mechanistic interpretation and is used only as a free fit parameter.

We derive a new model that uses the colonization rule consistent with METE and adds a density dependent death rule. This means our model can be viewed as a density dependent extension of METE, and in that sense hybridizes MaxEnt and mechanism. Our model introduces a parameter α which quantifies the degree of intraspecific negative density dependence. This parameter can be fit to empirical spatial data to predict the strength of underlying density dependence. However, as with all models inferring process from pattern, there are many underlying mechanisms that lead to similar spatial patterns (Vellend 2016; Leibold and Chase 2018), and we cannot definitively attribute any pattern to a single process.

More generally, our model predicts a more random spatial arrangement with stronger negative density dependence and more spatial aggregation with weaker density dependence. While empirically there is an apparent qualitative relationship between species density and aggregation (Condit et al. 2000; Bagchi et al. 2011; Comita et al. 2014), our aim here is to establish a general quantitative statement relating density dependence and spatial aggregation.

2.2 Methods

In this section, we review the Maximum Entropy Theory of Ecology (METE) and its prediction for the species-level abundance distribution, $\Pi(n)$. We then contrast this prediction of strong aggregation to the well known random placement model (Coleman 1981), which predicts no spatial aggregation. Given that most species are aggregated (He and Gaston 2000; Kitzes 2019), but not all are as aggregated as predicted by METE (Conlisk et al. 2012), the aggregation of most species should fall somewhere between these two predictions for $\Pi(n)$.

We then introduce a density dependent death rule to combine with the colonization rule consistent with METE, and derive the resulting $\Pi(n)$ distribution. This derivation assumes a steady state between deaths and new individuals in a single species, but our results should hold if this assumption is relaxed (see Discussion).

Finally, we discuss the techniques used to compare our predicted distribution to data, and describe the datasets used in our analysis.

Relevant code for the resulting $\Pi(n)$ distribution and data analysis is available at https://github.com/micbru/density_dependence_public.

The Maximum Entropy Theory of Ecology (METE)

In METE, the $\Pi(n)$ function is given by maximizing the information entropy of the $\Pi(n)$ distribution given the following constraint (Harte et al. 2008; Harte 2011, Chapter 7.4):

$$\sum_{n=0}^{n_0} n\Pi(n|A, A_0, n_0) = \frac{n_0 A}{A_0}. \tag{2.1}$$

This leads to the following distribution

$$\Pi(n|A, A_0, n_0) = \frac{e^{-\lambda_\Pi n}}{Z_\Pi} \tag{2.2}$$

where Z_Π is a normalization factor, and λ_Π is the Lagrange multiplier determined by the constraint condition.

In the case of a bisection, $A = A_0/2$ and the Π function simplifies to

$$\Pi(n|n_0) = \frac{1}{n_0 + 1}, \tag{2.3}$$

which is independent of n. This means that given n_0 individuals, any arrangement of them on the two sides of a bisected plot or quadrat is just as likely as any other. In other words, this is equivalent to equal probability for each unique spatial arrangement of unlabeled individuals (Haegeman et al. 2010).

Ecologically this prediction translates to very strong spatial aggregation, as individuals are equally as likely to all be on one side of the bisection as to be evenly divided on each half. This is in agreement with many datasets (Harte et al. 2008; Harte 2011, Chapter 8.3) but fails in others, where the theory over-predicts aggregation (Conlisk et al. 2007; McGlinn et al. 2015). This empirical agreement is why we choose the METE distribution as our starting point.

The prediction from METE is equivalent to the distribution obtained from using the Laplace rule of succession as a colonization rule (Harte 2011, Chapter 4.1.2). This rule states that in a colonization process, the probability of placing an individual on one side of the bisected area is roughly proportional to the fraction of individuals already there. This "rich get richer" effect results in strong spatial aggregation. The probability for placing an individual on the left half of a bisected plot with n_L individuals on the left and n_R individuals on the right is

$$p_L = \frac{n_L + 1}{n_L + n_R + 2}.$$

To make our notation consistent with that above, let the number on the left be n and the total number to be n_0. The probabilities of a new individual arriving on the left or on the right are then:

$$\begin{aligned} p_L(n|n_0) &= \frac{n+1}{n_0+2}, \\ p_R(n|n_0) &= \frac{n_0 - n + 1}{n_0 + 2}. \end{aligned} \tag{2.4}$$

If we place n_0 individuals using this rule, the resulting probability distribution is given by Eq. 2.3.

Random placement

Another model for spatial ecology, perhaps the simplest, is the random placement model (Coleman 1981). Instead of the placement rules in Eq. 2.4, each individual is placed randomly. In a bisected plot this means each individual has a 50 percent chance of being placed on either side, $p_L = p_R = 0.5$. Placing n_0 individuals this way gives the binomial distribution

$$\Pi_{\text{RP}}(n|n_0) = \binom{n_0}{n}\left(\frac{1}{2}\right)^{n_0} \tag{2.5}$$

which, if n_0 is large, means we are very likely to have roughly half the individuals on each side. This is equivalent to having no spatial aggregation; there is no preference for any new individual to stay close to any previous individual as each placement is a random coin flip.

Deriving the $\Pi(n)$ distribution with a density dependent death rule

We now introduce an intraspecific density dependent death rule in addition to the METE colonization rule in Eq. 2.4. To allow for general density dependence, we set the death rate proportional to n^α. The parameter α determines the strength of the density dependence, and can be inferred from the data. Density dependence may result from resource limitation, or some other mechanism (e.g. the Janzen-Connell effect (Janzen 1970; Connell 1971)).

In the case of a bisected plot, each death must be on the left or right. Thus, given that we have one death in a species, the probabilities that the death is on the left, $p_{D,L}$, or on the right, $p_{D,R}$, are

$$\begin{aligned} p_{D,L}(n|n_0) &= \frac{n^\alpha}{n^\alpha + (n_0 - n)^\alpha} \\ p_{D,R}(n|n_0) &= \frac{(n_0 - n)^\alpha}{n^\alpha + (n_0 - n)^\alpha}. \end{aligned} \tag{2.6}$$

Now that we have the colonization and death rules (Eqs. 2.4 and 2.6), we can derive the general $\Pi_\alpha(n|n_0)$ for bisections. We will assume the population size of the species is constant and step the model forward over time, where at each step in the model we will have one death followed by the placement of one new individual within a species. Each placement can be interpreted ecologically as a birth or as the immigration of an individual from the same species. We can then solve for the resulting steady state distribution where we reach an equilibrium in the spatial pattern.

There are several approaches for deriving the steady state solution for such a system. Here, we equate the rates leaving and entering any individual state $\Pi_\alpha(n|n_0)$. We take the probability that we start with n individuals on the left, one on the right dies and then one is

placed on the left resulting in $n+1$ individuals on the left, and equate that to the probability that we have $n+1$ individuals on the left, one on the left dies and then one is placed on the right resulting in n individuals on the left. We could have equivalently done the same thing with n and $n-1$. Equating these rates using the probabilities in Eq. 2.4 and Eq. 2.6 leads to a recursion relation. Solving it gives a general stationary solution for Π_α with a given n_0 and α:

$$\Pi_\alpha(n|n_0) = \frac{n^\alpha + (n_0 - n)^\alpha}{C(n_0, \alpha) n_0^\alpha} \binom{n_0}{n}^{\alpha-1} \tag{2.7}$$

where $C(n_0, \alpha)$ is the overall normalization that does not have a closed analytic form. In the case that $n_0(\alpha - 1)$ is large, an approximate form for the normalization is

$$C = \frac{2^{n_0(\alpha-1)} \pi n_0}{\sqrt{\alpha - 1}} \left(\frac{1}{2\pi n_0}\right)^{\alpha/2}. \tag{2.8}$$

See Supplementary Material B.1 for the details of this derivation.

If $\alpha = 2$, we can solve for the normalization explicitly to get

$$\Pi_{\alpha=2}(n|n_0) = \frac{n^2 + (n_0 - n)^2}{2^{n_0-1} n_0(n_0+1)} \binom{n_0}{n}, \tag{2.9}$$

and if $\alpha = 1$, we recover the METE prediction $\Pi_{\alpha=1}(n|n_0) = \frac{1}{n_0+1}$.

Comparing to data

Inferring the degree of intraspecific density dependence in empirical data requires obtaining a value of α consistent with the data. Bisection predictions can be compared to data by rank ordering the fraction of individuals present in one half of the plot for each species (e.g. Harte (2011)). This method, however, ignores species abundance and does not account for the likelihood of individual data points. This can lead to incorrect conclusions about which model is preferred (see Supplementary Material B.2).

We instead find the maximum likelihood α given the data, where we minimize the sum of the negative logs of the probabilities given data points n_i and $n_{0,i}$, where i labels each quadrat for a given species. Inferring α using this method gives us the values that are the most consistent with the data, even if they may not look like they agree with the rank ordered fractions (Supplementary Material Fig. B.1 and Table B.1).

The statistical error in estimating α this way goes as $1/\sqrt{p}$ where p is the sample size (see Supplementary Material B.3). We can also get some idea of error from the maximum likelihood estimate itself by considering the width of the likelihood distribution, however for Fig. 2.3 and 2.4, we do not include these error bars as they are smaller than the data points.

For determining α for individual species, we will require multiple bisections and the sample size p will be roughly the number of cell pairings, $p \approx 2^{b-1}$, where b is the number of bisections. There will be fewer data points in practice as some cell pairings will be empty.

We can also define a community α, assuming each species follows the same death rule with identical α. In this case, we will have a larger sample size. For a single bisection, we will have a sample size p equal to the number of species, $p = S_0$. For multiple bisections where we consider the species on aggregate, the sample size will be roughly equal to the number of species multiplied by the number of cells, $p \approx S_0 2^{b-1}$. Again, the equality is not exact as not all cell pairings beyond the first bisection will have all of the species present at the single bisection level.

In both the case of the species-level and community-level α, we will bisect single plots more than once (into quadrants, then into 8 cells, etc.) when comparing the model to data. In our analysis, we begin by bisecting the plot in half in one direction, then bisecting each of the resulting plots in the opposite direction. We alternate this bisection pattern until we have 2^b cells. We can then combine adjacent cells (either left/right or up/down) as if they were single plots with abundance $n_{0,i}$, where i will index the plots and range from 1 to 2^{b-1}. We then choose the abundance on one half to be n_i. This method gives us 2^{b-1} points.

Additionally, in this analysis we will only consider species that could have at least one individual per bisection ($n_0 > 2^{b-1}$). The smallest scale we consider in our datasets is $b = 8$, so when we bisect the plot more than once we will only consider species with more than 128 individuals. This restriction ensures that we do not have too many plots with only a few individuals present. If n_0 is very small, $\Pi_\alpha(n)$ is not particularly sensitive to α and it becomes very difficult to reliably infer α from the data. For $n_0 \leq 2$, $\Pi_\alpha(n)$ does not depend on α.

Data used

We will compare our results to two contrasting datasets. First, we will use data from a sparse Californian serpentine grassland site (Green et al. 2003; Green et al. 2019) at the McLaughlin University of California Natural Reserve censused in 1998. This is a 64 m^2 plot divided into 256 cells with 24 species and 37 182 individuals. There are 10 species with abundance greater than 128 individuals that constitute 36 783 individuals.

Second, we will use data from Barro Colorado Island (BCI) in Panama (Condit 1998; Hubbell et al. 1999; Hubbell et al. 2005; Condit et al. 2019), a 50 ha plot in a moist tropical forest. We will work with the 2005 census and consider plants with a diameter at breast height (dbh) greater than 10 cm. This dataset has 229 species and 20 852 individuals, and 40 species with abundance greater than 128 individuals that constitute 15 960 individuals.

2.3 Results

Comparison to METE and random placement

Figure 2.1 compares the bisection predictions for $\Pi(n)$ from METE, random placement, and our density dependent model for various α, at $n_0 = 10$ and 50. In general, our model

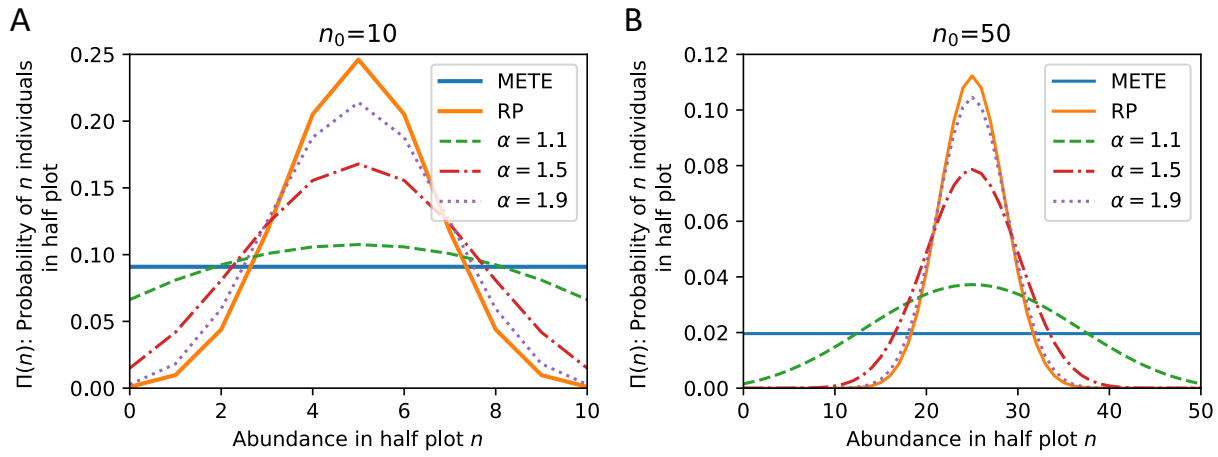


Figure 2.1: Comparison of the bisection probability distributions $\Pi(n)$ from METE, random placement (RP), and our density dependent model with varying α at **(A)** $n_0 = 10$ and **(B)** $n_0 = 50$. At $\alpha = 1$, our model corresponds exactly to METE. At larger n_0, $\alpha \to 2$ approaches the random placement distribution. Our model varies continuously between METE and random placement for $1 < \alpha < 2$.

predicts that increasing negative density dependence (larger α) leads to more random spatial patterning, and less density dependence (smaller α) leads to stronger aggregation.

We can relate our distribution directly to both the METE and random placement distributions for different values of α. $\alpha = 1$ corresponds exactly to the METE solution, which makes sense given that the placement and death rules are both linear in n. As $\alpha \to 2$, our distribution approaches the random placement prediction if n_0 is large enough (Supplementary Material B.4 shows this result analytically). For $1 < \alpha < 2$, we vary continuously between METE and random placement. We can make the distribution even more spatially aggregated than METE with $\alpha < 1$ and even less than random placement (overdispersed) with $\alpha > 2$.

We can also relate this distribution to the commonly used conditional negative binomial distribution (Bliss and Fisher 1953; He and Gaston 2000; He and Gaston 2003; Green and Plotkin 2007) in the limit of large n_0, assuming that matching the peak of the distributions is a good approximation for the entire distribution. In that limit, the aggregation parameter

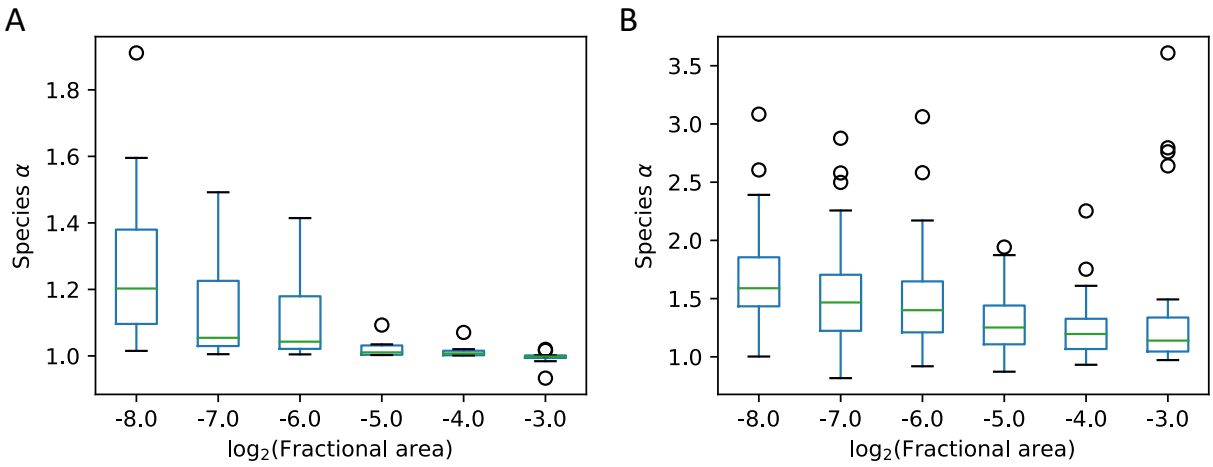


Figure 2.2: Boxplots for α among the species at different scales at both sites, where **(A)** shows 10 species from the serpentine dataset, and **(B)** shows 40 species from the BCI dataset. In both cases at smaller scales α is larger, and we see a relatively large spread in α across species at the same scale. The boxplots show boxes from quartile 1 (Q1) to quartile 3 (Q3) with a line at the median. The whiskers extend to 1.5×(Q3-Q1). The remaining points are plotted as individual circles.

k is approximately related to the density dependent parameter α by

$$k \approx \frac{n_0}{2}\left(\frac{\alpha - 1}{2 - \alpha}\right) + 1. \tag{2.10}$$

Note that this approximation holds for $1 \leq \alpha \leq 2$, which should be the ecologically relevant range for most species as most species will be more aggregated than random placement, and less aggregated than METE. This also allows the aggregation parameter k to be interpreted mechanistically as the degree of density dependence, in that higher k corresponds to higher α and greater density dependence. See Supplementary Material B.5 for the derivation.

Individual species

Since the Π function is defined on the species level, we can consider each species separately and find the maximum likelihood α for each. To do this we have to go beyond the first bisection to get multiple data points for the same species at smaller scales.

For the serpentine data, we exclude *Eriogonum nudum* from the following figures as an outlier (see Discussion). This leaves 9 species with abundance greater than 128 individuals.

Figure 2.2 shows the distribution of α values among the species at each scale, for both the serpentine and BCI data. The median α increases at smaller scales for both datasets, and is higher overall at the BCI dataset, even though the absolute scale is much larger. The spread in α is quite large, but this variation is expected considering the small number of data points, especially for rarer species. Most species have an α between 1 and 2, which is somewhere between the aggregation predicted by METE and random placement.

Community α

We can instead treat α as a community parameter, using each species as a single data point to recover a community α. Figure 2.3 shows the direct comparison between our model prediction and the serpentine and BCI datasets at the single bisection level. Each data point is the observed fraction of individuals in one half of the plot versus the species abundance. The curves in this figure show the 95% contour intervals for the $\Pi(n|n_0)$ distributions predicted by METE, random placement, and our density dependent model with the maximum likelihood α value. We can see that with increasing n_0, the random placement model narrows quickly to having most of its probability weight around 0.5, whereas the METE contours are very wide.

At the single bisection level, the maximum likelihood result for the serpentine dataset is nearly indistinguishable from $\alpha = 1$, so the confidence interval curves on the plot for METE and the density dependent model overlap for most n_0. For the BCI data, the maximum likelihood value is $\alpha = 1.12$, slightly larger than 1. In this case, where $1 < \alpha < 2$, we see the width of the predicted distribution is between METE and random placement. The likelihoods for each of the models are shown in Table 2.1.

Serpentine		**BCI**	
Model	Log-likelihood	Model	Log-likelihood
METE	-114.8	METE	-729
RP	-5188.6	RP	-963
$\alpha = 1.0003$	-114.6	$\alpha = 1.12$	-660

Table 2.1: Log-likelihood values for the three different models, with α as a community parameter. We can compare our model to METE using the deviance in log-likelihood and obtain a p-value. The deviance is defined as twice the difference in log-likelihood. For the serpentine dataset, the deviance is 0.6 which corresponds to a p-value of 0.45. For the BCI dataset, the deviance is 138 which corresponds to a p-value of $< 10^{-30}$.

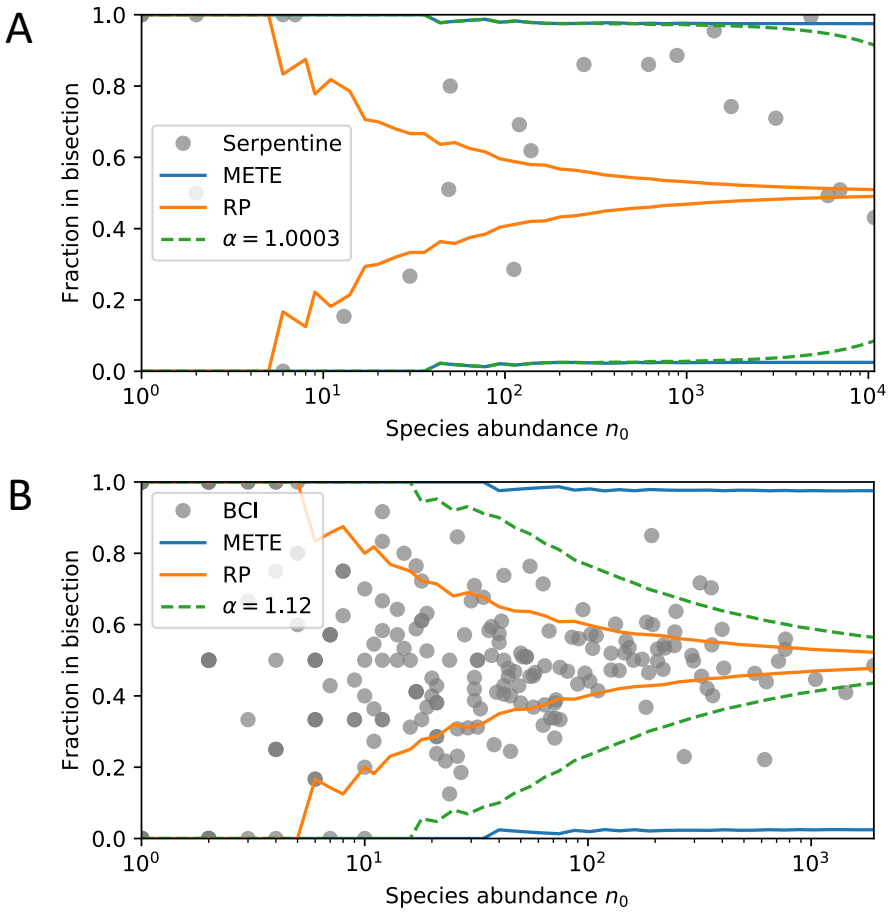


Figure 2.3: 95% contour intervals for the predicted bisection probability distributions $\Pi(n|n_0)$ from METE, random placement, and our density dependent model with maximum likelihood community α, and bisection data for each species in **(A)** the serpentine dataset, and **(B)** the BCI dataset. The data is plotted for each species, where the y-axis is the fraction in one half plot and the x-axis is the total species abundance in that plot. The contours are calculated at each n_0. For our density dependent model with a community α, $\alpha = 1.0003$ maximizes the log-likelihood for the serpentine dataset, and $\alpha = 1.12$ maximizes log-likelihood for BCI dataset.

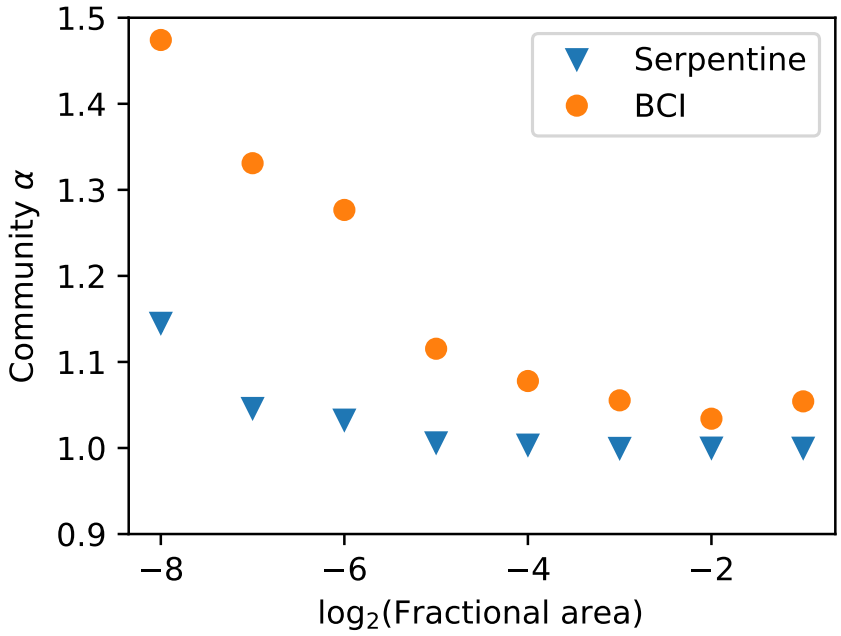


Figure 2.4: Community α scaling with area for species with abundance $n_0 > 128$. The density dependence again increases at smaller scales and the trend is similar to the single species analysis. The serpentine dataset has 36 783 individuals and the BCI dataset has 15 960 individuals.

Scale dependence in community α

Going beyond the first bisection allows us to see how α varies depending on the scale of our plot. Figure 2.4 shows how α scales with fractional area for both the serpentine and BCI plots. Density dependence increases at smaller scales in both datasets. The trend in community α across scales is similar to the median α in the single species analysis, though the median α is in general slightly larger than the community α. Note that here we restrict our analysis to species with $n_0 > 128$ for all scales so that we are including the same species across scales.

2.4 Discussion

Our model establishes a quantitative relationship between the spatially explicit distribution $\Pi(n)$ and the parameter α, which measures the strength of negative density dependence. This can be seen in Fig. 2.1, where the $\Pi_\alpha(n)$ distribution flattens with smaller α, indicating greater aggregation, and broadens as α increases. Importantly, the parameter α has a direct interpretation as quantifying the strength of negative density dependence. Further, our

relationship in Eq. 2.10 allows us to interpret the parameter k in the negative binomial distribution in the same intuitive way.

Comparing species and community α

In our analysis, we consider α both as a separate parameter for each species (as in Fig. 2.2), and as a community parameter where each species has the same α (as in Fig. 2.3 and Fig. 2.4). The community α is harder to interpret ecologically, but we include it to allow for comparisons with models with community level aggregation parameters (e.g. Conlisk et al. (2007) and Volkov et al. (2005)). To analyze and compare the accuracy of the species-level α and the community α, we considered the Akaike Information Criterion (AIC) in both cases across scales (Table 2.2). This was calculated for single species as the negative log-likelihood summed over each species with the number of parameters equal to the number of species, whereas for the community α there was only a single α parameter. For both serpentine and BCI at all scales considered, we find that the AIC is lower with species-level α compared to a single community α, despite the inclusion of 9 more parameters in the case of the serpentine data and 228 more parameters in the case of the BCI data. We therefore conclude that a separate α for each species describes the data better than a single community α.

	Scale (A/A_0)	2^{-8}	2^{-7}	2^{-6}	2^{-5}	2^{-4}	2^{-3}
Serpentine	Species α, AIC	474	769	1294	1931	3208	4881
	Community α, AIC	485	777	1321	2079	3420	5182
BCI	Species α, AIC	10133	7541	5079	3409	2109	1271
	Community α, AIC	10207	7621	5148	3522	2171	1307

Table 2.2: Comparison of the Akaike Information Criterion (AIC) for α defined at the individual species level and at the community level in both the serpentine and BCI data and across scales. At the individuals species level, the number of parameters is equal to the number of species, whereas at the community level there is only a single parameter. The AIC is lower at the species level in all cases.

Comparing serpentine and BCI

We use our model to directly compare our results between our two contrasting datasets, serpentine and BCI. Because the serpentine site was very sparse, whereas the BCI forest is tightly packed, we expect higher α values and greater density dependence at BCI than at the serpentine site. This is consistent with our inferred values of α at both the individual species level and at the community level.

Another difference between the serpentine and BCI sites is how well other macroecological distributions agree with METE. METE well describes other patterns at the serpentine

site, and does less well at explaining the BCI data. Given that $\alpha = 1$ corresponds to the METE prediction for $\Pi(n)$, we might expect that ecosystems well described by other METE predictions will have $\alpha \approx 1$, as these systems will generally be consistent with METE. This is consistent with our analysis here as the median and community αs for the serpentine data are approximately 1 at the largest scale, whereas at BCI the median and community αs are larger than 1.

Because METE predictions seem to hold for relatively static and undisturbed ecosystems (Newman et al. 2020), this suggests interpreting an increase in density dependence away from METE ($\alpha > 1$) as a kind of ecological disturbance. A biological example of strong density dependent mortality as a result of disturbance could be the self-thinning of trees in forest recovery from wildfire, such as bishop pines in coastal California (Harvey et al. 2014). This interpretation is in line with the recently proposed DynaMETE theory (Harte et al. 2021), which models specific mechanistic disturbances away from METE to predict macroecological patterns.

Scaling

Our scaling results in both Figs. 2.2 and. 2.4 make ecological sense. We expect that at smaller scales, the density dependence would be larger as individuals compete more for resources at that scale. At large scales, we expect α to be close to 1 as the individuals do not compete over large distances. This means that the spatial distributions look more aggregated on large scales than on small scales as the individuals within species broadly group together, but repel each other at small scales. We see this trend at the individual level in Fig. 2.2 as the medians increase at smaller scales, and for the community α in Fig. 2.4.

Our repeated bisection analysis also indicates at which scale density dependence becomes important. This will appear as a shoulder in the data where α moves away from ≈ 1. We could do this for individual species by tracing α and looking for a shoulder in Fig. 2.2, but here we will look at the community results in order to compare to Conlisk et al. (2007) and Volkov et al. (2005). Looking at Fig. 2.4, we find that the shoulder in absolute scale corresponds to $< 0.5\text{m}^2$ for the serpentine plot and < 1.6 ha for the BCI dataset. This again makes sense given that the serpentine grassland is much more sparse than the BCI forest.

We first compare to Conlisk et al. (2007), who introduce a fit parameter ϕ that modifies the colonization rule Eq. 2.4 and allows the Π distribution to vary continuously between random placement and METE. They compare their estimated community ϕ parameter to both the serpentine and BCI data in their Fig. 6. For the serpentine data, they find that at scales larger than around 0.5 m^2 (the 8th bisection), ϕ approaches 0.5, which corresponds to the METE prediction. At scales around 0.5 m^2 or smaller, $\phi \approx 0.25$, where $\phi = 0$ corresponds to random placement. This is consistent with our scaling results in Fig. 2.4. For BCI, they find that at all but the largest scales $\phi \approx 0.25$. Our result that α is larger at BCI than at the serpentine site across scales, which corresponds to less spatial aggregation, is not consistent with their findings. We believe this is due to a difference in how the data are analyzed.

In Conlisk et al. (2007), the species abundance n_0 is measured at the scale of the full plot, and the bisection prediction is recursively iterated to smaller scales (see their Theorem 2). Here, we treat each bisection at smaller scales separately. For example, after dividing the plot into 128 quadrats (8 bisections), we look at the species abundance in each individual quadrat without considering n_0 at the scale of the entire plot. In principle, we could conduct our analysis in the same way and anchor at the largest scale, though this would be difficult analytically, and our approach makes use of the empirical data available at each scale rather than only at the largest scale. Further, the method in Conlisk et al. (2007) depends implicitly on the chosen size of the overall study plot. This is not true in our analysis as a bisection studied at any scale does not depend on information at any other scale. In practice, this means that in our analysis there is no difference between studying species in a 1 m^2 subplot embedded in a 100 m^2 plot versus studying the same 1 m^2 plot independently. This difference in how n_0 is treated across scales could lead to different predictions for α (or ϕ).

We can further compare our results to Conlisk et al. (2007) by relating our α to ϕ, using their relationship between ϕ and k and our Eq. 2.10. This relationship depends on n_0, which may affect comparisons between these parameters across scales. Finally, an additional difference between our analyses is our different cutoff of $n_0 > 128$, and for BCI, dbh $>$ 100mm, however this does not explain all of the difference between our results. Supplementary Material B.6 derives an approximate relationship between α and ϕ, Supplementary Material Fig. B.4 uses that relationship to transform our Fig. 2.4 to a relationship in ϕ, and Supplementary Material Fig. B.5 shows how our result changes if we remove our abundance threshold. A takeaway from this comparison is that these scaling results depend at least in part on the choice of model and the data analysis methods.

Volkov et al. (2005) showed that intraspecfic and symmetric density dependence can explain different shapes for the species-abundance distribution. Their added parameter c is interpreted as a measure of the strength of symmetric density dependence, where $c = 0$ corresponds to no density dependence. This parameter is therefore similar to our community α in that all species have the same degree of negative density dependence. They then show at what density these effects become important in their Fig. 3. For BCI, they find $c = 1.80$, and the density dependent effects are visible for species with $n > 27$. To convert this to area, we need to look at scales of the total area divided by the abundance where density dependent effects become visible. Thus, this corresponds to density dependence entering at scales smaller than a fractional area of $1/27 = 1/2^{4.75}$, which is close to the same scale where we see α increase away from 1 in Fig. 2.4.

Across these results, we interpret increasing α at small scale as an increase in density dependence. However, at smaller spatial scales where there are fewer individuals it becomes more difficult to distinguish between different patterns of aggregation. In particular, when $n_0 A/A_0 << 1$, it is difficult to determine if the empirical pattern is due to noise or a specific clustering process (Harte 2011, pg 63). This sampling effect should be small here, as even at the smallest scale the median $n_0 A/A_0$ is greater than 1 for both datasets (Supplementary Material B.7).

Trends for individual species at BCI

At the individual species level at BCI, we find overall that most species at all scales are more aggregated than random ($\alpha < 2$ in Fig. 2.2). This is consistent with results from Condit et al. (2000). We also find that species tend to be more aggregated at large scales than at small scales (median $\alpha > 1$ at small scales and $\alpha \approx 1$ at large scales in Fig. 2.2), which makes sense as we expect some species to only be present in certain areas of the plot.

More broadly, we might expect to find trends in inferred density dependence with species abundance or size. More abundant species may be competing more for the same resources, or larger species may compete over larger distances. For example, Condit et al. (2000) find that both rarer species and smaller individuals tend to be more aggregated, however at a much smaller scale (within a 10 m radius). We looked for trends in abundance, mean dbh, and total energy for each species at BCI with $n_0 > 128$ across all scales considered (as in Fig. 2.2).

In terms of abundance (Supplementary Material Fig. B.7 and Table B.2), we do not find any species with high α and high abundance (no highly density dependent high abundance species), and we find that the variance in α decreases with abundance. We also find that at all scales except the two smallest, α decreases slightly with increasing log of abundance. Thus, we find that at larger scales, more abundant species are slightly more aggregated than less abundant species.

We find no evidence of a trend with species' mean dbh (Supplementary Material Fig. B.8 and Table B.3), though it is possible this trend is obscured by variance in individual size within a species, or that the range of mean dbh we considered (about $100 - 500$ mm) is too small to see its effect. Finally, we looked for an overall energy effect. Considering that the most abundant species tend to be smaller, it may be that density dependence depends on the total metabolic rate of a species. Plotting this relationship (Supplementary Material Fig. B.9 and Table B.4) again does not reveal a significant scaling relationship at all scales except one ($\log_2(A/A_0) = -6$).

A plausible mechanism for the observed density dependence at BCI is the Janzen-Connell effect (Janzen 1970; Connell 1971), whereby areas near parent trees are inhospitable for offspring, resulting in density dependence. Various studies (Harms et al. 2000; Carson et al. 2008; Comita et al. 2014) have observed this effect at BCI, which is consistent with our result that $\alpha > 1$ for most species at smaller scales there.

See Supplementary Material B.8 for more information on these trends.

Notable species

For individual species at the BCI dataset, *Gustavia superba* stood out with an average α of 1.001 across scales. This species is largely limited to 2 hectares of young secondary forest along the edge of the plot, (J. Wright, personal communication, 2019) making it look especially aggregated and resulting in a maximum likelihood α close to 1.

In the serpentine dataset, we excluded *Eriogonum nudum* as an outlier for part of our analysis. The maximum likelihood α was > 6 at the smallest scale and the maximum itself was very shallow. This species has a large canopy compared to the other grassland plants, and tends to be found far from other individuals. It makes sense that it would be overdispersed with $\alpha > 2$.

Implications for the species-area relationship

In METE, the spatial distribution is used together with the species-abundance distribution to predict the species-area relationship (Harte 2011, Chapter 7.5), and to upscale predictions of biodiversity (Harte and Kitzes 2015). These predictions should hold in ecosystems like the serpentine grassland analyzed here, as the observed species aggregation agrees with the METE prediction. However, different levels of aggregation will impact the species-area relationship. The impact of aggregation is discussed in Wilber et al. (2015). They find that increasing randomization decreases the predicted slope of the species-area relationship at the same scale, and therefore upscaling METE will overpredict species richness. In addition, they analyze the effect of variation in aggregation among species, which slightly decreases the slope at small scales and increases the slope at larger scales. This results in a species-area relationship that more closely resembles a power law. They also consider the effect of decreasing aggregation across scale, which results in a species-area relationship that no longer displays scale collapse. We observe both of these effects here.

Limitations and assumptions

As with all models inferring process from pattern, we can never be sure the pattern we observe can be completely attributed to the process we model. There are many different underlying processes that can lead to aggregation, including environmental filtering and dispersal limitation (Vellend 2016; Leibold and Chase 2018), and it is not possible for any one model to include every effect. Our empirical results here are consistent with our interpretation of α as a parameter that relates to the strength of intraspecific negative density dependence, however there are certainly other important mechanisms in these datasets. Regardless of our ability to infer process from pattern, our theoretical result that increasing density dependence increases spatial randomization holds.

Our model is also limited in that it only considers bisections, and it would be useful to extend it to be more general. There are many spatial arrangements that can not be accurately captured by dividing plots into bisection, and in general a single functional summary statistic does not completely describe the observed spatial pattern (Wiegand et al. 2013). For example, if we divide our plot into an m by m grid, and have one individual per cell, we would see exactly 0.5 as the fraction for each bisection. This result would be consistent with random placement with a large number of individuals, which does not well describe this exceptionally uniform arrangement. There could also be different degrees of spatial aggregation within a cell that we will not accurately capture with a bisection. Despite these

limitations, bisections are useful for understanding commonly observed macroscopic spatial patterns.

A conceptually simple extension to our model is to divide plots into quadrisections rather than bisections. The colonization and death rules then have three unknowns rather than one (the number of individuals in each quadrant, where the fourth is determined by constraining the sum to be n_0). This makes it hard to solve analytically, however we can simulate the birth-death process until it reaches steady state. We find no significant difference in our simulation compared to our prediction from two bisections, and find that a community $\alpha = 1.12$ is still consistent with the BCI data.

Because we consider the steady state solution in our model, we are assuming that the density dependence time scale is longer than the time scale of individual births or deaths. That is, α must not change too rapidly in time. This assumption is justified for many systems roughly in steady state with their environment and not undergoing rapid change (Newman et al. 2020).

Solving for the steady state solution also assumes that births and deaths are in balance. We assume here that there is a single death followed by a placement, however simulating two deaths followed by two placements gives a probability distribution consistent with our analytic prediction. We expect our result to hold with other numbers of deaths and placements. Assuming that births and deaths are in balance also implicitly assumes some amount of negative density dependence, and here α provides a quantitative measure of the degree of density dependence.

Another assumption in our model is the choice of colonization rule itself, though if we had chosen a different colonization rule many of our conclusions would remain the same. We use the colonization rule consistent with METE because of its good empirical agreement (Harte 2011, Chapter 8.3). This allows us to interpret the $\alpha = 1$ case as consistent with METE. This is useful as METE can be thought of as a null theory that holds in ecosystems that are undisturbed and relatively static (White et al. 2012; Xiao et al. 2015; Harte et al. 2017; Newman et al. 2020), and $\alpha \neq 1$ can be thought of as a density dependent correction, away from the MaxEnt distribution. In this sense, this model hybridizes MaxEnt and mechanism.

Instead, as an example, we could have chosen the colonization rule resulting in the random placement distribution. For a bisection this rule is just $p_L = p_R = 0.5$. In this case, $\alpha = 1$ would recover the binomial distribution, which we know does not well describe most spatial data (He and Gaston 2000; Condit et al. 2000), and so we cannot interpret $\alpha \neq 1$ as a density dependent correction. As another example, if we had chosen the more general colonization rule in (Conlisk et al. 2007) we would have two parameters to tune, making it difficulty to differentiate between colonization and death. In ecosystems where we suspect a different colonization rule may be in play, we could modify our theory appropriately. In any of these cases, our general results would remain largely unchanged.

Future work

One advantage of the bisection approach is that it can make predictions about inter-quadrat correlations. McGlinn et al. (2015) examined these correlations and compared empirical distance-decay relationships with the spatial predictions of METE ($\alpha = 1$ in this model). They found that the predicted distance-decay was much stronger than observed. We would expect the predicted distance-decay relationship to be weaker with $\alpha > 1$ in our model. Conlisk et al. (2007) note that $\phi > 0$ in their model produces more realistic looking distance-decay than random placement. Together, this means that with $1 < \alpha < 2$ our model should predict a more realistic shape for the distance-decay relationship compared to random placement, but with less steep of a slope than predicted by METE. However, Conlisk et al. (2007) also note that the analysis of these inter-quadrat correlations makes use of distance between cell pairs rather than physical distance, which limits the analysis (though note Ostling et al. (2004) provides a set of user rules to reduce this effect). This issue is also present in our model. Future comparisons to empirical distance-decay relationships could provide another method of estimating α and testing this framework.

Another advantage of our approach is that it only requires static spatial data. However, analyzing a single dataset over time could provide an interesting test of our interpretation of α as a measure of density dependence. This would be particularly appropriate with data where strong density dependent mortality is known to occur, for example a self-thinning forest recovering from wildfire (Harvey et al. 2014; Newman et al. 2020).

Finally, while our analysis here compares two contrasting datasets, future work could analyze more ecosystems to look for effects of habitat type, species richness, or average density.

Conclusion

Our model robustly predicts that increased intraspecific negative density dependence leads to more random spatial patterning, and establishes a quantitative relationship between the degree of density dependence described by the parameter α and spatial patterning described by the metric $\Pi(n)$. We predict that this result is general across ecosystems and taxonomic groups. We find that at all but the smallest scales, the serpentine grassland site is consistent with the absence of a density dependent correction and has the strong spatial aggregation predicted by METE. This is true for both the median individual species and at the community level. At the tropical forest site, our results indicate that negative density dependence is important: the median species α and the community α are both greater than 1 at even the largest scales. Both ecosystems show scaling of α consistent with its interpretation as the strength of negative density dependence. Median species α and community α are larger at smaller scales, and increase away from 1 at scales consistent with other analyses. Overall, our analysis of α is consistent with the interpretation of density dependence at both sites. Because this model uses only static spatial patterning, it can be applied in any ecosystem with spatially explicit data.

Acknowledgments

I would like to thank John Harte for his guidance throughout this work, as it was my first project in ecology. This material is based upon work is supported by the National Science Foundation under Grant No. DEB-1751380. I acknowledge the support of the Natural Sciences and Engineering Research Council of Canada (NSERC), [PGSD2-517114-2018]. John Harte thanks the Santa Fe Institute and the Rocky Mountain Biological Laboratory for their hospitality. We thank Kaito Umemura for valuable discussion and feedback, and Egbert Leigh and Joseph Wright for their help with the BCI dataset. We thank Jessica Green for the serpentine data. The BCI forest dynamics research project was founded by S.P. Hubbell and R.B. Foster and is now managed by R. Condit, S. Lao, and R. Perez under the Center for Tropical Forest Science and the Smithsonian Tropical Research in Panama. Numerous organizations have provided funding, principally the U.S. National Science Foundation, and hundreds of field workers have contributed. Publication made possible in part by support from the Berkeley Research Impact Initiative (BRII) sponsored by the UC Berkeley Library.

Chapter 3

Implementing the iteration scheme for DynaMETE, a dynamic extension of the Maximum Entropy Theory of Ecology

Abstract

Macroecological theory has largely focused on static large scale patterns, and there is not yet a theoretical framework for predicting how these patterns change over time. The Maximum Entropy Theory of Ecology (METE) predicts the shape of macroecological patterns in relatively static ecosystems by imposing constraints using state variables and then inferring distributions using MaxEnt. Its predictions have largely been successful in ecosystems where the state variables are relatively static, but have failed in disturbed ecosystems where the state variables are changing rapidly on ecological time scales. We here present a theory called DynaMETE, which extends METE to predict time trajectories of state variables and macroecological patterns. The theory hybridizes MaxEnt with underlying biological mechanism to determine the effects of disturbance. We explore the iterative scheme for this theory and develop a semi-analytic approach to iteration we call λ dynamics. We also present code for numerical iteration of the theory. We test the iteration scheme and find that it is not stable against all perturbations. Finally, we discuss how the theory could be altered to increase stability and alternative iteration schemes. We hope that further development of DynaMETE will improve its predictive ability, and allow it to be connected more directly to data.

3.1 Introduction

Macroecology seeks to understand the shape and origin of patterns in the abundance, energetics, and spatial distributions of individuals and species (Brown 1995; Rosenzweig 1995; Gaston and Blackburn 2000). However, to this point, macroecological theory has largely focused on large scale patterns observed at a single time point, and ignored how these patterns change over time (Fisher et al. 2010). There are some exceptions (e.g. Hill and Hamer 1998; Dornelas 2010; Turner 2010; Newman 2019), but there is no theory that provides a larger theoretical framework for predicting how macroecological patterns change in ecosystems out of steady state.

While theory has largely not advanced past static predictions, there is increasing empirical evidence that macroecological patterns are not static (Kempton and Taylor 1974; Carey et al. 2006; Harte 2011; Supp et al. 2012; Rominger et al. 2016; Newman et al. 2020; Franzman et al. 2021). The patterns observed in these studies across largely different ecosystems and with different study organisms indicate that macroecological patterns do change over time, and that they appear to be different between disturbed ecosystems and those that are roughly in steady state. Additionally, the way that these patterns change seem to depend on how the site is disturbed. For example, the slope of the species-area relationship remains relatively constant with scale in the bishop pine forest recovering from wildfire studied by Newman et al. (2020), but decreases sharply with scale in the declining alpine meadow studied by Franzman et al. (2021). The evenness of the species abundance distribution also changes differently in these two studies, where the disturbed pine forest is less even than the undisturbed forest while the declining meadow becomes more even over time.

Here I describe DynaMETE (Harte et al. 2021), a theory that attempts to build a broad theoretical framework for how disturbances change the shape of macroecological patterns and how they change over time. It is built using a maximum entropy (MaxEnt) framework, which selects the least informative distribution possible given constraints (Jaynes 1957; Jaynes 1982). The MaxEnt form of a distribution is obtained by maximizing the Shannon information entropy (Shannon and Weaver 1949). The Maximum Entropy Theory of Ecology (METE) is a static theory based on the MaxEnt framework (Harte 2011; Harte and Newman 2014), whose predictions have been able to describe macroecological patterns across diverse habitats and taxa (Harte 2011; White et al. 2012; Xiao et al. 2015).

METE assumes that state variables vary slowly on ecological time scales, and therefore their instantaneous values are sufficient for describing macroecological patterns. In a disturbance regime, this assumption often no longer holds as the state variables vary rapidly with time, and so it makes sense that in disturbed ecosystems macroecological patterns cannot be characterized by state variables alone. This is analogous to thermodynamics, where the macroscopic state variables of pressure, volume, and temperature can be used to derive the Boltzmann distribution for molecular kinetic energies using MaxEnt (Jaynes 1982), but for a gas out of steady state, the values of these state variables are no longer sufficient to determine this distribution. Note here the difference between steady state and equilibrium. While ecological systems are unlikely to be in true equilibrium as there will always be birth,

death, and other dynamical processes, it is possible for them to maintain a steady state over time if these processes are balanced. In the case of METE, we can reasonably assume that the ecosystem is in steady state as long as the state variables that characterize the system change slowly, even though there are many underlying biological processes.

DynaMETE uses METE as a starting point, and additionally includes mechanism to model disturbance. In steady state, its predictions are the same as METE, but away from steady state its predictions depend on the disturbance itself. DynaMETE uses the same state variables as METE, the number of species S, the number of individuals N, and the total metabolic rate E, but additionally adds the time derivatives of these state variables as constraints. These derivatives are related to underlying mechanism through transition functions that characterize the ecosystem. Because the constraints themselves are now time dependent, we can iterate DynaMETE forward in time by updating constraints and then redoing the MaxEnt procedure.

In this chapter, I first review the formulation of DynaMETE, and then present the proposed iteration scheme along with accompanying code run simulations of the theory. I then discuss ongoing work exploring alternative forms for the transition functions, and alternative iteration schemes.

3.2 The structure of DynaMETE

The core of DynaMETE, as with METE, is the structure function $R(n, \varepsilon)$. This distribution is a function of n, the abundance of a single species, and ε, the metabolic rate of an individual. Thus $Rd\varepsilon$ is the probability that a species picked at random from the species pool has abundance n, and that an individual picked at random from that species has a metabolic rate in the interval $(\varepsilon, \varepsilon + d\varepsilon)$. Note that n is therefore discrete as it represents abundance, and ε is continuous. We normalize the metabolic rate such that the smallest metabolic rate is $\varepsilon = 1$. The structure function is normalized so that $\sum_{n=1}^{N} \int_{\varepsilon=1}^{E} d\varepsilon \; R(n, \varepsilon) = 1$.

The structure function $R(n, \varepsilon)$ describes how individuals are distributed amongst species, and how metabolism is distributed over individuals. From an information theoretic perspective, this makes sense given our fixed set of state variables, N, E, and S. From this perspective, we want to know the least biased way of distributing individuals given these constraints. From an ecological perspective, these variables also make sense as the total energy available in an ecosystem is fixed, at least to some extent, and the number of individuals is measurable. The microscopic variables are also well motivated ecologically, in that the number of individuals in a given species is of great interest in macroecology, and the individual metabolic rate is related to the body size of individuals. The relationship between the state variables and microscopic variables has an analogy in thermodynamics. As an example, in an ideal gas the state variables are the pressure, volume, number of moles, and temperature. Given that, using MaxEnt (Jaynes 1957; Jaynes 1982), we can predict the underlying Boltzmann energy distribution, where the microscopic variable is the energy of an individual particle in the gas.

In METE, the distribution R depends on the state variables S, N, and E, but as mentioned in the introduction DynaMETE has additional constraints corresponding to the derivatives of the state variables. To simplify the notation, we write X_i to indicate state variables with $X = (N, E, S)$. The full set of constraints is then:

$$\frac{N}{S} = \sum_{n=1}^{N} \int_{\varepsilon=1}^{E} d\varepsilon \; nR(n, \varepsilon | X, dX/dt) \tag{3.1}$$

$$\frac{E}{S} = \sum_{n=1}^{N} \int_{\varepsilon=1}^{E} d\varepsilon \; n\varepsilon R(n, \varepsilon | X, dX/dt) \tag{3.2}$$

$$\frac{1}{S}\frac{dN}{dt} = \sum_{n=1}^{N} \int_{\varepsilon=1}^{E} d\varepsilon \; f(n, \varepsilon, X) R(n, \varepsilon | X, dX/dt) \tag{3.3}$$

$$\frac{1}{S}\frac{dE}{dt} = \sum_{n=1}^{N} \int_{\varepsilon=1}^{E} d\varepsilon \; h(n, \varepsilon, X) R(n, \varepsilon | X, dX/dt) \tag{3.4}$$

$$\frac{dS}{dt} = \sum_{n=1}^{N} \int_{\varepsilon=1}^{E} d\varepsilon \; q(n, \varepsilon, X) R(n, \varepsilon | X, dX/dt). \tag{3.5}$$

The functions f, h and q we call transition functions. These functions describe the mechanisms at the micro scale that are relevant for the change in derivative of the corresponding macroscopic state variable. In the initial formulation of the theory, f includes parameters for birth and death rate as well as immigration, h includes terms for the growth of an individual (ontogenetic growth) as well as death and immigration, and q includes terms for immigration and extinction. We emphasize that the transition functions included in Harte et al. 2021 are just one possible form for these functions, and other underlying mechanisms could be included. These transition functions are where the theory is hybridized between MaxEnt and mechanism: the parameters included here characterize the processes happening at the scale of individuals or species, and by summing over the structure function we see their effect in terms of changing the state variables.

Again to simplify notation, we rewrite the constraints as

$$F_\mu = \sum_{n=1}^{N} \int_{\varepsilon=1}^{E} d\varepsilon \; f_\mu(n, \varepsilon, X) R(n, \varepsilon | X, dX/dt), \tag{3.6}$$

where F_μ are the constraints in terms of the state variables and their derivatives, and the f_μ are the corresponding transition functions. The index μ here runs from 1 to 5. Throughout this text, we will use Greek indices (ie. μ, ν) when the index runs from 1 to 5, as is the case with the constraints and transition functions (F_μ and f_μ), and Latin indices (ie. i, j) when the index runs from 1 to 3, as is the case with the state variables and their derivatives (X_i and dX_i/dt). Note that f_μ includes n and $n\varepsilon$ from standard METE, and that these can also

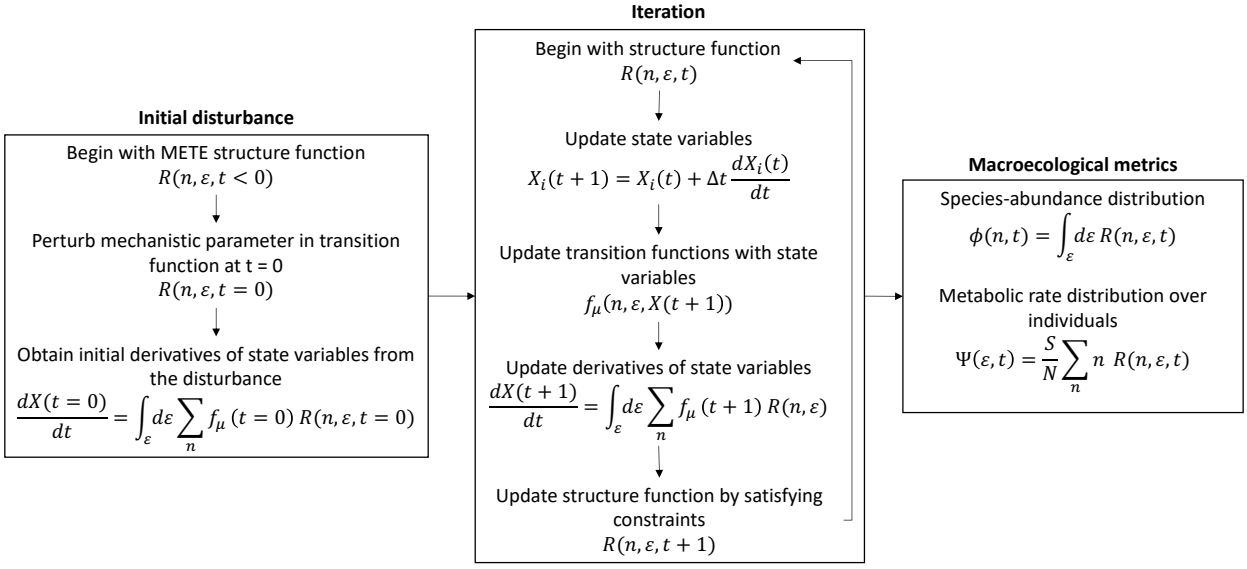


Figure 3.1: The structure of DynaMETE. The first panel describes an initial disturbance away from the standard METE solution, assuming that $R(t<0)$ is the standard METE R. The central panel then describes the iteration process to take $R(t)$ to $R(t+1)$. Finally, the third panel shows how to sum over R to obtain predictions for macroecological patterns at any time t.

be viewed as transition functions. We can simplify this even further by writing $F_\mu = \sum f_\mu R$ where the dependencies are implicit and the sum represents the sum and integral.

The information entropy is then defined as $H = -\sum_n \int_\varepsilon R\log(R)$. The solution to this set of equations obtained by applying MaxEnt is

$$R = \frac{e^{-\sum_\mu f_\mu \lambda_\mu}}{Z} \tag{3.7}$$

where the λ_μ are the corresponding Lagrange multipliers that can be solved using the constraint conditions in Eqs. 3.1–3.5, and Z is the normalization constant. Note that in the case $\lambda_{3-5} = 0$, this simplifies to the METE solution.

Figure 3.1 shows the overall structure of DynaMETE. The first panel describes an initial disturbance from steady state, where at time $t < 0$ we begin from the METE solution with $\lambda_{3-5} = 0$. The second panel shows the overall iteration process described below. Finally, as in METE, macreocological distributions can be calculated by summing appropriately over R for DynaMETE, which is shown in the third panel in Fig. 3.1.

3.3 Iteration scheme

In order to calculate how the the structure function varies in time, we need to construct an iteration scheme for the theory. The central panel in Fig. 3.1 shows the iteration scheme for DynaMETE. In it, we will assume that the transition functions do not depend on the derivatives of the state variables, but do depend on the state variables themselves. We will also assume that any rate parameters for underlying process, such as birth rate or death rate, are constant in time. This assumption could easily be relaxed if we assume a form for the time evolution of the parameters.

We begin at a discrete time t, where the constraint conditions will be satisfied. We can then update the X_i directly using their time derivatives as

$$X_i(t+1) = X_i(t) + \frac{dX_i(t)}{dt}\Delta t. \tag{3.8}$$

We then have to update the time derivatives of the state variables. Since the transition functions f_μ are are allowed to depend on X_i, we first update the transition functions to $f_\mu(t+1)$ by plugging in the updated state variables. We can then update the time derivatives using Eqs. 3.3, 3.4, and 3.5 as

$$\frac{dN(t+1)}{dt} = S(t+1)\sum_{n=1}^{N(t+1)}\int_{\varepsilon=1}^{E(t+1)} d\varepsilon\ f(n,\varepsilon,X(t+1))R(n,\varepsilon|X(t+1),dX(t)/dt) \tag{3.9}$$

$$\frac{dE(t+1)}{dt} = S(t+1)\sum_{n=1}^{N(t+1)}\int_{\varepsilon=1}^{E(t+1)} d\varepsilon\ h(n,\varepsilon,X(t+1))R(n,\varepsilon|X(t+1),dX(t)/dt) \tag{3.10}$$

$$\frac{dS(t+1)}{dt} = \sum_{n=1}^{N(t+1)}\int_{\varepsilon=1}^{E(t+1)} d\varepsilon\ q(n,\varepsilon,X(t+1))R(n,\varepsilon|X(t+1),dX(t)/dt), \tag{3.11}$$

where the Lagrange multipliers in R are evaluated at t, not at $t+1$. Note that in theory the upper limits of the sum and integral should be evaluated at $t+1$, as noted in these equations, however in practice with a small enough time step the difference will be minimal. Since R is an exponential function it is much smaller at large n and ε, so small changes in the upper limits of the sum are not significant.

Finally then, we can update the structure function R by solving for the Lagrange multipliers at $t+1$ with the initial constraint conditions. Note that at this step, Eqs. 3.3, 3.4, and 3.5 will already be satisfied as we derived the updated time derivatives using these expressions. We therefore need to obtain a new set of λs that satisfy the first two constraints while keeping the constraints for the time derivatives satisfied. Doing this maximization process will give us $R(t+1)$.

In practice, this is quite challenging to solve numerically as it involves simultaneously solving large sums and integrals. We therefore developed a semi-analytic method for iterating the theory, which we call λ dynamics.

3.4 λ dynamics

At each step of our iteration scheme, we have to maximize entropy for the function R. This is quite computationally intensive due to the complexity of the constraint conditions, and the large sum and integral for larger ecosystems. We here present a semi-analytic method to solve for the Lagrange multipliers at each new step. This means that rather than maximizing entropy numerically, we need only invert a matrix to solve for the derivatives of the λs at each step.

To demonstrate how this method works, let us consider only the equations relating to the state variable N: Eqs. 3.1, 3.3, and 3.9. We additionally have the equation

$$\frac{dN(t+1)}{dt} = S(t+1)\sum_{n=1}^{N(t+1)}\int_{\varepsilon=1}^{E(t+1)} d\varepsilon\ f(n,\varepsilon,X(t+1))R(n,\varepsilon|X(t+1),dX(t+1)/dt) \quad (3.12)$$

from solving the constraints at time $t+1$. Subtracting Eq. 3.9 from this equation gives

$$0 = \sum_n \int_\varepsilon d\varepsilon\ f(n,\varepsilon,X(t+1))\left(R(n,\varepsilon|X(t+1),\lambda(t+1)) - R(n,\varepsilon|X(t+1),\lambda(t))\right) \quad (3.13)$$

where we have written the explicit time dependence of the λ rather than dX/dt, as there is no explicit dependence on dX/dt in R. We can expand the second term as

$$R(n,\varepsilon|X(t+1),\lambda(t)) \approx R(n,\varepsilon|X(t+1),\lambda(t+1)) - \sum_\mu \frac{\partial R(n,\varepsilon|X(t+1),\lambda(t+1))}{\partial\lambda_\mu(t+1)}\frac{d\lambda_\mu(t+1)}{dt} \quad (3.14)$$

and substitute in to get

$$\sum_n \int_\varepsilon d\varepsilon\ \sum_\mu f(n,\varepsilon,X(t+1))\frac{\partial R(n,\varepsilon|X(t+1),\lambda(t+1))}{\partial\lambda_\mu(t+1)}\frac{d\lambda_\mu(t+1)}{dt} = 0. \quad (3.15)$$

Taking these partial derivatives with respect to λ gives

$$\frac{\partial R}{\partial\lambda_\mu} = (-f_\mu + F_\mu)R \quad (3.16)$$

where the second term, F_μ, comes from the partial derivative of Z. Rewriting Eq. 3.15 then gives

$$\sum_\mu \left(\sum_n \int_\varepsilon d\varepsilon\ f(n,\varepsilon,X)f_\mu(n,\varepsilon,X)R(n,\varepsilon|X,\lambda) - F_\mu\left(X,dX/dt\right)\frac{N}{S}\right)\frac{d\lambda_\mu}{dt} = 0 \quad (3.17)$$

where all variables are evaluated at $t+1$, but it is omitted for readability. We can rewrite the above in terms of covariances, where $\mathrm{Cov}(X,Y) = \langle XY\rangle - \langle X\rangle\langle Y\rangle$, where angular brackets represent the average over the distribution R. The end result is

$$\sum_\mu \mathrm{Cov}\left(f,f_\mu\right)\frac{d\lambda_\mu}{dt} = 0. \quad (3.18)$$

This derivation can be replicated for E and S to obtain

$$\sum_{\mu} \mathrm{Cov}\left(h, f_{\mu}\right) \frac{d\lambda_{\mu}}{dt} = 0 \tag{3.19}$$

$$\sum_{\mu} \mathrm{Cov}\left(q, f_{\mu}\right) \frac{d\lambda_{\mu}}{dt} = 0. \tag{3.20}$$

To solve for $d\lambda/dt$ we need two additional equations. We can obtain these by taking the time derivative of Eqs. 3.1 and 3.2. These derivatives give us another form for dN/dt and dE/dt. Using similar methods to the derivation above, we obtain

$$\frac{1}{S}\frac{dN}{dt} - \frac{N}{S^2}\frac{dS}{dt} + \sum_{\mu}\left(\mathrm{Cov}\left(n, f_{\mu}\right)\frac{d\lambda_{\mu}}{dt} + \mathrm{Cov}\left(n, df_{\mu}/dt\right)\lambda_{\mu}\right) = 0 \tag{3.21}$$

$$\frac{1}{S}\frac{dE}{dt} - \frac{E}{S^2}\frac{dS}{dt} + \sum_{i}\left(\mathrm{Cov}\left(n\varepsilon, f_{\mu}\right)\frac{d\lambda_{\mu}}{dt} + \mathrm{Cov}\left(n\varepsilon, df_{\mu}/dt\right)\lambda_{\mu}\right) = 0 \tag{3.22}$$

where we note that $df_0/dt = df_1/dt = 0$, so the second term in the brackets is only non-zero for $\mu = 3, 4, 5$.

Thus, we have five equations for the five derivatives of the lambdas. We can then iterate the theory forward in time using the same iteration scheme as normal, but rather than maximizing entropy to obtain a new set of λs, we obtain the new set of λs as $\lambda(t+1) = \lambda(t) + d\lambda(t)/dt\Delta t$. For small step sizes, this should converge to the same solution as if we were to maximize entropy at each step. We can test this numerically if we specify transition functions.

3.5 Specifying the transition functions

To this point, we have left the framework of DynaMETE intentionally very general. Any form of the transition functions that depend on X but not dX/dt will work. These functions are meant to capture the small scale processes that will change the macro state variables.

As an example, we here consider the transition functions presented in Harte et al. 2021. They are

$$f(n,\varepsilon) = \left(b_0 - d_0\frac{E}{E_c}\right) n\varepsilon^{-1/3} + \frac{m_0}{N}n \tag{3.23}$$

$$h(n,\varepsilon) = \left(w_0 - d_0\frac{E}{E_c}\right) n\varepsilon^{2/3} - \frac{w_{10}}{\ln^{2/3}(1/\beta)}n\varepsilon + \frac{m_0}{N}n \tag{3.24}$$

$$q(n,\varepsilon) = m_0 e^{-\mu S - \gamma} - d_0 S\frac{E}{E_c}\delta_{n,1}\varepsilon^{-1/3}. \tag{3.25}$$

The parameters are the birth rate b_0, the death rate d_0, the metabolic carrying capacity of the ecosystem E_c, the immigration rate m_0, the ontogenetic growth rates w_0 and w_{10}, and a

parameter for the size of the metacommunity μ. We now explain the justification for these forms for the transition functions.

First we consider the form of f. Taking the birth and death rates to be proportional to $n\varepsilon^{-1/3}$ in f is consistent with metabolic scaling theory (West et al. 2001; Brown et al. 2004). The E/E_c term represents a constraint that operates at the community level, rather than at the individual or species level. More broadly, note that d_0E/E_c always appears together in the transition functions, so d_0/E_c is effectively a single parameter characterizing the death rate with a metabolic rate carrying capacity. The final term in f characterizes the immigration rate. If an individual immigrant is a member of a species already present in the local community, we assume that the probability that it is from a species with abundance n is simply n/N. Given that the vast majority of immigrants will be from an existing species, we can assume the overall immigration rate m_0 needs only to be multiplied by n/N to give the overall rate of immigration to a species with n individuals.

The expression $w_0\varepsilon^{2/3} - w_1\varepsilon$ is consistent with ontogenetic growth (West et al. 2001). Here we have modified it in h by multiplying by the abundance of the species n, since that expression is at the individual level, and we have added a scaling constant $1/\ln^{2/3}(1/\beta)$, where $\beta = \lambda_1 + \lambda_2$. The immigration term is the same as for f as we assume immigrants have $\varepsilon = 1$. Finally, we multiply the death term in f by ε to account for the metabolic rate of the individual that is dying, and we ignore birth as it primarily partitions rather than adds to metabolism.

Finally, for q, we ignore the speciation models discussed in Harte et al. 2021 and consider only immigration. The first term represents the rate of immigration of new species. The full derivation is in SI-D in Harte et al. 2021, but essentially this considers the probability that an immigrant from a metacommunity with number of species S_m and abundance N_m is from a species not already present in the local community. The derivation assumes the metacommunity species abundance distribution follows the METE prediction, and so is a log-series parameterized by β_m. The parameter μ in q is equal to $\ln(1/\beta_m)/S_m$. We then multiply this by the overall immigration rate m_0 to obtain the rate of immigration of new species. Note that this term does not depend on n or ε, so it will be cancelled out in the normalization of R. However, it still contributes to dS/dt. The second term in q is the rate at which species with abundance $n = 1$ go extinct. This is the overall death rate multiplied by the Kronecker delta $\delta_{n,1}$, which is 1 when $n = 1$ and 0 otherwise.

3.6 Iteration code

With the transition functions and iteration scheme specified, we can now numerically iterate the theory. We have outlined two numerical approaches for doing so: brute force maximum entropy at every step, and λ dynamics. I have written code for each case which can be found at https://github.com/micbru/dynamete_iteration. The code `DynaMETE_Rfunctions.py` implements the necessary sums and integrals over R, `mean_covariances.py` uses these calculates to get the means and covariances of the transition functions, and `brute_force.py` and

`lambda_dynamics.py` implement the iteration with the corresponding schemes. The jupyter notebook `Iteration.ipynb` reproduces all of the results here and shows how to use these functions.

For both codes, note that the transition functions specified above depend only linearly on n, with the exception of q. However the dependence in q is a δ function which can be pulled out of the sum. Therefore, the sum over n can be done analytically, which greatly increasing code performance. The sum over n for Eqs. 3.1–3.4 (not including the normalization Z) is of the form

$$\sum_n n e^{-c(\varepsilon)n - q(n,\varepsilon)\lambda_5}, \tag{3.26}$$

where $c(\varepsilon)$ is equal to $\sum_{\mu=1}^{4} f_\mu \lambda_\mu / n = \lambda_1 + \lambda_2 \varepsilon + \lambda_3 f(n=1,\varepsilon) + \lambda_4 h(n=1,\varepsilon)$. Note that μ only runs until 4 in this sum as the q dependence is pulled out of the sum. Given the first term in q is not n dependent, it can be pulled out of the sum. The second term has $\delta_{n,1}$, so we can separate the sum as

$$e^{-q(n=0,\varepsilon)\lambda_5} \sum_{n=1}^{N} n e^{-c(\varepsilon)n} + e^{-c(\varepsilon) - q(n=1,\varepsilon)\lambda_5}. \tag{3.27}$$

The solution is then

$$Z \sum_n nR = e^{-c(\varepsilon) - q(n=1,\varepsilon)\lambda_5} + e^{-2c(\varepsilon) - q(n=0,\varepsilon)\lambda_5} \frac{2 - e^{-c(\varepsilon)} + e^{-c(\varepsilon)(N-1)} \left(N e^{-c(\varepsilon)} - N - 1 \right)}{\left(1 - e^{-c(\varepsilon)}\right)^2}. \tag{3.28}$$

The sum over n for the normalization Z can also be calculated analytically using the same technique. The form is the same but we do not multiply the structure function R by n. The solution is

$$Z \sum_n R = e^{-c(\varepsilon) - q(n=1,\varepsilon)\lambda_5} + e^{-2c(\varepsilon) - q(n=0,\varepsilon)\lambda_5} \frac{1 - e^{-c(\varepsilon)(N-1)}}{1 - e^{-c(\varepsilon)}}. \tag{3.29}$$

We then integrate this over ε to get the normalization Z.

These sums are implemented as functions in `DynaMETE_Rfunctions.py` as `nRsum` and `Rsum`. Note that for the covariances needed for λ dynamics, we also need the sum $\sum_n n^2 R$, which is done in a similar way and implemented as `n2Rsum`.

Now we only have to integrate over ε to solve the constraint equations. Again to speed up the calculations, we only calculate each required integral over ε once. The required combinations for the transition functions alone are the sum over $n\varepsilon^{-1/3}$, n, $n\varepsilon^{2/3}$, $n\varepsilon$, and $\delta_{n,1}\varepsilon^{-1/3}$. For the covariances for λ dynamics, we additionally need $n^2\varepsilon^{-2/3}$, $n^2\varepsilon^{-1/3}$, n^2, $n^2\varepsilon^{1/3}$, $n^2\varepsilon^{2/3}$, $n^2\varepsilon$, $n^2\varepsilon^{4/3}$, $n^2\varepsilon^{5/3}$, $n^2\varepsilon^2$, and $\delta_{n,1}\varepsilon^{-2/3}$, $\delta_{n,1}\varepsilon^{1/3}$, $\delta_{n,1}\varepsilon^{2/3}$. These are all implemented by the `get_means` function in `DynaMETE_Rfunctions.py`. This function takes the Boolean parameter `alln`, which if false only calculates the sums needed for the brute force maximization, and if true calculates all of the sums needed for λ dynamics.

To this point, we have only calculated the various sums and integrals over n and ε. The file `means_covariances.py` contains functions that take in this array and calculate individual means and covariances over the various transition functions and derivatives as required by the constraint equations (Eqs. 3.1–3.5), and the λ dynamics equations (Eqs. 3.18–3.22).

The final step is to use these calculations to iterate the theory as described in the previous section. This is implemented with the `iterate` function in both `brute_force.py` and `lambda_dynamics.py`. Note that the `lambda_dynamics.py` file has an additional method `get_dl_matrix_vector` which calculates the covariance matrix defined by Eqs. 3.18–3.22 that we need to invert in order to solve for the derivatives of the Lagrange multipliers.

Overall, the goal is that users need only call the `iterate` function, specifying the parameters for the transition functions as a labeled list `p`, the initial state variables `s0`, and the length of the iteration `t`. This function sets the default iteration step as `dt=0.2`, though this parameter can also be passed in.

In theory, we can use the iteration scheme from any initial conditions, which can be optionally passed in as the initial λs, `l0` and the initial derivatives of the state variables, `ds0`. However given that our goal with DynaMETE is to model disturbance, we will often assume the ecosystem begins in steady state before one or more of the parameters are perturbed. We can then iterate from there to see the effect of that type of disturbance.

To begin in steady state, we set some parameters based on biological plausibility, and then fix others so that the derivatives of the state variables are zero initially and $\lambda_{3-5} = 0$. This initial optimization of parameters at steady state is implemented in the function `get_ss_params` in `lambda_dynamics.py`, which is set up to solve for w_{10}, μ, and m_0 but can easily be changed to solve for different parameters. The initial λs in the iteration are then equivalent to the METE λs. To begin the iteration process, we perturb parameters and use Eqs. 3.9, 3.10, and 3.11 to derive the initial time derivatives. We then update the state variables as normal, beginning the first iteration step, which should now continue as the time derivatives are no longer zero. This method for establishing the initial disturbance is shown in the first panel in Fig. 3.1.

The default with the `iterate` function if initial conditions are not passed in is to assume the system begins at steady state, but the parameters have been perturbed. The function then calculates the corresponding METE λs, sets the initial time derivatives using Eqs. 3.9, 3.10, and 3.11, and then begins the iteration process. Therefore, to iterate the theory from a perturbed steady state, we need only pass in the perturbed parameters and the state variables at steady state.

This function then returns the λs, state variables, and their derivatives at each time step.

3.7 Exploring the theory

We now need to specify parameter values. We choose parameter values that are plausible for a tropical forest. More specifically, we use Barro Colorado Island (BCI), a tropical forest in Panama, as a guide to obtain realistic parameter values. This site has 30 years of census

data available (Condit 1998; Hubbell et al. 1999; Hubbell et al. 2005; Condit et al. 2019), allowing for comparison to the theory over time. Further, both N and S are declining over time at this site, which may be a consequence of disturbance as the formation of Gatun Lake has resulted in the semi-isolation of the forest from its metacommunity. This site is therefore a good starting point to test the theory as we have a hypothesis for what kinds of disturbance may be affecting this site, as well as available time series data.

The parameters and state variables presented in Table 3.1 are obtained as biologically plausible values for BCI. These values are meant to be in line with what BCI could have looked like before the formation of Gatun Lake, around 1905. The state variables are rounded from what we see in 1985, where $S = 309$, $N = 241\ 786$, and $E = 19\ 788\ 030$. E is in units such that a tree with 1 cm dbh has a metabolic rate of 1 and this number was calculated assuming that metabolic rate scales as dbh^2 (Brown et al. 2004; Marbà et al. 2007; but see Muller-Landau et al. 2006). The initial parameters are then chosen by first fixing the parameters E_c, b_0, d_0, and w_0 to be biologically plausible, and then solving for w_{10}, μ, and m_0 while setting the time derivatives in Eq. 3.3–3.5 to 0.

The choice of initial parameters here is as in SI-F in Harte et al. 2021. To summarize, we set E_c as a round number near the steady state values, $E_c = 2 \times 10^7$, and then take E to be slightly larger as a steady state value. Note that we take $E > E_c$ to balance against the positive migration term in Eq. 3.3, though note that if we are at a site with a net migration outflow then $m_0 < 0$ and $E < E_c$. The birth and death rates assume that average saplings with 1 cm dbh die at a rate of 0.2/year, and we set $b_0 = d_0$. The growth rates of trees are set assuming that a sapling of 1 cm dbh doubles in diameter in 3-4 years, and becomes a tree of 30 cm dbh in 80 years, which gives $w_0 \approx 1$ and $w_{10} \approx 0.4$. Finally, we can check the biological plausibility of μ by assuming the size of the metacommunity, which is this case is about 100 times the area of the BCI plot.

Note that our parameters w_{10}, μ, and m_0 are slightly different from the parameters in Harte et al. 2021, as we use more precise numerical solutions for the parameters that set the initial derivatives to zero. This ensures that if we iterate the theory from steady state, we remain in steady state.

We emphasize that the parameter set here is meant to serve as a starting point, and gives us a way to numerically test the structure of the theory as well as different iteration schemes. It is not meant to be compared to BCI data directly, or to serve as a set of best fit parameters describing the site.

We first test our semi-analytic iteration scheme, λ dynamics. Figure 3.2 shows trajectories for the λs and the derivatives of the state variables under a perturbation to the death rate $d_0 \to 0.25$ for both the brute force and λ dynamics iteration schemes for $\Delta t = \{0.1, 0.2, 0.5\}$. We can see that for small enough Δt the different iteration schemes give very similar results, and further that the solution overall seems to converge for small Δt.

We have omitted plots of the state variables themselves as their change is relatively small over this time window, as well as plots of the macroecological patterns themselves obtained from R as the difference between the predicted form for METE and DynaMETE after this many iterations is small. To get a larger effect, we could run the perturbation longer or

Parameter	Values	State Variables	Values
b_0	0.2	N	2.3×10^5
d_0	0.2	E	2.04×10^7
m_0	437.3	S	320
w_0	1.0		
w_{10}	0.42		
E_c	2×10^7		
μ	0.0215		

Table 3.1: The numerical parameters we use to explore DynaMETE initially.

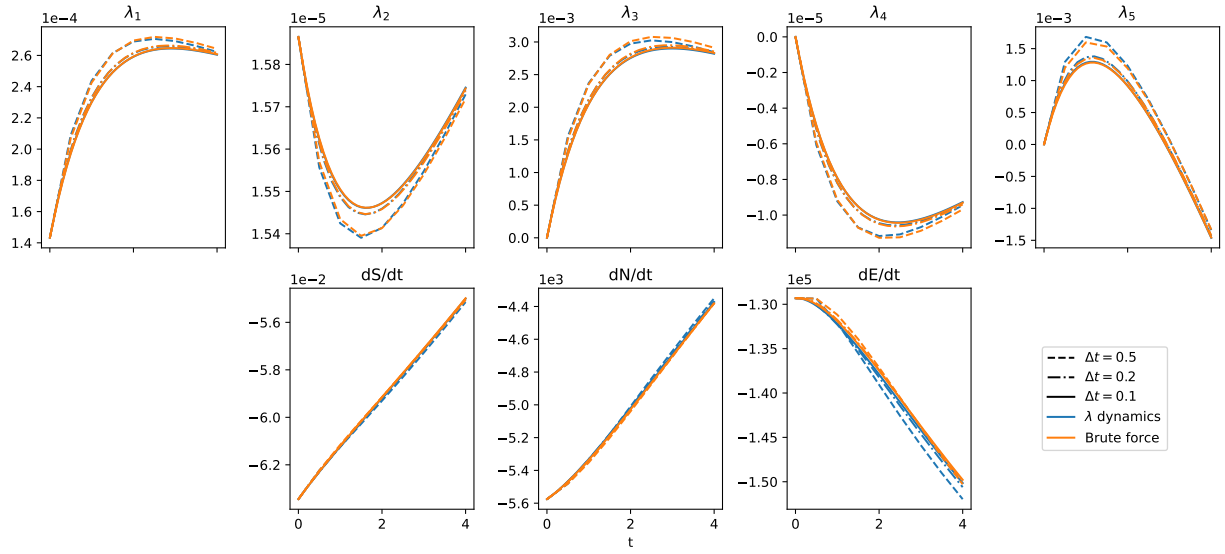


Figure 3.2: Time trajectories of the Lagrange multipliers λ_μ and the derivatives of the state variables dX_i/dt after an increase in death rate from $d_0 = 0.20$ to $d_0 = 0.25$, with the rest of the parameters given in Table 3.1. The color represents the iteration method, and the dashes represent the time step used. These are calculated using the `lambda_dynamics` and `brute_force` codes.

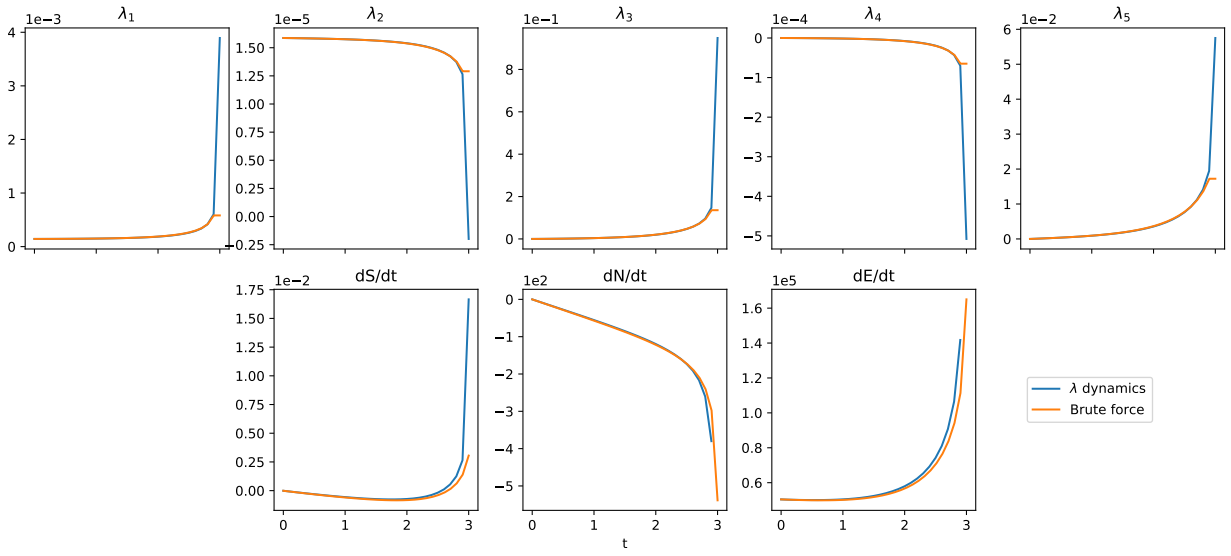


Figure 3.3: Time trajectories of the Lagrange multipliers λ_μ and the derivatives of the state variables dX_i/dt after an decrease in ontogenetic growth rate from $w_{10} = 0.42 \to 0.4096$, with the rest of the parameters given in Table 3.1. The color represents the iteration method, and overlap until the iteration fails to progress. These are calculated using the `lambda_dynamics` and `brute_force` codes.

make the perturbation larger. The iteration here is meant to serve as an example for how the iteration scheme works, as well as to show that λ dynamics is working as expected.

After setting up this code, we tried other perturbations and noticed that in some cases the theory seems to have issues with stability. An example of this is perturbing the ontogenetic growth rate w_{10} down to the initial value in Harte et al. 2021, $w_{10} = 0.42 \to 0.4096$. With the brute force iteration scheme, the maximization method can no longer find a solution that satisfies the constraints, and with λ dynamics, the constraints stop being satisfied. This becomes apparent looking at the trajectories for the λs and the derivatives of the state variables, which are shown in Fig. 3.3. Each of the dX/dt and λs blow up. This is not solved by taking a smaller time step and seems to be part of the theory, especially because the failure appears for both λ dynamics and the brute force iteration schemes.

Given that the same issue with the theory appears with both iteration schemes, we are led to believe that the issue is fundamental to the theory, or a result of the specific transition functions chosen here, rather than a numerical instability. We have a few approaches to determine what is causing the theory to fail. One potential issue is the initial condition of the time derivatives. In the d_0 perturbation, dN/dt is heading towards 0, and if we continue the iteration dE/dt also curves back towards 0. However in the w_{10} perturbation, all of the time derivatives and λs seem to continue to increase in the same direction before blowing up. We are currently investigating the second derivatives of the state variables to see if the

form of the transition functions causes some kind of runaway effect.

Another possibility is that the theory collapses when there are not enough degrees of freedom for the constraint conditions. This is because the transition functions, and therefore the exponent for R, depends only on n, $n\varepsilon$, $n\varepsilon^{-1/3}$, $n\varepsilon^{2/3}$, and $\delta_{n,1}\varepsilon^{-1/3}$, all of which are linear in n. This should be sufficient degrees of freedom to solve in that we have five different forms and five constraints, but perhaps there is overlap between the $n\varepsilon^{-1/3}$ term and the $\delta_{n,1}\varepsilon^{-1/3}$ term. In a toy model without ε, we can show that having only linear n cause the theory to collapse, and that adding a term with a different n dependence fixes this issue. We therefore explore different n dependence for some of the terms below.

3.8 Ongoing and future work

Alternative transition functions

Given we suspect the numerical failure of the iteration scheme is related to the exponent in R depending only on linear n and a few different powers of ε, we are working on different forms of the transition functions that are still biologically plausible but will increase the number of degrees of freedom. This requires writing new iteration code, as the iteration code above is specific to the transition functions in Harte et al. 2021. This code is available at https://github.com/micbru/dynamete_iteration and includes the python files `lambda_dynamics_FlexibleFunctions.py`, `DynaMETE_Rfunctions_FlexibleFunctions.py` as well as the jupyter notebook `FlexibleFunctions.ipynb` to replicate the results here. The primary difference is that the sum over n is no longer calculated analytically, which drastically reduces the number of hard coded equations. In this code, the integral over ε is done first in parallel over all n, before being summed over n. This means we can easily change the functional form of the transition functions by changing their definitions, and the definition of their derivatives. The rest of the code does not change, and the functions are all very similar or identical. While this makes the code much more flexible, the major drawback of this approach is that the iteration is much slower as the sums must now be done numerically and the R function takes much longer to compute. Because of this, testing this code with the parameter set in Table 3.1 is not feasible, and we instead test with a similar but smaller ecosystem.

To test this more flexible code, we change the death rate to depend on n^2 rather than n, so $d_0/E_c\, n/\varepsilon^{-1/3} \to d_0/E_c\, n^2/\varepsilon^{-1/3}$. This affects both f and h. This is biologically motivated as density dependent death, but is meant more as an example to show how modifying the code can work. The state variables and parameters in this case are given in Table 3.2. Note that the state variables are scaled down significantly, but are meant to be representative of a smaller sample at BCI and were obtained roughly from a patch roughly 1/256 the size of the entire 50 ha plot. The parameter d_0 was also changed in line with changing $n \to n^2$, and again m_0, w_{10}, and μ were obtained by setting the derivatives of the state variables to zero using the function `get_ss_params` in `lambda_dynamics_FlexibleFunctions.py`. Note that the μ

Parameter	Values	State Variables	Values
b_0	0.2	N	10^3
d_0	0.002	E	2.04×10^4
m_0	127.5	S	30
w_0	1.0		
w_{10}	0.6705		
E_c	2×10^4		
μ	0.3525		

Table 3.2: The numerical parameters we use to explore DynaMETE with alternative transition functions, where $d_0 n \to d_0 n^2$.

parameter here is not particularly realistic, as it would require an enormous metacommunity N_{meta} without that many additional species (for example, $N_{meta} = 10^6$, $S_{meta} = 36$). Again though, this test is primarily meant to demonstrate a possible alternative form for the transition functions.

Figure 3.4 shows trajectories for the derivatives of the state variables and the λs with a d_0 perturbation, $d_0 \to 0.0025$, with this form of the transition function and time step $\Delta t = 0.2$. Note the relative similarity between this figure and Fig. 3.2, which shows a d_0 perturbation for the initial theory. This is promising in that the particular form of the transition function does not seem to change the effects of one type of perturbation all that much, meaning that we may still be able to disentangle the underlying mechanism without having to know the exact function form for the transition function. However, this also means that this modification to the transition function may not solve the stability issues with the iteration, and there may be a different underlying issue with the theory.

Analytic solution to iterations

The iteration scheme presented here depends on using discrete time, since we are obtaining the iteration by advancing the state variables. We believe the solution should converge if we take Δt to be small enough, but have not yet found an analytic form for this solution. The semi-analytic method is a first step towards an analytic solution for the iteration of the λs. This is important given the difficulty of advancing the theory numerically, particularly with large N or E. Additionally, we may be able to see analytically how different transition functions will affect the trajectories in the theory if we can solve the iteration more generally.

For simple toy models of the theory, it is possible to invert Eqs. 3.18–3.22 analytically in terms of $d\lambda_\mu/dt$. However, these equations still depend on both the state variables and the Lagrange multipliers, and we need to invert the constraint conditions to solve for the full time dependence. Inverting the constraint conditions Eq. 3.1–3.5 themselves in the full theory is very challenging, but will be necessary to obtain any analytic expression for the

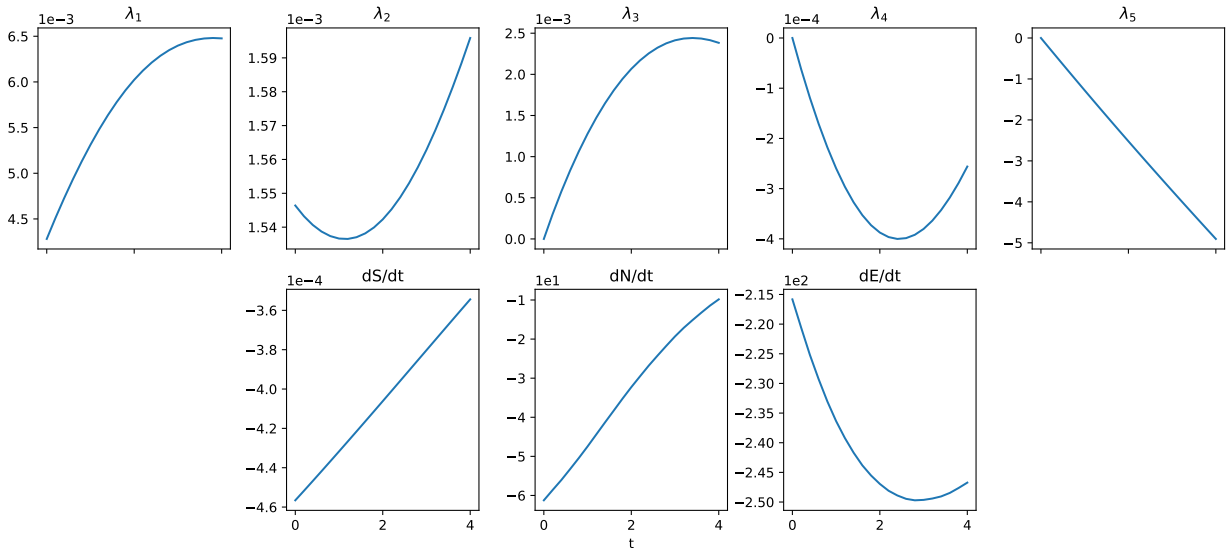


Figure 3.4: Time trajectories of the Lagrange multipliers λ_μ and the derivatives of the state variables dX_i/dt after an increase in death rate from $d_0 = 0.0020$ to $d_0 = 0.0025$, with the rest of the parameters given in Table 3.2 and $\Delta t = 0.2$. These are calculated with the λ dynamics iteration method for the transition functions with $d_0 n \to d_0 n^2$, using the `lambda_dynamics_FlexibleFunctions` code.

time evolution of the theory.

Alternative iteration schemes

Rather than solve the existing iteration scheme, another approach is to develop an alternative. Pessoa et al. 2021 presents a framework for dynamics on Gibbs statistical manifolds, which is the underlying manifold in the case of DynaMETE as it is obtained by maximizing entropy. This framework uses information geometry (Caticha 2015; Amari 2016), which assigns a geometric structure to the space of probability distributions such that the distance between neighbouring distributions $P(x|\theta + d\theta)$ and $P(x|\theta)$ characterized by parameters $\theta = \{\theta_\mu\}$ is given by $d\ell^2 = g_{\mu\nu} d\theta^\mu d\theta^\nu$. Note that in this section we will use the Einstein summation convention, so $g_{\mu\nu} d\theta^\mu d\theta^\nu = \sum_\mu \sum_\nu g_{\mu\nu} d\theta^\mu d\theta^\nu$ as the indices appear twice in a single term as both a subscript and a superscript. The metric $g_{\mu\nu}$ here is the Fisher-Rao information metric given by

$$g_{\mu\nu} = \int dx \; P(x|\theta) \frac{\partial \log P(x|\theta)}{\partial \theta^\mu} \frac{\partial \log P(x|\theta)}{\partial \theta^\nu}. \tag{3.30}$$

In the case of a MaxEnt distribution, θ^μ is equal to the constraints F^μ, and the metric simplifies to the Hessian of the entropy H

$$g_{\mu\nu} = -\frac{\partial^2 H}{\partial F^\mu \partial F^\nu}. \tag{3.31}$$

The inverse of the metric is given by the covariance matrix

$$C^{\mu\nu} = \mathrm{Cov}(f^\mu f^\nu), \tag{3.32}$$

and $C^{\mu\nu} g_{\nu\gamma} = \delta^\mu_\gamma$.

We can use this information geometry framework to iterate the theory by moving around on this statistical manifold with coordinates defined by the constraints F^μ. This has already been explored to some extent for METE in Pessoa 2021, but with DynaMETE the fact that the constraints F^μ are themselves time dependent should allow us to iterate the theory.

It is also worth noting that there are similarities between these equations and λ dynamics. First, we will rewrite λ dynamics in terms of the information metric. Eqs. 3.18–3.20 can be rewritten as

$$C^{\mu\nu}\frac{d\lambda_\nu}{dt} = 0 \text{ for } \mu = \{3, 4, 5\}, \tag{3.33}$$

and Eqs. 3.21–3.22 can be written as

$$C^{\mu\nu}\frac{d\lambda_\nu}{dt} = -\frac{dF^\mu}{dt} - \mathrm{Cov}\left(f^\mu, \frac{df^\nu}{dt}\right)\lambda_\nu \text{ for } \mu = \{1, 2\}. \tag{3.34}$$

If we do not allow the f_μ to depend on X, and so assume $df_\mu/dt = 0$, then this simplifies to

$$C^{\mu\nu}\frac{d\lambda_\nu}{dt} = -\frac{dF^\mu}{dt}. \tag{3.35}$$

Further, with this assumption, we can take the time derivatives of both sides of the constraints Eqs. 3.3, 3.4, and 3.5 to get

$$\frac{dF^\mu}{dt} = -\sum_\nu \mathrm{Cov}(f^\mu, f^\nu)\frac{d\lambda_\nu}{dt} = -C^{\mu\nu}\frac{d\lambda_\nu}{dt} \text{ for } \mu = \{3, 4, 5\}. \tag{3.36}$$

In λ dynamics, this expression is equal to zero. Thus we can rewrite λ dynamics for all μ under this assumption as

$$C^{\mu\nu}\frac{\lambda_\nu}{dt} = -\frac{dF^\mu}{dt}. \tag{3.37}$$

This is as Eq. 19 in Pessoa et al. 2021, where $dF^\mu = -C^{\mu\nu}d\lambda_\nu$, but we have divided by dt and assumed that $df_\mu/dt = 0$ and therefore $dF^\mu/dt = 0$ for $\mu = \{3, 4, 5\}$. However, in the case that this assumption does not hold, these approaches are no longer equivalent. Further work is needed to see if we can develop a more general iteration scheme using this information geometry framework, perhaps using Eq. 50 in Pessoa et al. 2021,

$$\frac{\langle \Delta F^\mu \rangle}{\Delta t} = C^{\mu\nu}\frac{\partial S}{\partial F^\nu} - \frac{\Gamma^\mu}{2}, \tag{3.38}$$

where $\Gamma^\mu = \Gamma^\mu_{\nu\gamma} C^{\nu\gamma}$ and $\Gamma^\mu_{\nu\gamma}$ is the Christoffel symbol.

3.9 Conclusion

DynaMETE is an ambitious theoretical framework that seeks to predict how state variables and macroecological patterns change over time in response to disturbance. However, there is still work to do to connect the framework as presented to ecological data. The iteration scheme with the transition functions initially presented in Harte et al. 2021 does not appear to be stable to many perturbations. This could indicate something biological, or could be a larger issue with the iteration scheme proposed. Numerical code to test the theory as written, as well as to test a more flexible version of the theory, is available at https://github.com/micbru/dynamete_iteration. We are just beginning to explore alternative transition functions and iteration schemes, which will allow for a more complete understanding of the theory. As the theory grows, we hope that DynaMETE will contribute to better understanding of disturbed ecosystems, will help identify processes driving ecological change, and will improve conservation and management strategies.

Acknowledgments

I want to acknowledge that the development of this theory was a largely collaborative effort with John Harte and Kaito Umemura, and I would like to thank them for their feedback and guidance on this chapter. I would also like to thank Nina Groleger and Julia Nicholson, who carried out numerical simulations with the iteration code, and Juliette Franzman and Roya Safeinili, who were involved in the early development of DynaMETE. Funding for this project was provided by grant DEB 1751380 from the US National Science Foundation, and I also acknowledge the support of the Natural Sciences and Engineering Research Council of Canada (NSERC), [PGSD2- 517114-452018].

Conclusion

Throughout this work, the importance of dynamical macroecology is clear. Patterns are not constant in time, and how they vary carries important information about underlying biology. These patterns can be analyzed empirically and compared to static predictions, as in the first chapter where I show how deviations from static theory can provide us with information about underlying biology, and how disturbance in the form of land use changes macroecological patterns.

We can also learn about underlying process by modelling explicit disturbances and comparing resulting predictions to data, as in chapter 2. There I show how modifying static theory to include a specific disturbance can be used to understand how disturbance affects patterns, and that in turn can be used to interpret data.

We can then move from modelling specific disturbance to building broad theoretical frameworks that model generic disturbance and out of steady state dynamics. As with dynamics out of steady state in statistical physics, building theory for macroecological patterns under disturbance is challenging. These types of models are generally the most difficult to connect directly to data, but are important in predicting how ecosystems will evolve over time. My third chapter presents one approach to building general theory for macroecological patterns out of steady state.

Understanding how macroecology changes over time has many important implications. It could allow us to connect changes in patterns to different types of disturbance, or other underlying ecology. Deviations from static patterns may also be useful in identifying when ecosystems are subject to disturbance. More robust theory could also predict how climate change will impact these systems over time, or predict how habitat loss will affect species extinctions. More broadly, macroecology may point us toward unifying principles and theory in ecology.

Future work will have to address most of these questions, as dynamic understanding of macroecological patterns is in its infancy. Given that we are at a time where humans are having more of an impact on the planet than ever before, understanding macroecology out of steady state will be increasingly important as the number of ecosystems around the world with significant disturbance continues to increase.

Bibliography

Amari, S. (2016). *Information Geometry and Its Applications* (Vol. 194). Springer Japan.

Arnold, T. B., & Emerson, J. W. (2011). Nonparametric Goodness-of-Fit Tests for Discrete Null Distributions. *The R Journal*, *3*(2), 34–39. https://doi.org/10.32614/RJ-2011-016

Arrhenius, O. (1921). Species and Area. *Journal of Ecology*, *9*(1), 95–99. https://doi.org/10.2307/2255763

Azaele, S., Suweis, S., Grilli, J., Volkov, I., Banavar, J. R., & Maritan, A. (2016). Statistical mechanics of ecological systems: Neutral theory and beyond. *Reviews of Modern Physics*, *88*(3), 035003. https://doi.org/10.1103/RevModPhys.88.035003

Bagchi, R., Henrys, P. A., Brown, P. E., Burslem, D. F. R. P., Diggle, P. J., Gunatilleke, C. V. S., Gunatilleke, I. A. U. N., Kassim, A. R., Law, R., Noor, S., & Valencia, R. L. (2011). Spatial patterns reveal negative density dependence and habitat associations in tropical trees. *Ecology*, *92*(9), 1723–1729. https://doi.org/10.1890/11-0335.1

Banavar, J. R., Maritan, A., & Volkov, I. (2010). Applications of the principle of maximum entropy: From physics to ecology. *Journal of Physics: Condensed Matter*, *22*(6), 063101. https://doi.org/10.1088/0953-8984/22/6/063101

Bertram, J., & Dewar, R. C. (2015). Combining mechanism and drift in community ecology: A novel statistical mechanics approach. *Theoretical Ecology*, *8*(4), 419–435. https://doi.org/10.1007/s12080-015-0259-7

Bliss, C. I., & Fisher, R. A. (1953). Fitting the Negative Binomial Distribution to Biological Data. *Biometrics*, *9*(2), 176–200. https://doi.org/10.2307/3001850

Borda-de-Água, L., Whittaker, R. J., Cardoso, P., Rigal, F., Santos, A. M. C., Amorim, I. R., Parmakelis, A., Triantis, K. A., Pereira, H. M., & Borges, P. A. V. (2017). Dispersal ability determines the scaling properties of species abundance distributions: A case study using arthropods from the Azores. *Scientific Reports*, *7*(1), 3899. https://doi.org/10.1038/s41598-017-04126-5

Borges, P. A. V., & Hortal, J. (2009). Time, area and isolation: Factors driving the diversification of Azorean arthropods. *Journal of Biogeography*, *36*(1), 178–191. https://doi.org/10.1111/j.1365-2699.2008.01980.x

Borges, P. A. V., Lobo, J. M., Azevedo, E. B. d., Gaspar, C. S., Melo, C., & Nunes, L. V. (2006). Invasibility and species richness of island endemic arthropods: A general model

of endemic vs. exotic species. *Journal of Biogeography*, *33*(1), 169–187. https://doi.org/10.1111/j.1365-2699.2005.01324.x

Borges, P. A. V., Vieira, V., Amorim, I. R., Bicudo, N., Fritzén, N., Gaspar, C., Heleno, R., Hortal, J., Lissner, J., Logunov, D., Machado, A., Marcelino, J., Meijer, S. S., Melo, C., Mendonça, E., Moniz, J., Pereira, F., Santos, A. M. C., Simões, A. M., & Torrão, E. (2010). List of arthropods (Arthropoda). In P. A. V. Borges, A. Costa, R. Cunha, R. Gabriel, V. Gonçalves, A. Martins, I. Melo, M. Parente, P. Raposeiro, P. Rodrigues, R. Santos, L. Silva, P. Vieira, & V. Vieira (Eds.), *A list of the terrestrial and marine biota from the Azores* (pp. 179–246). Principia.

Brown, C., Illian, J. B., & Burslem, D. F. R. P. (2016). Success of spatial statistics in determining underlying process in simulated plant communities. *Journal of Ecology*, *104*(1), 160–172. https://doi.org/10.1111/1365-2745.12493

Brown, C., Law, R., Illian, J. B., & Burslem, D. F. R. P. (2011). Linking ecological processes with spatial and non-spatial patterns in plant communities. *Journal of Ecology*, *99*(6), 1402–1414. https://doi.org/10.1111/j.1365-2745.2011.01877.x

Brown, J. H. (1995). *Macroecology*. University of Chicago Press.

Brown, J. H., Gillooly, J. F., Allen, A. P., Savage, V. M., & West, G. B. (2004). Toward a Metabolic Theory of Ecology. *Ecology*, *85*(7), 1771–1789. https://doi.org/10.1890/03-9000

Brummer, A. B., & Newman, E. A. (2019). Derivations of the Core Functions of the Maximum Entropy Theory of Ecology. *Entropy*, *21*(7), 712. https://doi.org/10.3390/e21070712

Brush, M., & Harte, J. (2021). Relating the Strength of Density Dependence and the Spatial Distribution of Individuals. *Frontiers in Ecology and Evolution*, *9*, 390. https://doi.org/10.3389/fevo.2021.691792

Cardoso, P., Aranda, S. C., Lobo, J. M., Dinis, F., Gaspar, C., & Borges, P. A. V. (2009). A spatial scale assessment of habitat effects on arthropod communities of an oceanic island. *Acta Oecologica*, *35*(5), 590–597. https://doi.org/10.1016/j.actao.2009.05.005

Carey, S., Harte, J., & Moral, R. D. (2006). Effect of community assembly and primary succession on the species-area relationship in disturbed ecosystems. *Ecography*, *29*(6), 866–872. https://doi.org/10.1111/j.2006.0906-7590.04712.x

Carson, W. P., Anderson, J., Leigh, E., & Schnitzer, S. A. (2008). Challenges Associated with Testing and Falsifying the Janzen–Connell Hypothesis: A Review and Critique. In W. P. Carson & S. A. Schnitzer (Eds.), *Tropical Forest Community Ecology* (pp. 210–241). Wiley-Blackwell.

Caticha, A. (2015). The basics of information geometry. *AIP Conference Proceedings*, *1641*(1), 15–26. https://doi.org/10.1063/1.4905960

Coleman, B. D. (1981). On random placement and species-area relations. *Mathematical Biosciences*, *54*(3-4), 191–215. https://doi.org/10.1016/0025-5564(81)90086-9

Comita, L. S., Queenborough, S. A., Murphy, S. J., Eck, J. L., Xu, K., Krishnadas, M., Beckman, N., & Zhu, Y. (2014). Testing predictions of the Janzen-Connell hypothesis: A meta-analysis of experimental evidence for distance- and density-dependent seed

and seedling survival (L. Gómez-Aparicio, Ed.). *Journal of Ecology*, *102*(4), 845–856. https://doi.org/10.1111/1365-2745.12232

Condit, R., Pérez, R., Aguilar, S., Lao, S., Foster, R., & Hubbell, S. (2019). Complete data from the Barro Colorado 50-ha plot: 423617 trees, 35 years. *Dryad Digital Repository*. https://doi.org/10.15146/5xcp-0d46

Condit, R. (1998). *Tropical Forest Census Plots: Methods and Results from Barro Colorado Island, Panama and a Comparison with Other Plots*. Springer-Verlag.

Condit, R., Ashton, P. S., Baker, P., Bunyavejchewin, S., Gunatilleke, S., Gunatilleke, N., Hubbell, S. P., Foster, R. B., Itoh, A., LaFrankie, J. V., Lee, H. S., Losos, E., Manokaran, N., Sukumar, R., & Yamakura, T. (2000). Spatial Patterns in the Distribution of Tropical Tree Species. *Science*, *288*(5470), 1414–1418. https://doi.org/10.1126/science.288.5470.1414

Conlisk, E., Bloxham, M., Conlisk, J., Enquist, B., & Harte, J. (2007). A New Class of Models of Spatial Distribution. *Ecological Monographs*, *77*(2), 269–284. https://doi.org/10.1890/06-0122

Conlisk, J., Conlisk, E., Kassim, A. R., Billick, I., & Harte, J. (2012). The shape of a species' spatial abundance distribution. *Global Ecology and Biogeography*, *21*(12), 1167–1178. https://doi.org/10.1111/j.1466-8238.2011.00755.x

Connell, J. (1971). On the role of natural enemies in preventing competitive exclusion in some marine animals and in rain forest trees. In P. J. d. Boer & G. R. Gradwell (Eds.), *Dynamics of populations: Proceedings of the Advanced Study Institute on Dynamics of numbers in populations, Oosterbeek, the Netherlands, 7-18 September 1970* (pp. 298–312). Pudoc.

Connolly, S. R., & Dornelas, M. (2011). Fitting and empirical evaluation of models for species abundance distributions. In A. E. Magurran & B. J. McGill (Eds.), *Biological Diversity: Frontiers in Measurement and Assessment*. Oxford University Press.

Díaz, S., Settele, J., Brondízio, E. S., Ngo, H. T., Agard, J., Arneth, A., Balvanera, P., Brauman, K. A., Butchart, S. H. M., Chan, K. M. A., Garibaldi, L. A., Ichii, K., Liu, J., Subramanian, S. M., Midgley, G. F., Miloslavich, P., Molnár, Z., Obura, D., Pfaff, A., … Zayas, C. N. (2019). Pervasive human-driven decline of life on Earth points to the need for transformative change. *Science*, *366*(6471). https://doi.org/10.1126/science.aax3100

Diggle, P. (2014). *Statistical analysis of spatial and spatio-temporal point patterns* (Third edition). CRC Press, Taylor & Francis Group.

Dornelas, M. (2010). Disturbance and change in biodiversity. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *365*(1558), 3719–3727. https://doi.org/10.1098/rstb.2010.0295

Dornelas, M., Moonen, A. C., Magurran, A. E., & Bàrberi, P. (2009). Species abundance distributions reveal environmental heterogeneity in modified landscapes. *Journal of Applied Ecology*, *46*(3), 666–672. https://doi.org/10.1111/j.1365-2664.2009.01640.x

Drakare, S., Lennon, J. J., & Hillebrand, H. (2006). The imprint of the geographical, evolutionary and ecological context on species-area relationships. *Ecology Letters*, *9*(2), 215–227. https://doi.org/10.1111/j.1461-0248.2005.00848.x

Elias, R. B., Gil, A., Silva, L., Fernández-Palacios, J. M., Azevedo, E. B., & Reis, F. (2016). Natural zonal vegetation of the Azores Islands: Characterization and potential distribution. *Phytocoenologia*, *46*(2), 107–123. https://doi.org/10.1127/phyto/2016/0132

Fahrig, L. (2003). Effects of Habitat Fragmentation on Biodiversity. *Annual Review of Ecology, Evolution, and Systematics*, *34*(1), 487–515. https://doi.org/10.1146/annurev.ecolsys.34.011802.132419

Fahrig, L. (2019). Habitat fragmentation: A long and tangled tale. *Global Ecology and Biogeography*, *28*(1), 33–41. https://doi.org/10.1111/geb.12839

Fattorini, S., Rigal, F., Cardoso, P., & Borges, P. A. V. (2016). Using species abundance distribution models and diversity indices for biogeographical analyses. *Acta Oecologica*, *70*, 21–28. https://doi.org/10.1016/j.actao.2015.11.003

Fischer, J., & Lindenmayer, D. B. (2007). Landscape modification and habitat fragmentation: A synthesis. *Global Ecology and Biogeography*, *16*(3), 265–280. https://doi.org/10.1111/j.1466-8238.2007.00287.x

Fisher, J. A. D., Frank, K. T., & Leggett, W. C. (2010). Dynamic macroecology on ecological time-scales. *Global Ecology and Biogeography*, *19*(1), 1–15. https://doi.org/10.1111/j.1466-8238.2009.00482.x

Fisher, R. A., Corbet, A. S., & Williams, C. B. (1943). The Relation Between the Number of Species and the Number of Individuals in a Random Sample of an Animal Population. *Journal of Animal Ecology*, *12*(1), 42–58. https://doi.org/10.2307/1411

Florencio, M., Cardoso, P., Lobo, J. M., Azevedo, E. B. d., & Borges, P. A. V. (2013). Arthropod assemblage homogenization in oceanic islands: The role of indigenous and exotic species under landscape disturbance. *Diversity and Distributions*, *19*(11), 1450–1460. https://doi.org/10.1111/ddi.12121

Foley, J. A., DeFries, R., Asner, G. P., Barford, C., Bonan, G., Carpenter, S. R., Chapin, F. S., Coe, M. T., Daily, G. C., Gibbs, H. K., Helkowski, J. H., Holloway, T., Howard, E. A., Kucharik, C. J., Monfreda, C., Patz, J. A., Prentice, I. C., Ramankutty, N., & Snyder, P. K. (2005). Global Consequences of Land Use. *Science*, *309*(5734), 570–574. https://doi.org/10.1126/science.1111772

Franzman, J., Brush, M., Umemura, K., Ray, C., Blonder, B., & Harte, J. (2021). Shifting macroecological patterns and static theory failure in a stressed alpine plant community. *Ecosphere*, *12*(6). https://doi.org/10.1002/ecs2.3548

Gaston, K. J., & Blackburn, T. M. (2000). *Pattern and process in macroecology*. Blackwell Science.

Gaston, K. J., Borges, P. A. V., He, F., & Gaspar, C. (2006). Abundance, spatial variance and occupancy: Arthropod species distribution in the Azores. *Journal of Animal Ecology*, *75*(3), 646–656. https://doi.org/10.1111/j.1365-2656.2006.01085.x

Gillespie, R. G., Claridge, E. M., & Roderick, G. K. (2008). Biodiversity dynamics in isolated island communities: Interaction between natural and human-mediated pro-

cesses. *Molecular Ecology*, *17*(1), 45–57. https://doi.org/10.1111/j.1365-294X.2007.03466.x

Gillespie, R. G., & Roderick, G. K. (2002). Arthropods on Islands: Colonization, Speciation, and Conservation. *Annual Review of Entomology*, *47*(1), 595–632. https://doi.org/10.1146/annurev.ento.47.091201.145244

Gouveia, S. F., Rubalcaba, J. G., Soukhovolsky, V., Tarasova, O., Barbosa, A. M., & Real, R. (2020). Ecophysics reload—exploring applications of theoretical physics in macroecology. *Ecological Modelling*, *424*, 109032. https://doi.org/10.1016/j.ecolmodel.2020.109032

Gouws, E. J., Gaston, K. J., & Chown, S. L. (2011). Intraspecific Body Size Frequency Distributions of Insects. *PLOS ONE*, *6*(3), e16606. https://doi.org/10.1371/journal.pone.0016606

Gray, J. S., Bjørgesæter, A., & Ugland, K. I. (2006). On plotting species abundance distributions. *Journal of Animal Ecology*, *75*(3), 752–756. https://doi.org/10.1111/j.1365-2656.2006.01095.x

Green, J., Harte, J., & Ostling, A. (2019). Data from: Species richness, endemism, and abundance patterns: Tests of two fractal models in a serpentine grassland. *Dryad Digital Repository*. https://doi.org/10.6078/D1MQ2V

Green, J. L., Harte, J., & Ostling, A. (2003). Species richness, endemism, and abundance patterns: Tests of two fractal models in a serpentine grassland. *Ecology Letters*, *6*(10), 919–928. https://doi.org/10.1046/j.1461-0248.2003.00519.x

Green, J. L., & Plotkin, J. B. (2007). A statistical theory for sampling species abundances. *Ecology Letters*, *10*(11), 1037–1045. https://doi.org/10.1111/j.1461-0248.2007.01101.x

Haegeman, B., Etienne, R. S., Rossberg, A. E. A. G., & McPeek, E. M. A. (2010). Entropy Maximization and the Spatial Distribution of Species. *The American Naturalist*, *175*(4), E74–E90. https://doi.org/10.1086/650718

Harms, K. E., Wright, S. J., Calderón, O., Hernández, A., & Herre, E. A. (2000). Pervasive density-dependent recruitment enhances seedling diversity in a tropical forest. *Nature*, *404*(6777), 493–495. https://doi.org/10.1038/35006630

Harte, J., Zillio, T., Conlisk, E., & Smith, A. B. (2008). Maximum Entropy and the State-Variable Approach to Macroecology. *Ecology*, *89*(10), 2700–2711. https://doi.org/10.1890/07-1369.1

Harte, J. (2011). *Maximum entropy and ecology : A theory of abundance, distribution, and energetics*. Oxford University Press.

Harte, J., & Kitzes, J. (2015). Inferring Regional-Scale Species Diversity from Small-Plot Censuses. *PLOS ONE*, *10*(2), e0117527–e0117527. https://doi.org/10.1371/journal.pone.0117527

Harte, J., & Newman, E. A. (2014). Maximum information entropy: A foundation for ecological theory. *Trends in Ecology & Evolution*, *29*(7), 384–389. https://doi.org/10.1016/j.tree.2014.04.009

Harte, J., Newman, E. A., & Rominger, A. J. (2017). Metabolic partitioning across individuals in ecological communities. *Global Ecology and Biogeography*, *1*(5), 993–997. https://doi.org/10.1111/geb.12621

Harte, J., Umemura, K., & Brush, M. (2021). DynaMETE: A hybrid MaxEnt-plus-mechanism theory of dynamic macroecology. *Ecology Letters*, *24*(5), 935–949. https://doi.org/10.1111/ele.13714

Harvey, B. J., Holzman, B. A., & Forrestel, A. B. (2014). Forest resilience following severe wildfire in a semi-urban national park. *Fremontia*, *42*(3), 14–18.

He, F., & Gaston, K. J. (2000). Estimating Species Abundance from Occurrence. *The American Naturalist*, *156*(5), 553–559. https://doi.org/10.1086/303403

He, F., & Gaston, K. J. (2003). Occupancy, Spatial Variance, and the Abundance of Species. *The American Naturalist*, *162*(3), 366–375. https://doi.org/10.1086/377190

Hill, J. K., & Hamer, K. C. (1998). Using species abundance models as indicators of habitat disturbance in tropical forests. *Journal of Applied Ecology*, *35*(3), 458–460. https://doi.org/10.1046/j.1365-2664.1998.00310.x

Hubbell, S. P., Foster, R. B., O'Brien, S. T., Harms, K. E., Condit, R., Wechsler, B., Wright, S. J., & Lao, S. L. d. (1999). Light-Gap Disturbances, Recruitment Limitation, and Tree Diversity in a Neotropical Forest. *Science*, *283*(5401), 554–557. https://doi.org/10.1126/science.283.5401.554

Hubbell, S., Condit, R., & Foster, R. (2005). BCI 50-ha plot data.

Hubbell, S. P. (2001). *The unified neutral theory of biodiversity and biogeography*. Princeton University Press.

Janzen, D. H. (1970). Herbivores and the Number of Tree Species in Tropical Forests. *The American Naturalist*, *104*(940), 501–528. https://doi.org/10.1086/282687

Jaynes, E. T. (1957). Information Theory and Statistical Mechanics. *Physical Review*, *106*(4), 620–630. https://doi.org/10.1103/PhysRev.106.620

Jaynes, E. T. (1982). On The Rationale of Maximum-Entropy Methods. *Proceedings of the IEEE*, *70*(9), 939–952. https://doi.org/10.1109/PROC.1982.12425

Kempton, R. A., & Taylor, L. R. (1974). Log-Series and Log-Normal Parameters as Diversity Discriminants for the Lepidoptera. *Journal of Animal Ecology*, *43*(2), 381–399. https://doi.org/10.2307/3371

Kitzes, J. (2019). Evidence for power-law scaling in species aggregation. *Ecography*, *42*(6), 1224–1225. https://doi.org/10.1111/ecog.04159

Kitzes, J., & Shirley, R. (2016). Estimating biodiversity impacts without field surveys: A case study in northern Borneo. *Ambio*, *45*(1), 110–119. https://doi.org/10.1007/s13280-015-0683-3

Kitzes, J., & Wilber, M. (2016). Macroeco: Reproducible ecological pattern analysis in Python. *Ecography*, *39*(4), 361–367. https://doi.org/10.1111/ecog.01905

Kitzes, J., Wilber, M., Lewis, C., & White, E. P. (2015). Jkitzes/macroeco.

Klein Goldewijk, K., Beusen, A., Doelman, J., & Stehfest, E. (2017). Anthropogenic land use estimates for the Holocene – HYDE 3.2. *Earth System Science Data*, *9*(2), 927–953. https://doi.org/10.5194/essd-9-927-2017

Law, R., Illian, J., Burslem, D. F. R. P., Gratzer, G., Gunatilleke, C. V. S., & Gunatilleke, I. a. U. N. (2009). Ecological information from spatial patterns of plants: Insights from point process theory. *Journal of Ecology*, *97*(4), 616–628. https://doi.org/10.1111/j.1365-2745.2009.01510.x

Lawton, J. H. (1999). Are There General Laws in Ecology? *Oikos*, *84*(2), 177–177. https://doi.org/10.2307/3546712

Leibold, M. A., & Chase, J. M. (2018). *Metacommunity ecology*. Princeton University Press.

Levin, S. A. (1992). The Problem of Pattern and Scale in Ecology: The Robert H. MacArthur Award Lecture. *Ecology*, *73*(6), 1943–1967. https://doi.org/10.2307/1941447

Macías-Hernández, N., Ramos, C., Domènech, M., Febles, S., Santos, I., Arnedo, M., Borges, P., Emerson, B., & Cardoso, P. (2020). A database of functional traits for spiders from native forests of the Iberian Peninsula and Macaronesia. *Biodiversity Data Journal*, *8*, e49159. https://doi.org/10.3897/BDJ.8.e49159

Magurran, A. E., & McGill, B. J. (2011). *Biological Diversity: Frontiers in Measurement and Assessment*. Oxford University Press.

Marbà, N., Duarte, C. M., & Agustí, S. (2007). Allometric scaling of plant life history. *Proceedings of the National Academy of Sciences*, *104*(40), 15777–15780.

Martins, I. S., Proença, V., & Pereira, H. M. (2014). The unusual suspect: Land use is a key predictor of biodiversity patterns in the Iberian Peninsula. *Acta Oecologica*, *61*, 41–50. https://doi.org/10.1016/j.actao.2014.10.005

Matthews, T. J., & Whittaker, R. J. (2014). Fitting and comparing competing models of the species abundance distribution: Assessment and prospect. *Frontiers of Biogeography*, *6*(2). https://doi.org/10.21425/F5FBG20607

Matthews, T. J., & Whittaker, R. J. (2015). REVIEW: On the species abundance distribution in applied ecology and biodiversity management. *Journal of Applied Ecology*, *52*(2), 443–454. https://doi.org/10.1111/1365-2664.12380

Maxwell, S. L., Fuller, R. A., Brooks, T. M., & Watson, J. E. M. (2016). Biodiversity: The ravages of guns, nets and bulldozers. *Nature*, *536*(7615), 143–145. https://doi.org/10.1038/536143a

McGill, B. J. (2010). Towards a unification of unified theories of biodiversity. *Ecology Letters*, *13*(5), 627–642. https://doi.org/10.1111/j.1461-0248.2010.01449.x

McGill, B. J., Chase, J. M., Hortal, J., Overcast, I., Rominger, A. J., Rosindell, J., Borges, P. A. V., Emerson, B. C., Etienne, R. S., Hickerson, M. J., Mahler, D. L., Massol, F., McGaughran, A., Neves, P., Parent, C., Patiño, J., Ruffley, M., Wagner, C. E., & Gillespie, R. (2019). Unifying macroecology and macroevolution to answer fundamental questions about biodiversity. *Global Ecology and Biogeography*, *28*(12), 1925–1936. https://doi.org/10.1111/geb.13020

McGill, B. J., Etienne, R. S., Gray, J. S., Alonso, D., Anderson, M. J., Benecha, H. K., Dornelas, M., Enquist, B. J., Green, J. L., He, F., Hurlbert, A. H., Magurran, A. E., Marquet, P. A., Maurer, B. A., Ostling, A., Soykan, C. U., Ugland, K. I., & White, E. P. (2007). Species abundance distributions: Moving beyond single prediction theo-

ries to integration within an ecological framework. *Ecology Letters*, *10*(10), 995–1015. https://doi.org/10.1111/j.1461-0248.2007.01094.x

McGlinn, D. J., Xiao, X., Kitzes, J., & White, E. P. (2015). Exploring the spatially explicit predictions of the Maximum Entropy Theory of Ecology. *Global Ecology and Biogeography*, *24*(6), 675–684. https://doi.org/10.1111/geb.12295

McGlinn, D. J., Xiao, X., & White, E. P. (2013). An empirical evaluation of four variants of a universal species–area relationship. *PeerJ*, *1*, e212. https://doi.org/10.7717/peerj.212

Muller-Landau, H. C., Condit, R. S., Chave, J., Thomas, S. C., Bohlman, S. A., Bunyavejchewin, S., Davies, S., Foster, R., Gunatilleke, S., Gunatilleke, N., Harms, K. E., Hart, T., Hubbell, S. P., Itoh, A., Kassim, A. R., LaFrankie, J. V., Lee, H. S., Losos, E., Makana, J.-R., ... Ashton, P. (2006). Testing metabolic ecology theory for allometric scaling of tree size, growth and mortality in tropical forests. *Ecology Letters*, *9*(5), 575–588. https://doi.org/10.1111/j.1461-0248.2006.00904.x

Münkemüller, T., Gallien, L., Pollock, L. J., Barros, C., Carboni, M., Chalmandrier, L., Mazel, F., Mokany, K., Roquet, C., Smyčka, J., Talluto, M. V., & Thuiller, W. (2020). Dos and don'ts when inferring assembly rules from diversity patterns. *Global Ecology and Biogeography*, *29*(7), 1212–1229. https://doi.org/10.1111/geb.13098

Newbold, T., Hudson, L. N., Contu, S., Hill, S. L. L., Beck, J., Liu, Y., Meyer, C., Phillips, H. R. P., Scharlemann, J. P. W., & Purvis, A. (2018). Widespread winners and narrow-ranged losers: Land use homogenizes biodiversity in local assemblages worldwide. *PLOS Biology*, *16*(12), e2006841. https://doi.org/10.1371/journal.pbio.2006841

Newbold, T., Hudson, L. N., Hill, S. L. L., Contu, S., Lysenko, I., Senior, R. A., Börger, L., Bennett, D. J., Choimes, A., Collen, B., Day, J., De Palma, A., Díaz, S., Echeverria-Londoño, S., Edgar, M. J., Feldman, A., Garon, M., Harrison, M. L. K., Alhusseini, T., ... Purvis, A. (2015). Global effects of land use on local terrestrial biodiversity. *Nature*, *520*(7545), 45–50. https://doi.org/10.1038/nature14324

Newman, E. A. (2019). Disturbance Ecology in the Anthropocene. *Frontiers in Ecology and Evolution*, *7*. https://doi.org/10.3389/fevo.2019.00147

Newman, E. A., Wilber, M. Q., Kopper, K. E., Moritz, M. A., Falk, D. A., McKenzie, D., & Harte, J. (2020). Disturbance macroecology: A comparative study of community structure metrics in a high-severity disturbance regime. *Ecosphere*, *11*(1), e03022. https://doi.org/10.1002/ecs2.3022

Norder, S. J., de Lima, R. F., de Nascimento, L., Lim, J. Y., Fernández-Palacios, J. M., Romeiras, M. M., Elias, R. B., Cabezas, F. J., Catarino, L., Ceríaco, L. M., Castilla-Beltrán, A., Gabriel, R., de Sequeira, M. M., Rijsdijk, K. F., Nogué, S., Kissling, W. D., van Loon, E. E., Hall, M., Matos, M., & Borges, P. A. (2020). Global change in microcosms: Environmental and societal predictors of land cover change on the Atlantic Ocean Islands. *Anthropocene*, *30*, 100242. https://doi.org/10.1016/j.ancene.2020.100242

O'Dwyer, J. P., & Chisholm, R. (2014). A mean field model for competition: From neutral ecology to the Red Queen. *Ecology Letters*, *17*(8), 961–969. https://doi.org/10.1111/ele.12299

Ostling, A., Harte, J., Green, J. L., & Kinzig, A. P. (2004). Self-Similarity, the Power Law Form of the Species-Area Relationship, and a Probability Rule: A Reply to Maddux. *The American Naturalist*, *163*(4), 627–633. https://doi.org/10.1086/382663

Pereira, H. M., Navarro, L. M., & Martins, I. S. (2012). Global Biodiversity Change: The Bad, the Good, and the Unknown. *Annual Review of Environment and Resources*, *37*(1), 25–50. https://doi.org/10.1146/annurev-environ-042911-093511

Pessoa, P. (2021). Legendre transformation and information geometry for the maximum entropy theory of ecology. *arXiv*. https://doi.org/2103.11230

Pessoa, P., Costa, F. X., & Caticha, A. (2021). Entropic Dynamics on Gibbs Statistical Manifolds. *Entropy*, *23*(5), 494. https://doi.org/10.3390/e23050494

Pigolotti, S., Cencini, M., Molina, D., & Muñoz, M. A. (2018). Stochastic spatial models in ecology: A statistical physics approach. *Journal of Statistical Physics*, *172*(1), 44–73. https://doi.org/10.1007/s10955-017-1926-4

Pimm, S. L., Jenkins, C. N., Abell, R., Brooks, T. M., Gittleman, J. L., Joppa, L. N., Raven, P. H., Roberts, C. M., & Sexton, J. O. (2014). The biodiversity of species and their rates of extinction, distribution, and protection. *Science*, *344*(6187). https://doi.org/10.1126/science.1246752

Plotkin, J. B., Potts, M. D., Leslie, N., Manokaran, N., Lafrankie, J., & Ashton, P. S. (2000). Species-area Curves, Spatial Aggregation, and Habitat Specialization in Tropical Forests. *Journal of Theoretical Biology*, *207*(1), 81–99. https://doi.org/10.1006/jtbi.2000.2158

Preston, F. W. (1948). The Commonness, And Rarity, of Species. *Ecology*, *29*(3), 254–283. https://doi.org/10.2307/1930989

Rigal, F., Cardoso, P., Lobo, J. M., Triantis, K. A., Whittaker, R. J., Amorim, I. R., & Borges, P. A. V. (2018). Functional traits of indigenous and exotic ground-dwelling arthropods show contrasting responses to land-use change in an oceanic island, Terceira, Azores. *Diversity and Distributions*, *24*(1), 36–47. https://doi.org/10.1111/ddi.12655

Rigal, F., Whittaker, R. J., Triantis, K. A., & Borges, P. A. V. (2013). Integration of non-indigenous species within the interspecific abundance–occupancy relationship. *Acta Oecologica*, *48*, 69–75. https://doi.org/10.1016/j.actao.2013.02.003

Rominger, A. J., Goodman, K. R., Lim, J. Y., Armstrong, E. E., Becking, L. E., Bennett, G. M., Brewer, M. S., Cotoras, D. D., Ewing, C. P., Harte, J., Martinez, N. D., O'Grady, P. M., Percy, D. M., Price, D. K., Roderick, G. K., Shaw, K. L., Valdovinos, F. S., Gruner, D. S., & Gillespie, R. G. (2016). Community assembly on isolated islands: Macroecology meets evolution. *Global Ecology and Biogeography*, *25*(7), 769–780. https://doi.org/10.1111/geb.12341

Rosenzweig, M. L. (1995). *Species diversity in space and time*. Cambridge University Press.

Shannon, C. E., & Weaver, W. (1949). *The mathematical theory of communication*.

Supp, S. R., Xiao, X., Ernest, S. K. M., & White, E. P. (2012). An experimental test of the response of macroecological patterns to altered species interactions. *Ecology*, *93*(12), 2505–2511. https://doi.org/10.1890/12-0370.1

Thibault, K. M., White, E. P., Hurlbert, A. H., & Ernest, S. K. M. (2011). Multimodality in the individual size distributions of bird communities. *Global Ecology and Biogeography*, *20*(1), 145–153. https://doi.org/10.1111/j.1466-8238.2010.00576.x

Thomas, C. D., Cameron, A., Green, R. E., Bakkenes, M., Beaumont, L. J., Collingham, Y. C., Erasmus, B. F. N., Siqueira, M. F. d., Grainger, A., Hannah, L., Hughes, L., Huntley, B., Jaarsveld, A. S. v., Midgley, G. F., Miles, L., Ortega-Huerta, M. A., Peterson, A. T., Phillips, O. L., & Williams, S. E. (2004). Extinction risk from climate change. *Nature*, *427*(6970), 145–148. https://doi.org/10.1038/nature02121

Triantis, K. A., Hortal, J., Amorim, I., Cardoso, P., Santos, A. M. C., Gabriel, R., & Borges, P. A. V. (2012). Resolving the Azorean knot: A response to Carine & Schaefer (2010). *Journal of Biogeography*, *39*(6), 1179–1184. https://doi.org/10.1111/j.1365-2699.2011.02623.x

Turner, M. G. (2010). Disturbance and landscape dynamics in a changing world. *Ecology*, *91*(10), 2833–2849. https://doi.org/10.1890/10-0097.1

Ulrich, W., Ollik, M., & Ugland, K. I. (2010). A meta-analysis of species–abundance distributions. *Oikos*, *119*(7), 1149–1155. https://doi.org/10.1111/j.1600-0706.2009.18236.x

Vellend, M. (2016). *The theory of ecological communities*. Princeton University Press.

Volkov, I., Banavar, J. R., He, F., Hubbell, S. P., & Maritan, A. (2005). Density dependence explains tree species abundance and diversity in tropical forests. *Nature*, *438*(7068), 658–661. https://doi.org/10.1038/nature04030

West, G. B., Brown, J. H., & Enquist, B. J. (2001). A general model for ontogenetic growth. *Nature*, *413*(6856), 628–631. https://doi.org/10.1038/35098076

White, E. P., Thibault, K. M., & Xiao, X. (2012). Characterizing species abundance distributions across taxa and ecosystems using a simple maximum entropy model. *Ecology*, *93*(8), 1772–1778. https://doi.org/10.1890/11-2177.1

Wiegand, T., He, F., & Hubbell, S. P. (2013). A systematic comparison of summary characteristics for quantifying point patterns in ecology. *Ecography*, *36*(1), 92–103. https://doi.org/10.1111/j.1600-0587.2012.07361.x

Wiegand, T., & Moloney, K. A. (2013). *Handbook of spatial point-pattern analysis in ecology*. CRC Press, Taylor & Francis Group.

Wilber, M. Q., Kitzes, J., & Harte, J. (2015). Scale collapse and the emergence of the power law species–area relationship. *Global Ecology and Biogeography*, *24*(8), 883–895. https://doi.org/10.1111/geb.12309

Williamson, M., & Gaston, K. J. (2005). The lognormal distribution is not an appropriate null hypothesis for the species–abundance distribution. *Journal of Animal Ecology*, *74*(3), 409–422. https://doi.org/10.1111/j.1365-2656.2005.00936.x

Xiao, X., Mcglinn, D. J., & White, E. P. (2015). A Strong Test of the Maximum Entropy Theory of Ecology. *The American Naturalist*, *185*(3), E70–E80. https://doi.org/10.5061/dryad.5fn46

Zillio, T., & He, F. (2010). Modeling spatial aggregation of finite populations. *Ecology*, *91*(12), 3698–3706. https://doi.org/10.1890/09-2233.1

Appendix A

Appendix for Chapter 1 – "The influence of land use on arthropod macroecology in the Azores"

A.1 Community level analysis

We feel that our analysis treating the transects as replicates across land use is stronger, particularly when comparing data to METE. This is because METE predictions are made within a single community and aggregating data over disparate locations, even of the same land use, may create a mismatch between the theory expectations and the aggregated data. Additionally, the number of species scales differently when summing multiple small patches than in a large patch of comparable area.

Despite that, we present the analysis at the community level here. In this case, all transects with the same land use are aggregated together, and we compare that empirical data to the METE prediction made with the total number of species and individuals for that land use. The mean least squared error for the SAD and MRDI across land uses is shown in Fig. A.1. These results are similar to those obtained when the transects are analyzed individually, though note here that the MRDI is the worse fit at the intensive pasture rather than the semi-natural pasture. Another difference is that the MRDI is comparatively better fit than the SAD at the forest sites. This is primarily as the SAD is significantly worse fit at the community level. The semi-natural pasture is still the only site that is poorly fit by both metrics. This fits with our interpretation in the main text that this site is the most poorly described by METE. The SAD results in particularly are very similar when analyzed at the community level.

We also show the empirical rank ordered SADs along with the corresponding METE predictions in Fig. A.2. Note that again the pasture sites are characterized by a few very abundant species, and METE under predicts the number of singletons across sites.

The empirical rank ordered MRDIs along with the corresponding METE predictions in

Fig. A.3. Again here, METE over predicts the metabolic rate of the highest metabolic rate individuals. This is particularly unsurprising here, as METE will predict higher metabolic rate individuals in larger ecosystems, which we have created here by aggregating across transects. It is likely that the maximum size of arthropods is more constrained at the transect level than at this larger community level.

Finally, we show the mean least squared error across land uses for indigenous and introduced species separately in Fig. A.4, and the corresponding rank ordered SADs in Fig. A.5. Here again we see that the highly abundant species at the pasture sites are introduced, and the fit for introduced species is much worse than for indigenous species at the semi-natural pasture. One difference with this community level analysis is that at both forest sites, the introduced species are better fit by METE than the indigenous species. We see this to some extent in the transect level analysis at the exotic forest, but it is more obvious here. We note that the overall fit for the community level analysis is worse across all land uses when compared to the transect level analysis, except for the intensive pasture.

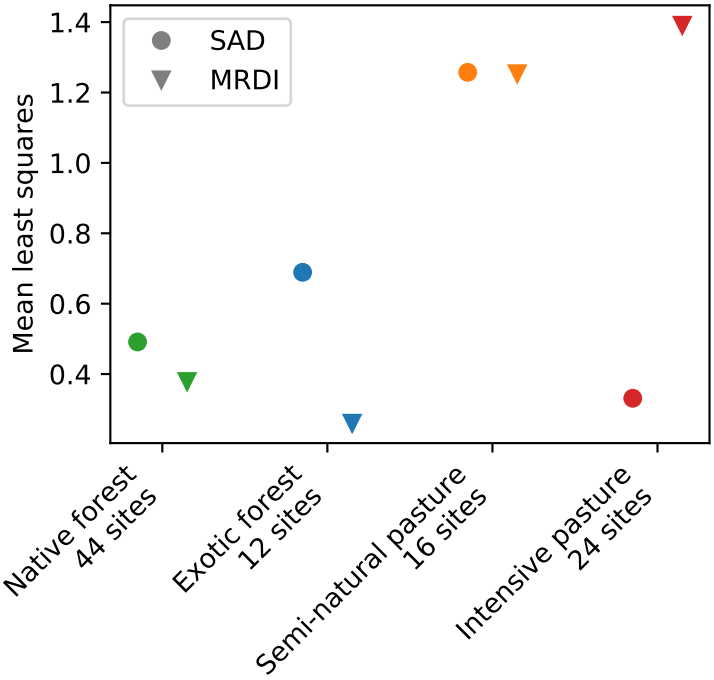


Figure A.1: The mean of the the mean least squared error for the SAD and the MRDI across land uses when transects are aggregated rather than analyzed individually. There are no error bars as there are no replicates when the data are analyzed this way.

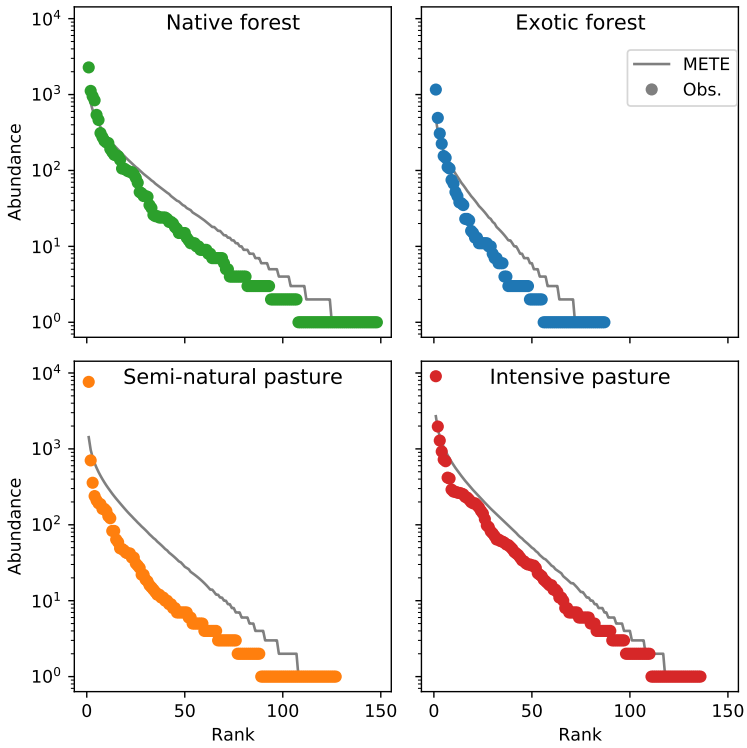


Figure A.2: Aggregated rank ordered SADs by land use. The solid line is the METE prediction and the points are observed rank abundance.

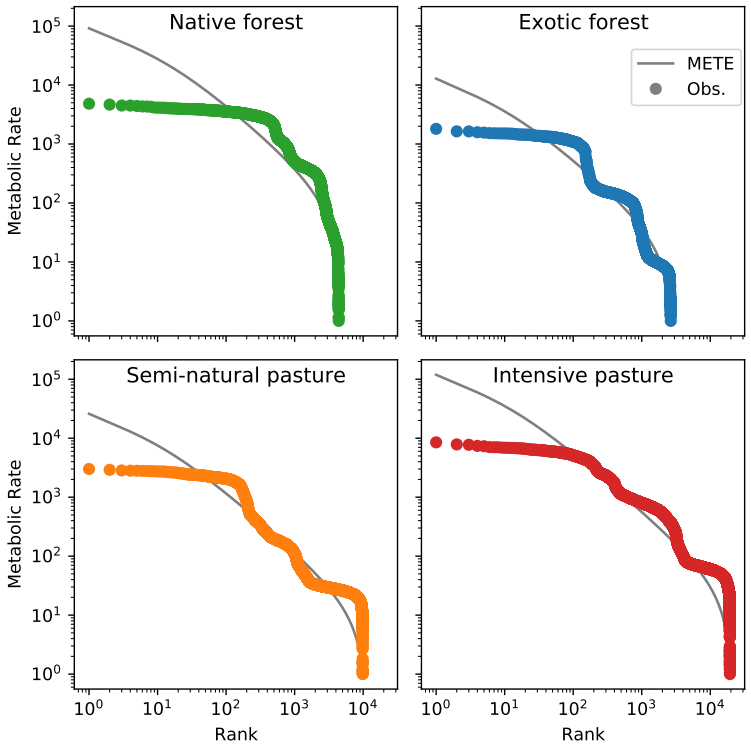


Figure A.3: Aggregated rank ordered MRDIs compared to METE predictions by land use. The empirical curve here is simulated from assuming a mean-variance relationship and adding variance to the mean body mass for each species.

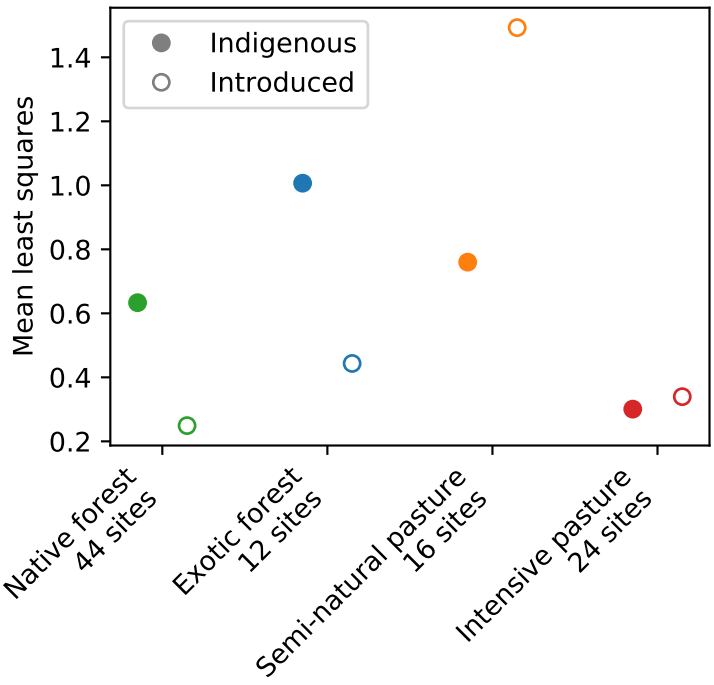
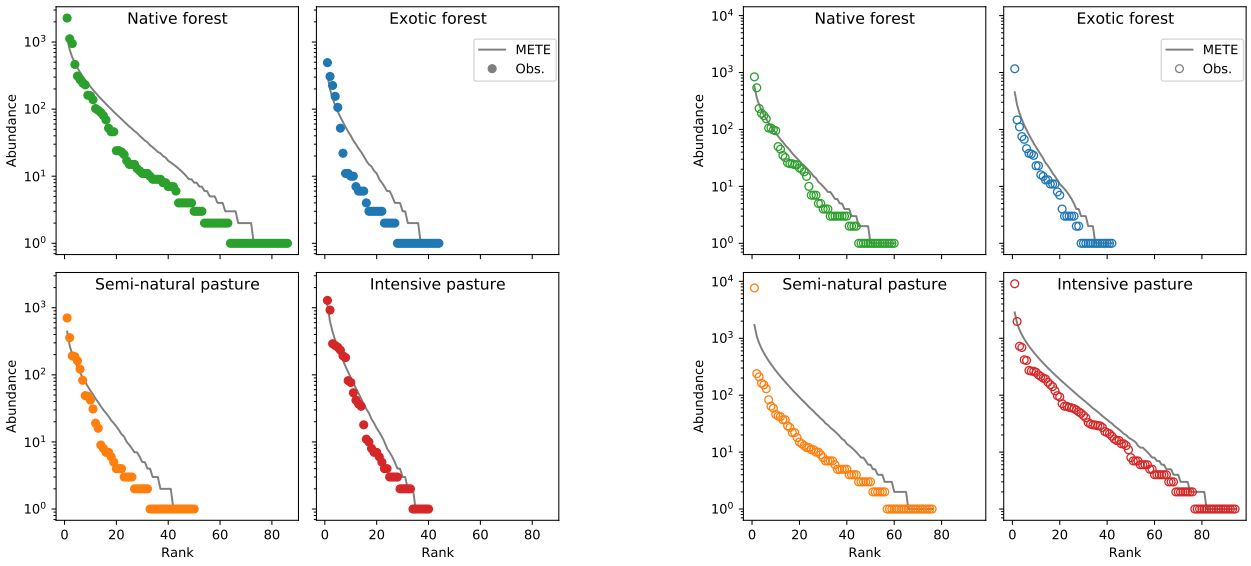


Figure A.4: Mean and its standard error for the mean least squares of the SAD for indigenous (closed circles) and introduced species (open circles) across land uses. Here, we have aggregated all transects rather than analyzed them individually, and so there are no error bars.



(a) Indigenous

(b) Introduced

Figure A.5: Aggregated rank ordered species abundance distributions for different land use, where species have been categorized and indigenous (filled circles) or introduced (open circles). Mete predictions are the solid line in each case.

A.2 Kolmogorov-Smirnov test

As a second goodness of fit test to check our results using mean least squares, we use the Kolmogorov-Smirnov test comparing the empirical CDF and the METE predicted CDF. For the SAD, the KS test must be adapted as the distribution is discrete with many duplicate values (ie. many singletons). We made use of the R package DGOF, which implements the KS test for discrete distributions. This is not an issue for the MRDI as it does not have repeated values. We plot the mean and standard error of the test statistic D_{KS} for both the SAD and MRDI for each land use in Fig. A.6.

The results here are comparable to those obtained with mean least squares. The MRDI is worse fit than the SAD, and the semi-intensive pasture is the worst fit for both the SAD and the MRDI. The intensive pasture results are also similar as it is among the best fit for the SAD, and intermediately well fit for the MRDI. That these results are comparable to the mean least squares results is important, as our conclusions about how well METE describes the data across land use do not appear to depend on the goodness of fit test itself.

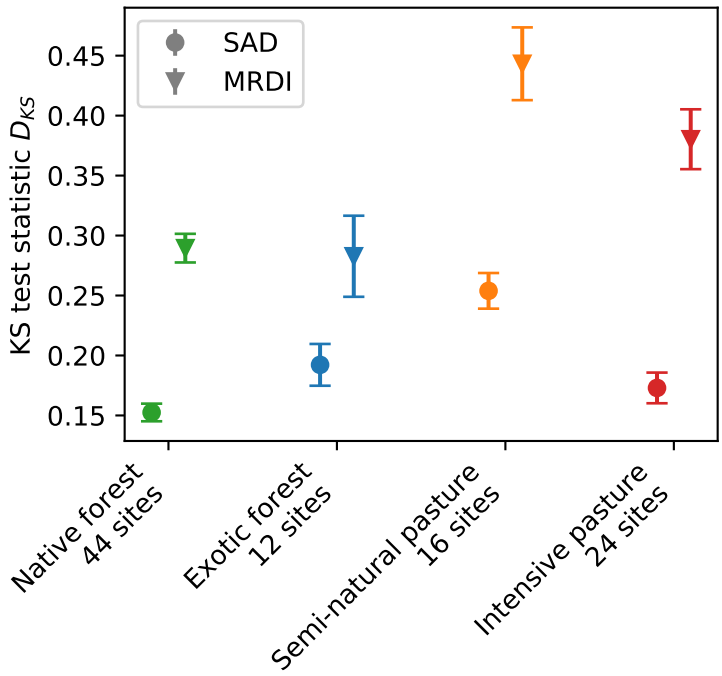


Figure A.6: The mean and standard error of the KS test statistic D_{KS} for the SAD and MRDI across transects for each land use.

A.3 Intraspecific body mass variation

For select species, we have individual measurements of multiple individuals within a species. We can use this intraspecific body length variation to estimate the variance in intraspecific body mass using the same scaling relationship. We can then convert this into a variance in metabolic rate, which we can use to reconstruct more realistic metabolic rate distributions.

We have individual body length measurements for 26 of the 115 species of Coleoptera (beetles) in the data, and for 26 of the 41 species of Araneae (spiders). These are the two most common orders in the data, which is where it is most important to have an idea of intraspecific variation in body mass. For less abundant species, this variation matters less as there are only a few of each species present in the data. We use the Coleoptera results for all other orders in the dataset. Note that this data actually has separate body length values for female and male spiders, given that sexual dimorphism is common. However, in the larger dataset we have only average body length, and therefore we will use only the overall mean and variance for spiders without separating by sex.

We plot the body mass distributions for the four most abundant species of Coleoptera and Araneae in this dataset in Fig. A.7, overlayed with the best fit normal distributions. We can see here that the Araneae species appear more bimodal, but the number of data points is quite small, and again this finer resolution is not available in the full data set.

We plot the relationship between the log of the mean and the log of the variance for both beetles and spiders in Fig. A.8. The slopes, intercepts, and R^2 correlation coefficient values are shown in Table A.1. We use these values to simulate variation in body mass for all species in our dataset and to reconstruct empirical MRDIs.

Order	Slope	Intercept	R^2
Coleoptera (beetles)	1.99 ± 0.12	-1.24	0.925
Araneae (spiders)	2.22 ± 0.13	-1.15	0.919

Table A.1: Results from the regression of $\log_{10}$ of variance versus $\log_{10}$ of mean body mass for data that includes intraspecific body mass variation for both Coleoptera and Araneae.

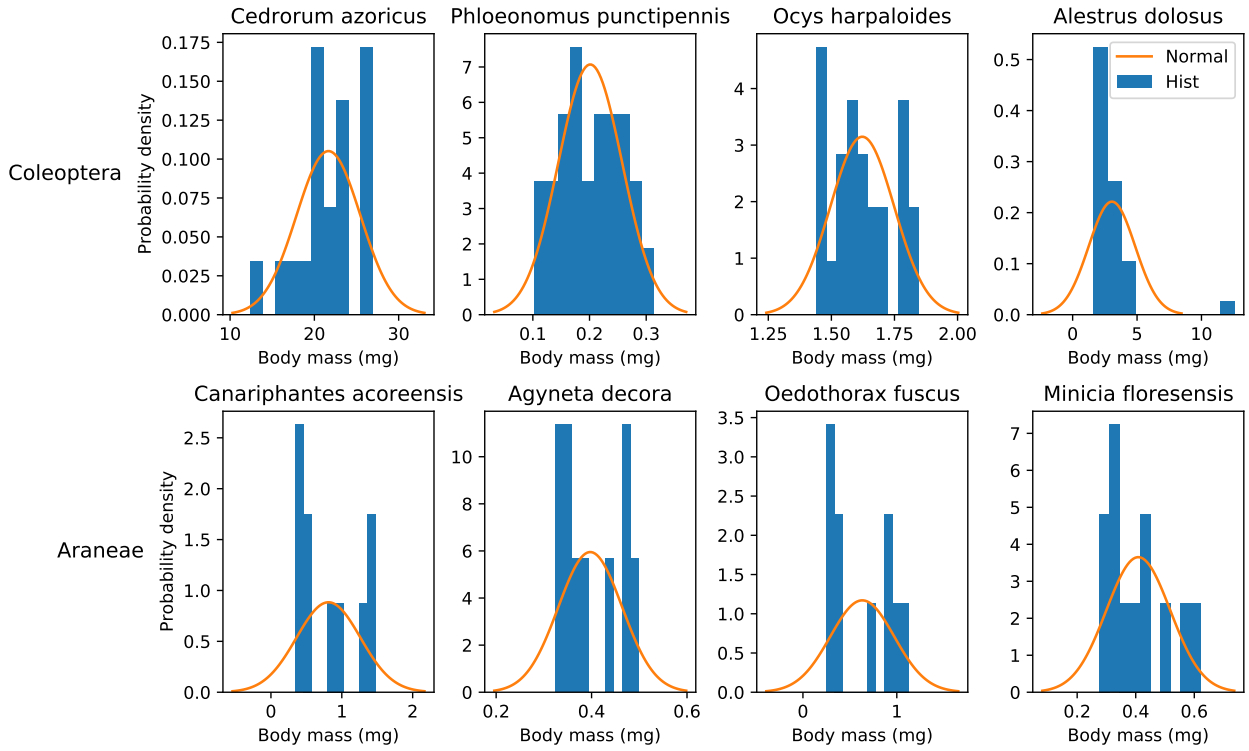


Figure A.7: Histograms and best fit normal distributions for the four most abundant species of Coleoptera and Araneae present in the data that includes intraspecific variation.

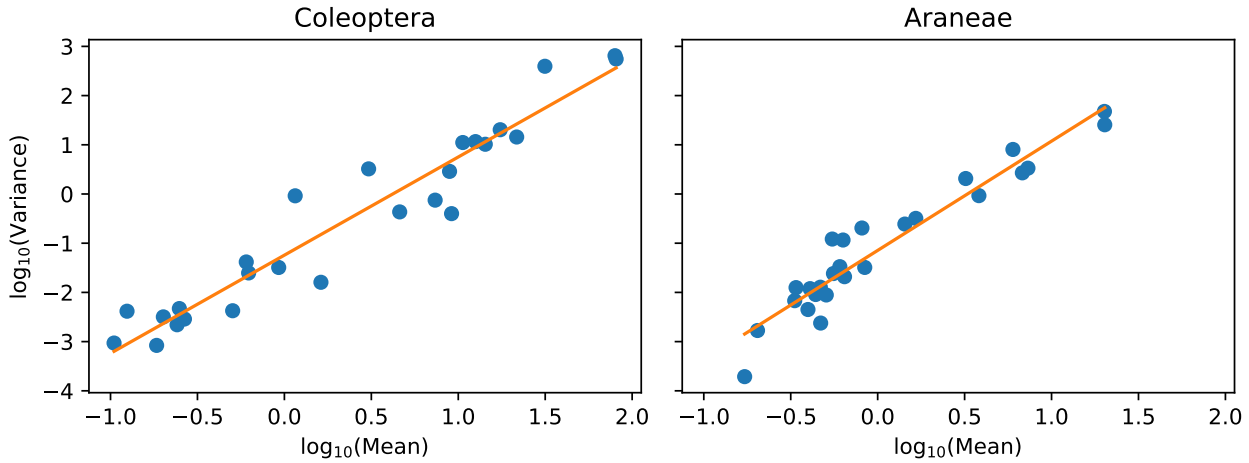


Figure A.8: $\log_{10}$ of variance versus $\log_{10}$ of mean body mass for data that includes intraspecific body mass variation for both Coleoptera and Araneae.

A.4 SAR comparison of number of species

In addition to the comparison of slopes in the main text, we can compare the predicted number of species at each scale directly. The mean least squares across transects at each land use are shown in Fig. A.9, again as the mean at the land use with its standard error. The direct plots of $\log(S)$ versus $\log(A)$ do not show the difference between the theoretical predictions and the observed values very well as the differences are relatively small, so we instead show the residuals of $\log(S_0)$ for each transect in Fig. A.10. Note that as mentioned in Methods, the largest scale corresponds exactly in all cases, and so we really only have 7 points of prediction in this case. Here again we see that METE over predicts S_0 at smaller scales for the pasture sites, and the residuals are relatively randomly distributed for the forest sites.

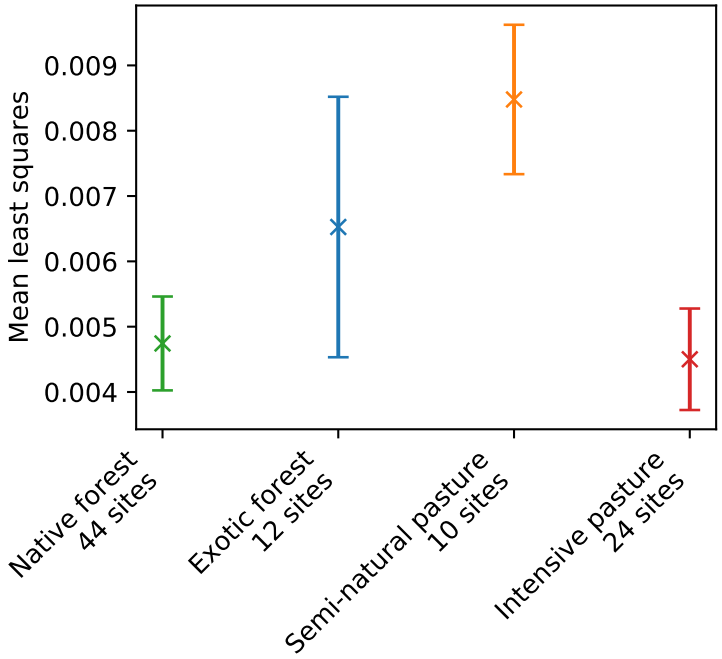


Figure A.9: The mean and its standard error for the mean least squares of the predicted number of species for each transect, organized by land use.

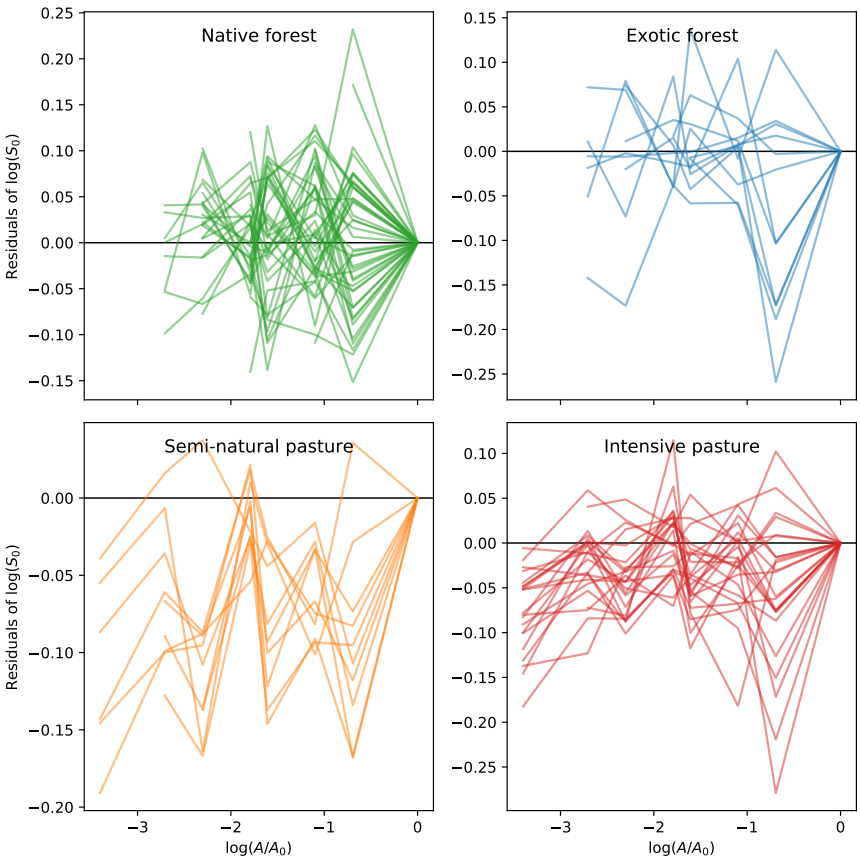


Figure A.10: The observed minus the predicted $\log(S_0)$ at each scale $\log(A/A_0)$. Each line traces out a single transect.

A.5 SADs at each transect

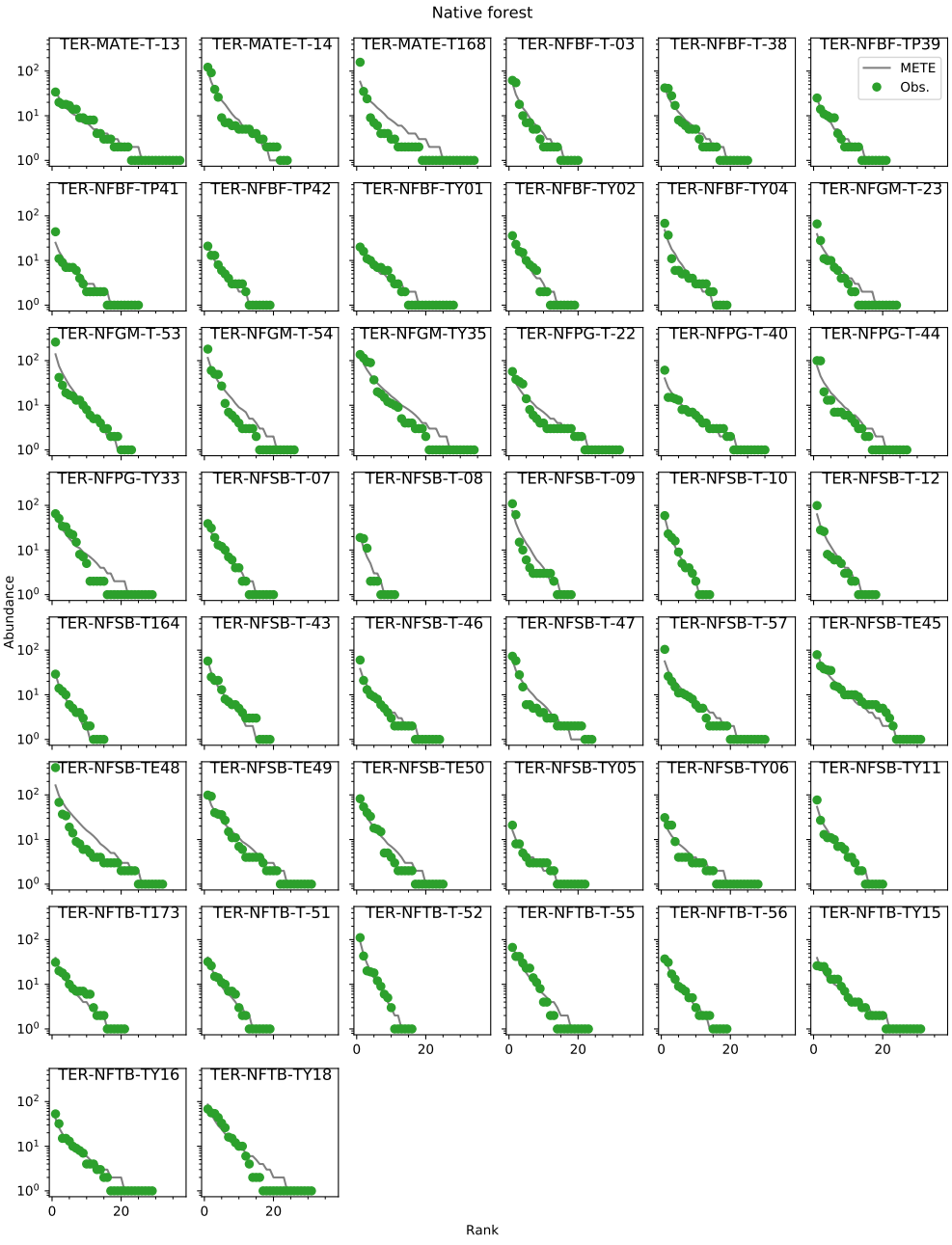


Figure A.11: The rank ordered SAD at each transect in the native forest.

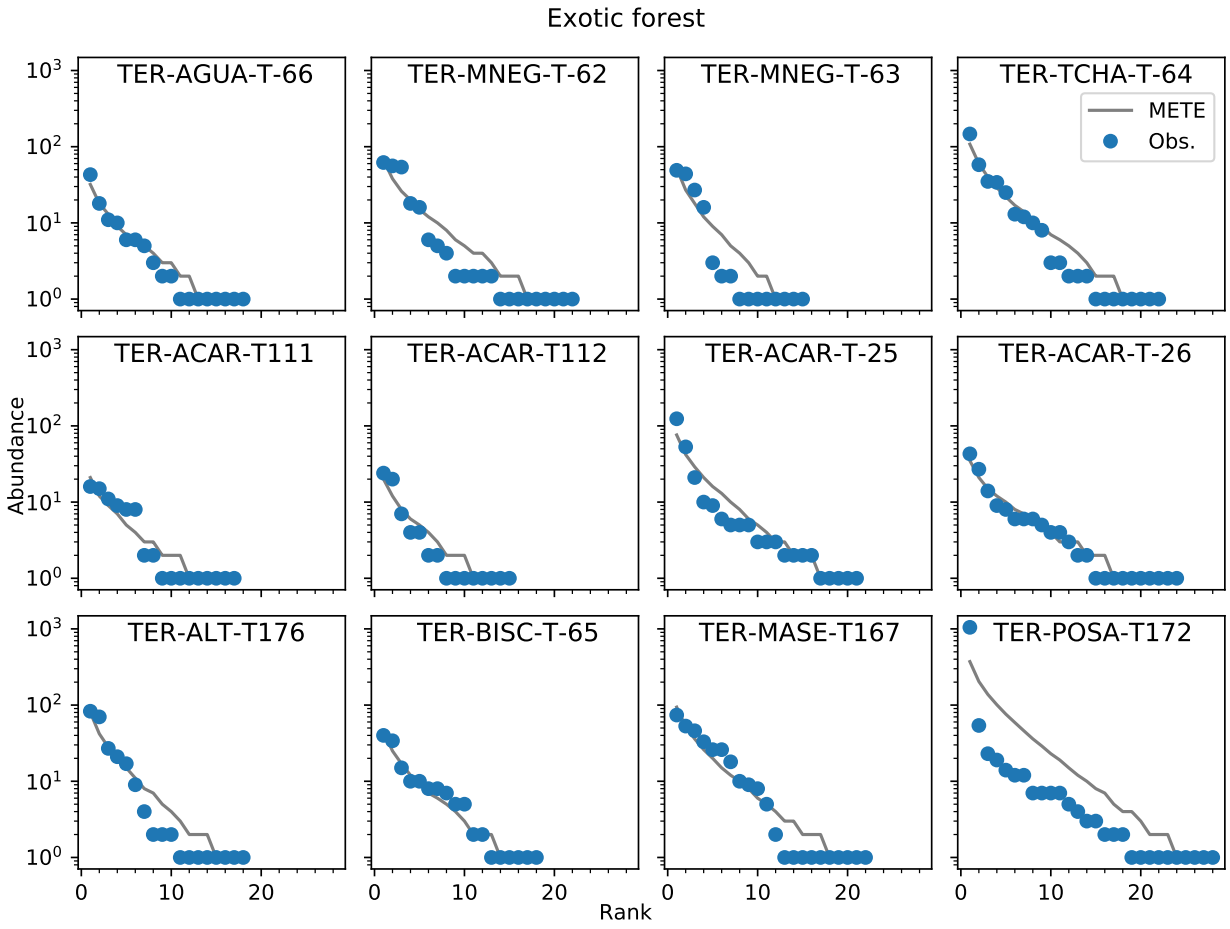


Figure A.12: The rank ordered SAD at each transect in the exotic forest.

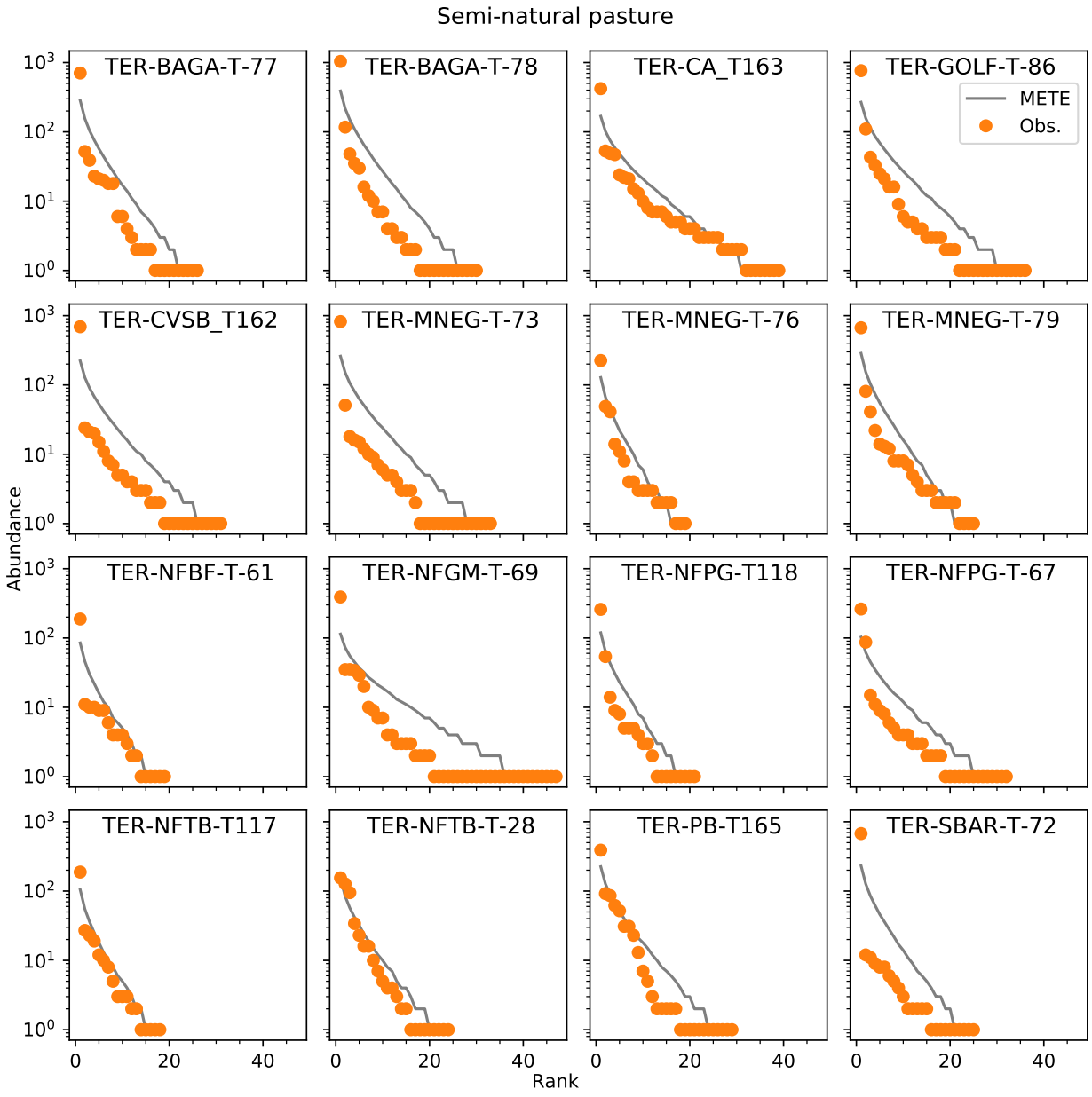


Figure A.13: The rank ordered SAD at each transect in the semi-natural pasture.

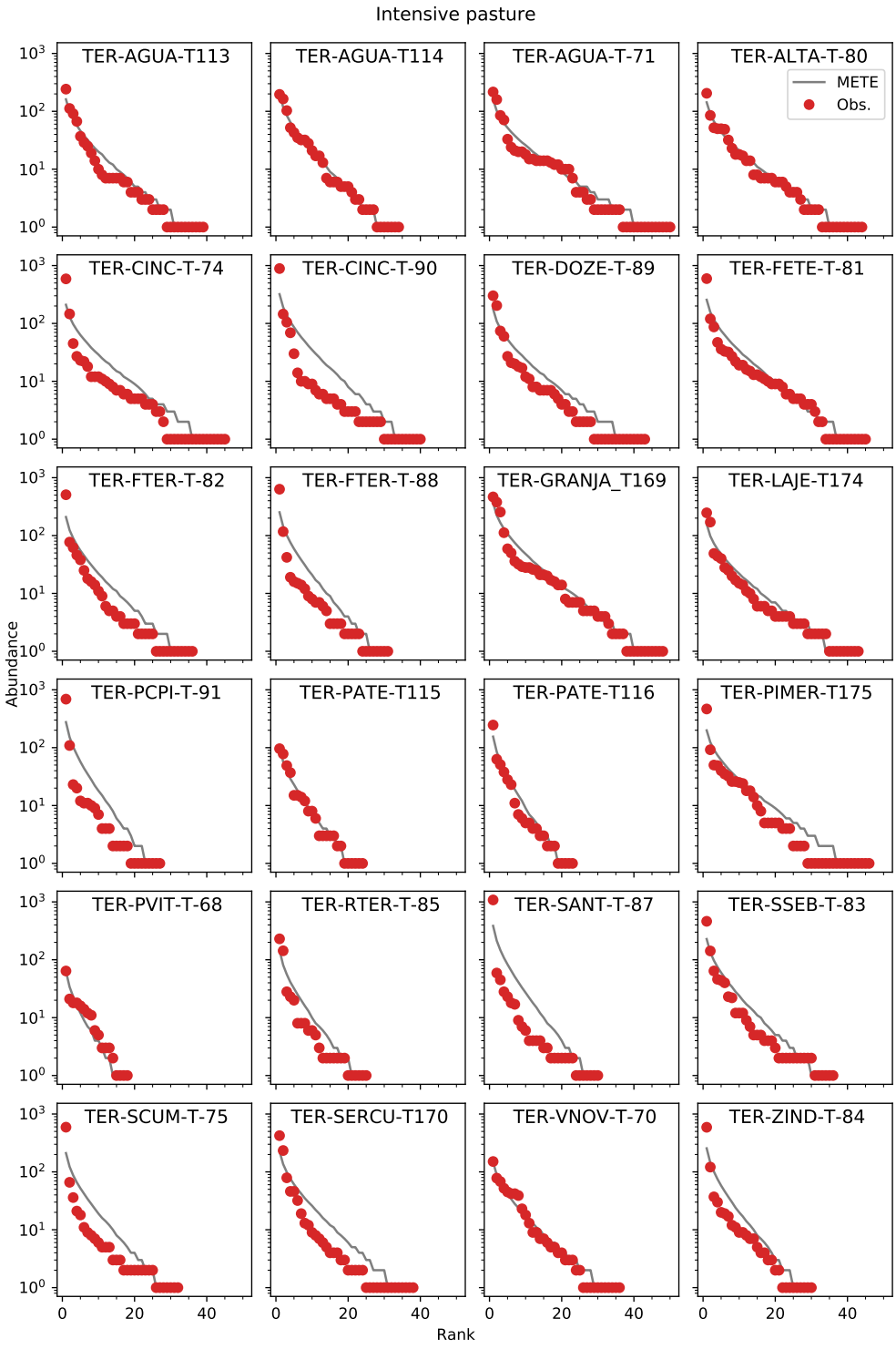


Figure A.14: The rank ordered SAD at each transect in the intensive pasture.

A.6 MRDIs at each transect

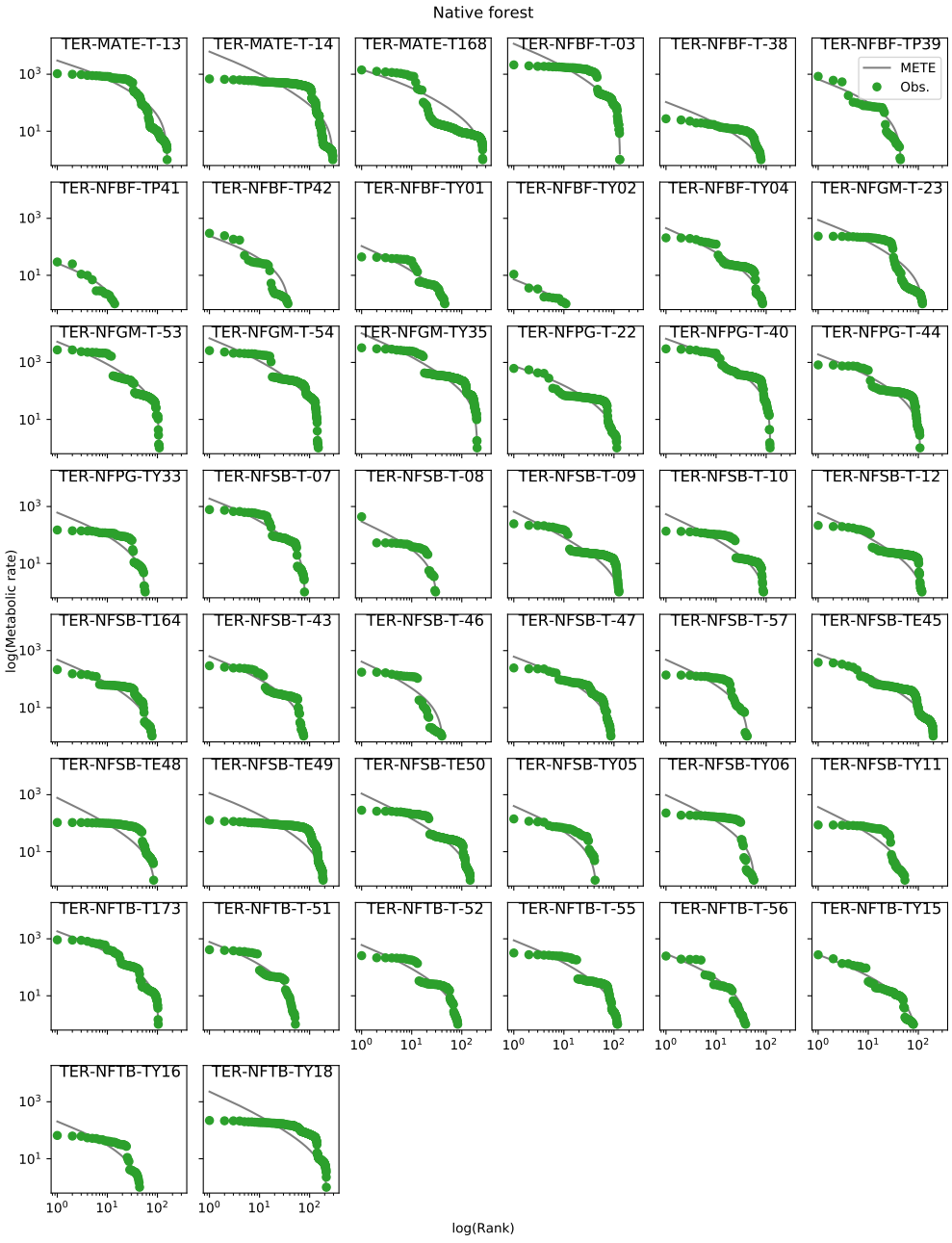


Figure A.15: The rank ordered MRDI at each transect in the native forest.

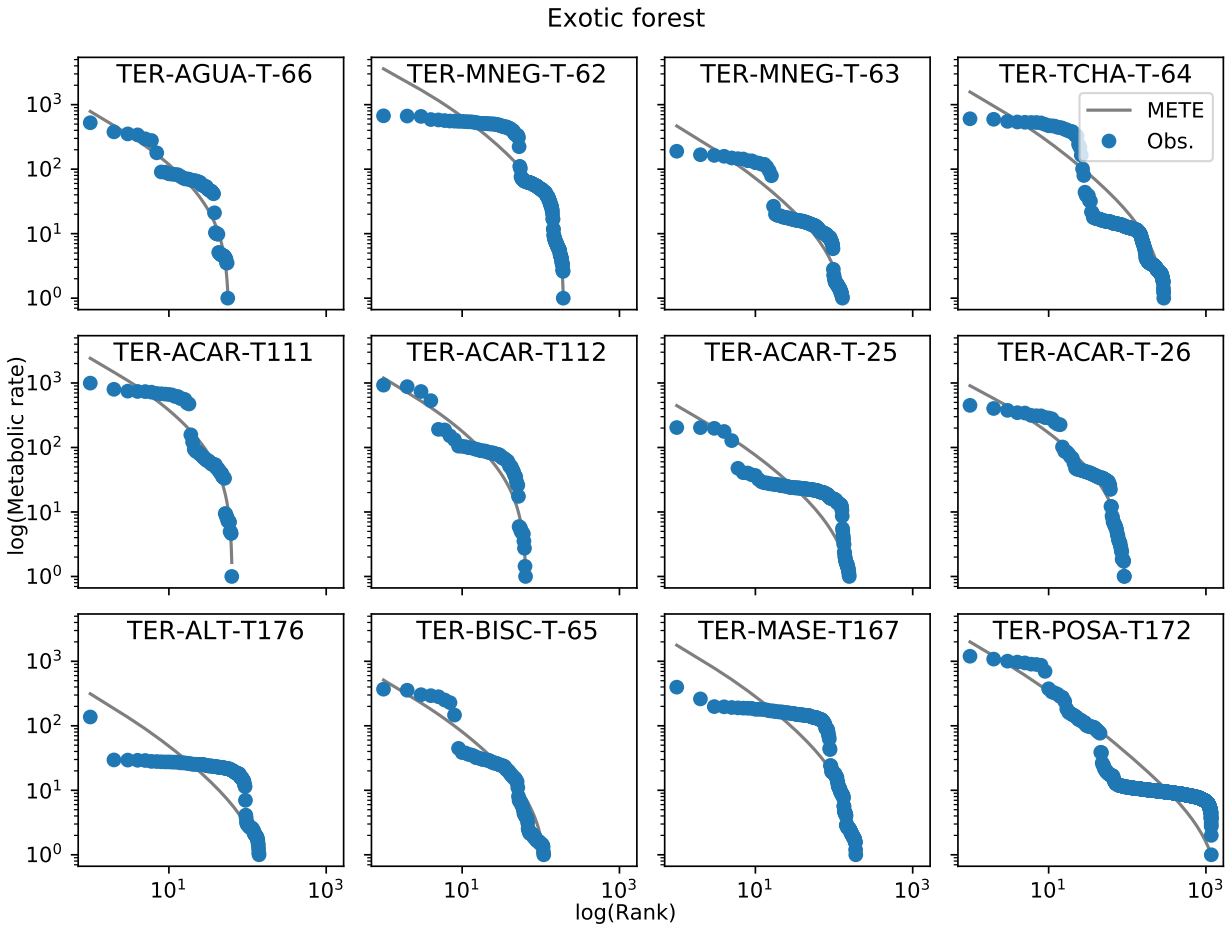


Figure A.16: The rank ordered MRDI at each transect in the exotic forest.

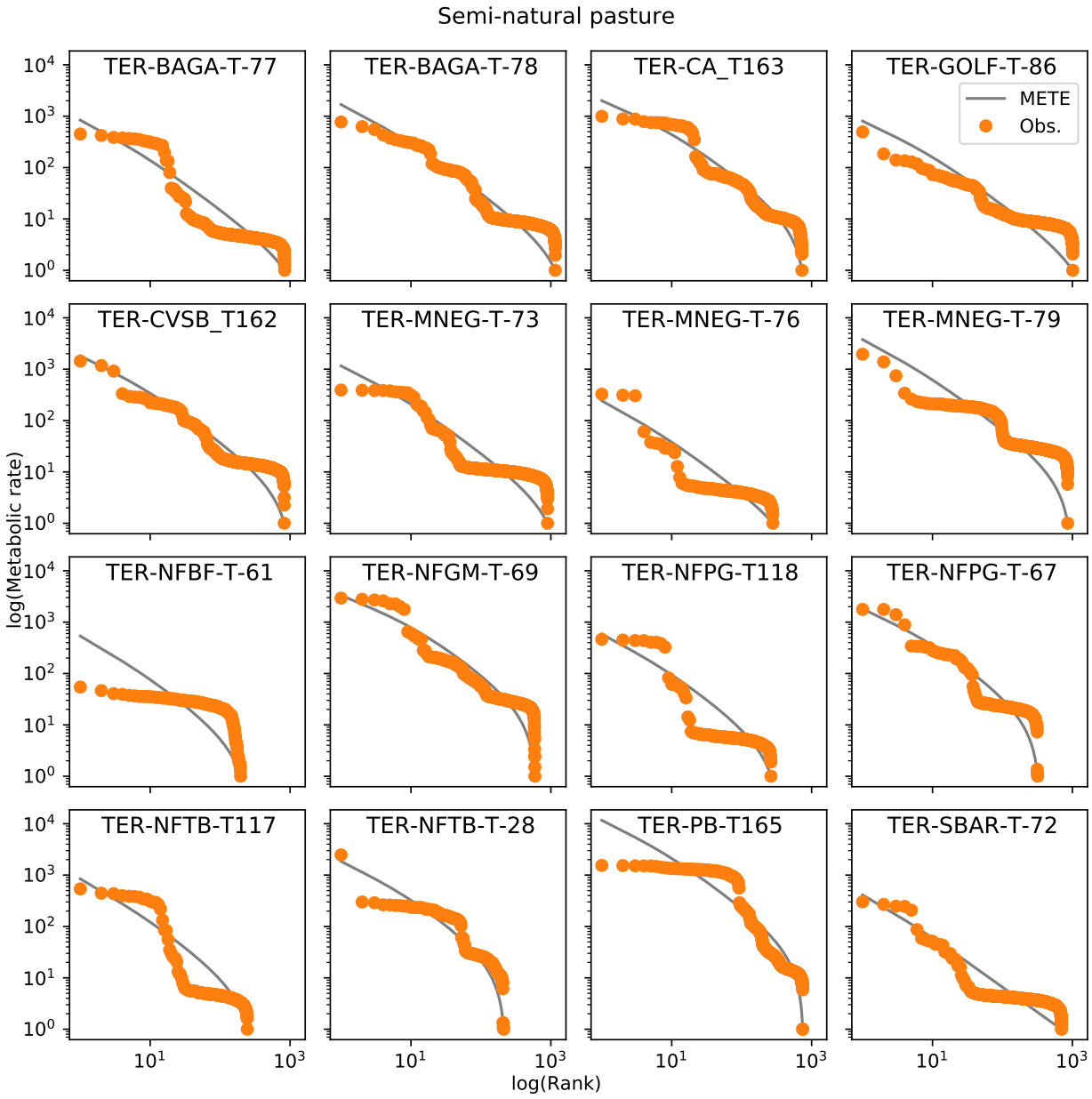


Figure A.17: The rank ordered MRDI at each transect in the semi-natural pasture.

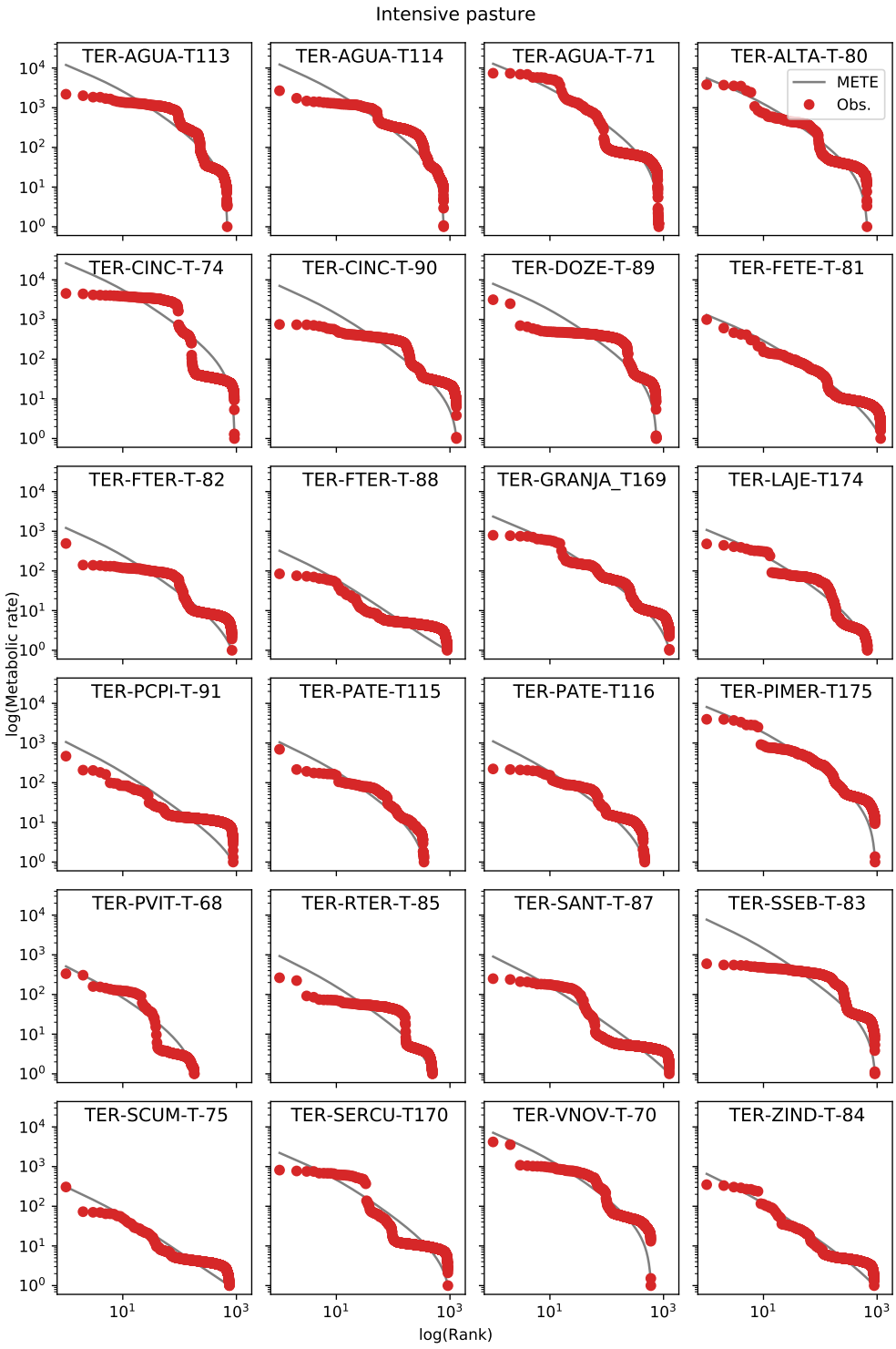


Figure A.18: The rank ordered MRDI at each transect in the intensive pasture.

Appendix B

Appendix for Chapter 2 – "Relating the strength of density dependence and the spatial distribution of individuals"

B.1 Derivation of the probability distribution

Following the methods in the main text and equating the rates entering and exiting $\Pi(n)$ gives

$$p_R(n|n_0-1)p_{D,L}(n+1|n_0)\Pi(n+1|n_0) = p_L(n|n_0-1)p_{D,R}(n|n_0)\Pi(n|n_0).$$

Plugging in with Eq. 2.4 and 2.6 gives

$$\frac{(n+1)^{\alpha-1}}{(n_0-n-1)^\alpha+(n+1)^\alpha}\Pi_\alpha(n+1) = \frac{(n_0-n)^{\alpha-1}}{(n_0-n)^\alpha+n^\alpha}\Pi_\alpha(n). \tag{B.1}$$

We can solve this recursion relation to obtain a general stationary solution. To do so, we define

$$P(n) = \frac{\Pi(n)}{(n_0-n)^\alpha+n^\alpha}$$

to get

$$(n+1)^{\alpha-1}P(n+1) = (n_0-n)^{\alpha-1}P(n).$$

We can fix $\Pi(0)$, since we can choose it freely before normalizing the distribution. For simplicity, choose $\Pi(0)=1$. Then we can write the first few P terms and generalize to $P(n)$:

$$
\begin{aligned}
P(0) &= \frac{1}{n_0^\alpha} \\
P(1) &= \frac{1}{n_0} \\
P(2) &= \left(\frac{n_0 - 1}{2}\right)^{\alpha-1} \frac{1}{n_0} \\
P(3) &= \left(\frac{(n_0-1)(n_0-2)}{6}\right)^{\alpha-1} \frac{1}{n_0} \\
&\vdots \\
P(n) &= \left(\frac{(n_0-1)(n_0-2)\dots(n_0-(n-1))}{n!}\right)^{\alpha-1} \frac{1}{n_0} \\
&= \left(\frac{n_0!}{n!(n_0-n)!}\right)^{\alpha-1} \frac{1}{n_0^\alpha}.
\end{aligned}
$$

Finally, we plug this back into our definition of $\Pi(n)$ to obtain the general stationary $\Pi(n)$ for a given n_0 and α

$$\Pi_\alpha(n|n_0) = \frac{n^\alpha + (n_0 - n)^\alpha}{C(n_0, \alpha) n_0^\alpha} \binom{n_0}{n}^{\alpha-1} \tag{B.2}$$

which is Eq. 2.7 in the main text.

We now want an approximate analytic form for the normalization $C(n_0, \alpha)$,

$$C(n_0, \alpha) = \sum_{n=0}^{n_0} \frac{(n^\alpha + (n_0 - n)^\alpha)}{n_0^\alpha} \binom{n_0}{n}^{\alpha-1}. \tag{B.3}$$

We begin by approximating the binomial term with a normal distribution, which is good for large n_0. This gives

$$\binom{n_0}{n}^{\alpha-1} \approx \exp\left(-\frac{2}{n_0}(\alpha-1)(n - n_0/2)^2\right) \left(2^{-n_0}\sqrt{\frac{\pi n_0}{2}}\right)^{-(\alpha-1)}.$$

We then use the symmetry of the sum over n to write $n^\alpha + (n_0 - n)^\alpha = 2n^\alpha$. Finally, we replace the sum with an integral over n from 0 to infinity, again assuming n_0 is very large. This integral gives

$$
\begin{aligned}
\int_0^\infty dn 2n^a \exp\left(-\frac{2}{n_0}(\alpha-1)(n - n_0/2)^2\right) = \\
\left(\frac{n_0}{2(\alpha-1)}\right)^{\alpha/2} \left(\frac{\alpha n_0}{2}\Gamma\left(\frac{\alpha}{2}\right) {}_1F_1\left(\frac{1-\alpha}{2}, \frac{3}{2}, -\frac{n_0}{2}(\alpha-1)\right) + \right. \\
\left. \sqrt{\frac{n_0}{2(\alpha-1)}}\Gamma\left(\frac{\alpha+1}{2}\right) {}_1F_1\left(-\frac{\alpha}{2}, \frac{1}{2}, -\frac{n_0}{2}(\alpha-1)\right)\right).
\end{aligned} \tag{B.4}
$$

Now if we assume $n_0(\alpha - 1)$ is large, we can approximate

$$ {}_1F_1\left(a, b, z\right) \approx \frac{\Gamma(b)}{\Gamma(b-a)}(-z)^{-a} $$

and Eq. B.4 becomes

$$ 2\sqrt{\frac{\pi}{\alpha - 1}}\left(\frac{n_0}{2}\right)^{\alpha+1/2}. $$

Putting the other contributions to the normalization back in, we get

$$ C(n_0, \alpha) = \frac{2^{n_0(\alpha-1)}\pi n_0}{\sqrt{\alpha - 1}}\left(\frac{1}{2\pi n_0}\right)^{\alpha/2}, \tag{B.5} $$

as Eq. 2.8 in the main text.

B.2 The problem with rank ordered fractions

As mentioned in the main text, Harte (2011) and others have compared bisection predictions to data by rank ordering the fraction of individuals present in half of the plot for each species. This gives a number of points equal to the number of species in the plot. Rank ordering in this way with the BCI data gives us 229 data points, one for each species.

Since our models only predicts the bisection curve for a single n_0, but our data incorporates species with different n_0, we still have to do a bit of work to compare our theory to the rank ordered data. For each species, we generate one value of n randomly from the desired theoretical distribution with abundance n_0. We then rank order these predicted fractions and compare to the rank ordered data. Figure B.1A shows the results using the BCI data.

By eye to the first decimal place, it looks like $\alpha = 1.4$ is a very good fit to the data. With this method, the BCI data deviated the most from the METE prediction for the datasets considered in Chapter 8.3 of Harte (2011).

To fit the free parameter α more rigorously, we maximize the log-likelihood of the density dependent distribution given n and n_0 from the data. We obtain $\alpha = 1.12$, which by eye does not appear to be as good of a fit to the rank ordered data.

The log-likelihood values from the rank-ordered best fit α, the maximum log-likelihood α, as well as random placement and METE are given in Table B.1. Note that even though our model with $\alpha = 1.4$ looks like it has a much better fit in Fig. B.1A, the log-likelihood is very close to that of METE, whereas $\alpha = 1.12$ provides a much better fit to the data in terms of maximizing log-likelihood.

Unfortunately, rank ordering the results from the distribution draws in Fig. B.1A hides the likelihood of individual points and ignores the effect of n_0. The rank ordered plot assigns the same weight to each data point, whereas maximizing the log-likelihood considers how likely each data point is. This is particularly problematic for rank ordering as the probability distribution itself changes with n_0, but Fig. B.1A puts all points on the same plot by fraction. This is misleading since the distribution itself depends on both the fraction n/n_0, and n_0.

We show this more explicitly in Fig. B.1B, which is similar to Fig. 2.3 in the main text, and shows the fraction and abundance together on one plot with 95% probability contours of each distribution overlayed. We can see that with increasing n_0, the random placement model narrows very quickly to having most of its probability weight around 0.5, whereas the METE contours are very wide. We can see that the BCI data does narrow with n_0, but not as much as predicted by random placement.

Most important for this section, we see that the contours for $\alpha = 1.4$ narrow much faster than the data, and many individual data points fall outside of the contours, particularly at moderate to high abundance. The individual points that fall outside of the contours have very low probability, and so when maximizing log-likelihood we obtain a smaller value of α where more of the points fit within the 95% contour intervals. This method is preferred to rank ordering, as it accounts for the likelihood of individual points and the probability distribution's dependence on n_0.

In summary, rank ordering by fraction obscures the abundance data as it only presents the fractions and does not properly account for how the distribution depends on n_0, or the likelihood of individual points. We do not recommend rank order comparisons across data with different n_0, and instead recommend maximizing log-likelihood.

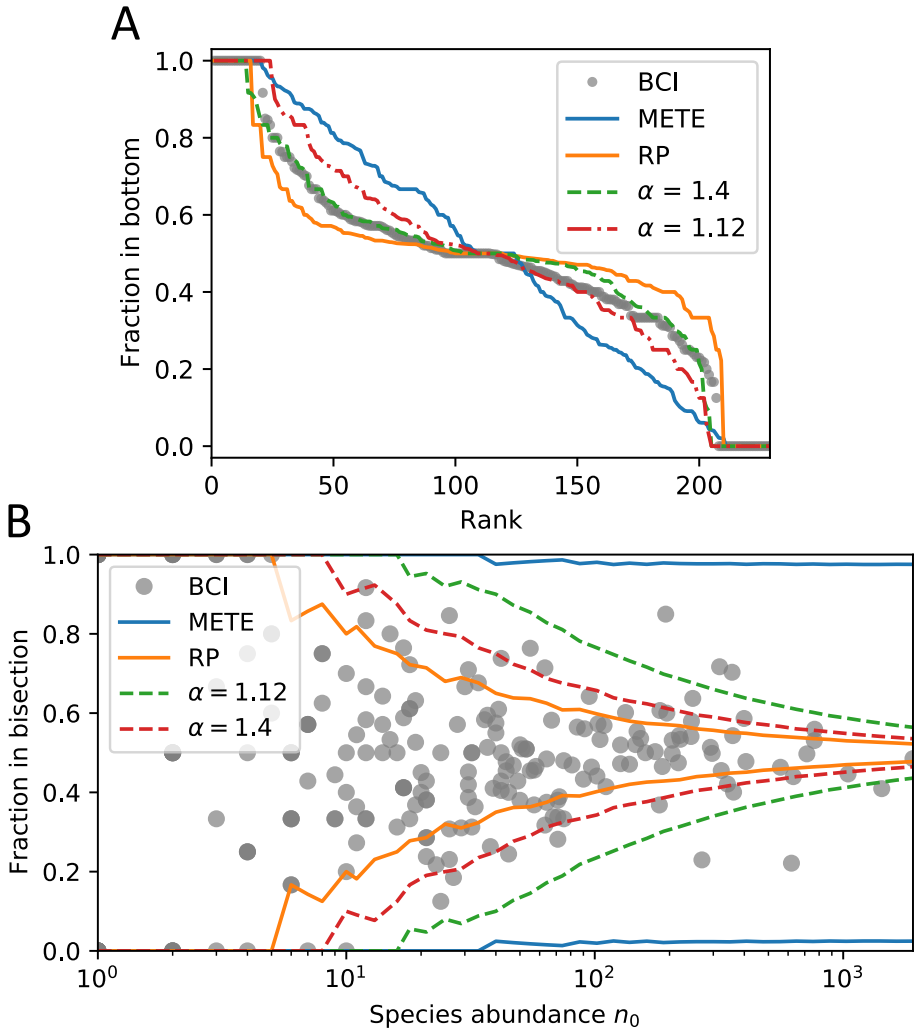


Figure B.1: Rank ordered fraction comparison of METE, random placement, and our density dependent model. By eye, **(A)** seems to indicate that our model with $\alpha = 1.4$ is a much better fit compared to the maximum likelihood value of $\alpha = 1.12$. In **(B)**, we plot the BCI data with the fraction in one half of the bisection versus the total abundance in that bisection. We overlay this data with the 95% contour intervals for each probability distribution at each n_0. We see that many individual data points, particularly at moderate to high abundance, fall outside of the 95% contour intervals for $\alpha = 1.4$, which is why the maximum likelihood approach prefers the smaller value of $\alpha = 1.12$.

Model	**Log-likelihood**
METE	-729
RP	-963
$\alpha = 1.4$	-718
$\alpha = 1.12$	-660

Table B.1: Log-likelihood values for the BCI data set for the three different models, including the by eye fit to the rank ordered plot and the maximum likelihood estimate of α. Despite the good apparent fit in the rank ordered plot, here we see that a maximum likelihood approach prefers a smaller α.

B.3 Error in recovering α

To test how well we could fit for a known α using our methods, we directly simulate the placement and death processes in Eqs. 2.4 and 2.6.

In our simulations, we start with $n_0 = 40$ and then assign a random initial n. We simulate 500 placements and deaths with $\alpha = 1.2$ and take the final n, repeating this p times to simulate the number of observed points (e.g. species at a single bisection). We then obtain the maximum likelihood α.

Figure B.2 shows the results of these simulations. Figure B.2A shows explicitly how the error scales with the number of points p, and Fig. B.2B shows the mean α we recover with standard deviations from 5 runs at each number of points.

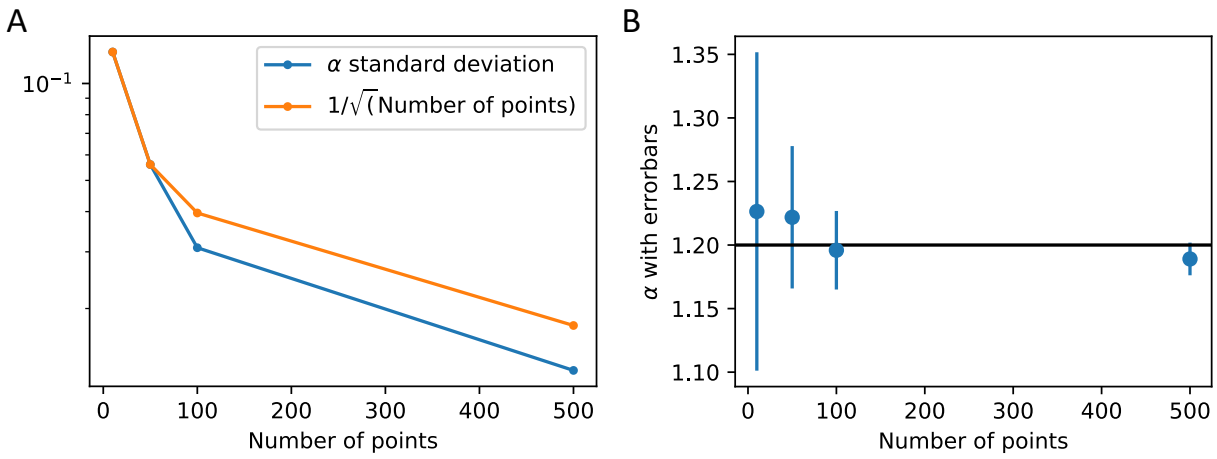


Figure B.2: The error in recovering α from our simulation scales as 1 over the square root of the number of points p. This is shown explicitly in **(A)**. **(B)** shows the value and standard deviation of α that we recover from a simulation with a known $\alpha = 1.2$ (the thick black line). The number of points on the x-axis is analogous to the number of data points we observe in data, and the standard deviation is obtained from doing 5 simulations with that number of runs. We simulated with $p = \{10, 50, 100, 500\}$.

B.4 Relating to the binomial distribution

One way to understand why Eq. 2.9 approaches the binomial distribution for large n_0 is to rewrite Eq. 2.9 as:

$$\Pi_{\alpha=2}(n) = \left(\frac{4(n - n_0/2)^2}{n_0(n_0+1)} + \frac{n_0}{n_0+1} \right) \mathrm{BD}(n, n_0, 1/2). \tag{B.6}$$

The first term in the parentheses is 0 at $n = n_0/2$ and approximately 1 (given $\frac{n_0}{n_0+1} \approx 1$) at $n = \{0, n_0\}$, and the second term is approximately 1. This means that there is a factor of 2 increase compared to the binomial in the tails of the distribution near $n = \{0, n_0\}$, but around the center the two distributions look identical. Since the tails are already so small at large n_0, this factor of 2 isn't really noticeable and so this distribution looks very similar to the binomial distribution.

B.5 Relating to the negative binomial distribution

The negative binomial distribution is very common in ecological modelling (Bliss and Fisher 1953; He and Gaston 2000; He and Gaston 2003), and many ecologists are familiar with the k parameter as a measure of aggregation. A standard way to write the negative binomial distribution is

$$\mathrm{NB}(n|p,k) = (1-p)^n p^k \frac{\Gamma(n+k)}{\Gamma(n+1)\Gamma(k)}, \tag{B.7}$$

where p is a parameter that relates to the mean μ of the distribution as $\mu = k(1-p)/p$.

For a bisection we use the conditional negative binomial distribution, where the negative binomial is conditioned on a total of n_0 individuals. Given n individuals on one side of the distribution and total number of individuals n_0, the corresponding conditional negative binomial is (Conlisk et al. 2007)

$$\mathrm{CNB}(n|n_0,p,k) = \frac{\Gamma(n_0+1)\Gamma(2k)}{\Gamma^2(k)\Gamma(n_0+2k)} \frac{\Gamma(n+k)\Gamma(n_0-n+k)}{\Gamma(n+1)\Gamma(n_0-n+1)}. \tag{B.8}$$

This distribution is equivalent to random placement as $k \to \infty$, and equivalent to METE when $k = 1$.

We want to compare this to our distribution, Eq. 2.7. However, we cannot compare directly as the unknown normalization C depends on α and n_0. Instead, we consider only the n dependence. If we equate it to the Π distribution:

$$\mathrm{CNB} = \frac{\Gamma(n+k)\Gamma(n_0-n+k)}{\Gamma(n+1)\Gamma(n_0-n+1)} \sim \frac{n^\alpha + (n_0-n)^\alpha}{(\Gamma(n+1)\Gamma(n_0-n+1))^{\alpha-1}} = \Pi.$$

Then we want to solve

$$\Gamma(n+k)\Gamma(n_0-n+k) \sim (n^\alpha + (n_0-n)^\alpha)(\Gamma(n+1)\Gamma(n_0-n+1))^{2-\alpha}. \tag{B.9}$$

To ensure this expression makes sense, we can directly consider the limiting cases of METE and the binomial distribution. The METE distribution corresponds to $\alpha = 1$ for our model, and we can see that for the LHS to have the same n dependence this means $k = 1$, which as noted above also corresponds to METE for the conditional negative binomial. For $\alpha = 2$, which corresponds roughly to the binomial distribution if we ignore the first term on the RHS (good up to a factor of 2, see Eq. B.6), the n dependence on the RHS drops out. For this to be true on the LHS, we need $k \to \infty$, which makes the conditional negative binomial equivalent to the binomial distribution. This means both limiting cases in Eq. B.9 make sense.

It is analytically challenging to equate the n dependence between these two forms more generally, however by plotting the two distributions (Fig. B.3) we can see that they are quite similar for specific k and α values.

We can make some comparison by equating the ratios between two different points n and n', eliminating the normalization problem. This is not ideal because we are now only

equating at two artificial points, and it is possible the distributions are different outside of that. However, by plotting the distributions (Fig. B.3) we can see that matching the peaks of the distributions roughly makes the distributions match at large enough n_0, so we will use this approximation.

Taking the ratio gives

$$\frac{\Gamma(n'+k)\Gamma(n_0-n'+k)}{\Gamma(n+k)\Gamma(n_0-n+k)} = \frac{(n'^{\alpha}+(n_0-n')^{\alpha})(\Gamma(n'+1)\Gamma(n_0-n'+1))^{2-\alpha}}{(n^{\alpha}+(n_0-n)^{\alpha})(\Gamma(n+1)\Gamma(n_0-n+1))^{2-\alpha}}.$$

We now equate the ratios of the central points by letting $n' \to n_0/2+1$, and $n \to n_0/2$. This gives

$$\frac{n_0/2+k}{n_0/2+k-1} = \left(1+\frac{2}{n_0}\right)^{2-\alpha}\left(\frac{1}{2}\left(1+\frac{2}{n_0}\right)^{\alpha}+\frac{1}{2}\left(1-\frac{2}{n_0}\right)^{\alpha}\right).$$

Expanding around large n_0 and keeping the first order term gives

$$k \approx \frac{n_0}{2}\left(\frac{\alpha-1}{2-\alpha}\right).$$

For $\alpha = 1$, we know $k = 1$, so we use that as the constant offset to get our approximate relationship

$$k \approx \frac{n_0}{2}\left(\frac{\alpha-1}{2-\alpha}\right)+1. \tag{B.10}$$

Figure B.3 plots the conditional negative binomial distribution with k calculated from Eq. B.10 compared to our distribution for a range of αs. Our derived approximate relationship results in good agreement between these distributions.

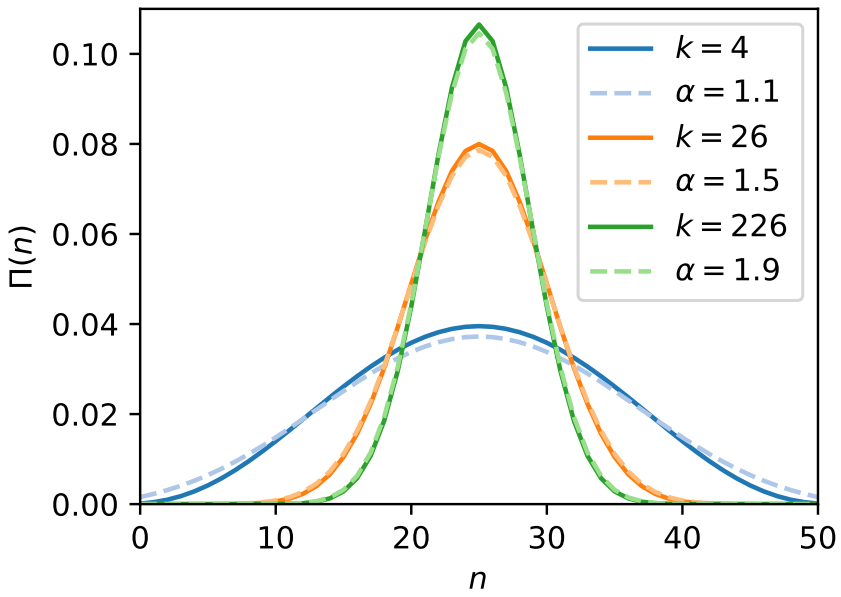


Figure B.3: Conditional negative binomial distributions (solid lines) with k calculated from Eq. B.10 compared to our predicted distributions with their corresponding α (dashed lines). $n_0 = 50$.

B.6 Comparison to Conlisk et al. (2007)

Using our approximate relationship between α and the negative binomial parameter k, Eq. 2.10, we can relate α to the ϕ parameter in Conlisk et al. (2007) at the level of a single bisection. For a single bisection, the probability distribution derived in Conlisk et al. (2007) is equivalent to the conditional negative binomial with $k = (1-\phi)/\phi$ (see their Theorem 1.3 and 1.5 with $c = 2$, and Theorem 2.3 with $i = 1$). Using Eq. 2.10, we find

$$\phi \approx \frac{2(2-\alpha)}{(\alpha-1)n_0 + 4(2-\alpha)}. \tag{B.11}$$

Note that this relationship agrees for METE and random placement – when $\alpha = 2$, $\phi = 0$, and when $\alpha = 1$, $\phi = 1/2$. Note also that this relationship depends on n_0. This makes it more complicated to compare their Fig. 6 to our results, as there is no single n_0 at each scale. In order to make this comparison, at each scale we use the median n_0 across quadrats and species to relate α to ϕ. The results are shown in Fig. B.4.

An additional difference between our analyses is our use of a threshold of $n_0 > 128$ for species to be included in the scaling relationship, and our cutoff of dbh $>$ 100mm rather than 10mm for the BCI data. However, this does not account for our different scaling results at the community level. Figure B.5 shows our scaling results without these cutoffs, as well as the transformed results using Eq. B.11, again using median n_0 at each scale. The serpentine results are largely the same, which makes sense as there are few species with $n_0 < 128$. While the BCI results change slightly, we still find a difference in scaling compared to Fig. 6 in Conlisk et al. (2007) as BCI looks more random across scales.

Overall, this shows that the scaling patterns we see in the aggregation parameter depend at least in part on the model of aggregation, and how the data are analyzed. In this case, the difference between our scaling results at BCI when compared to Conlisk et al. (2007) are likely due to a difference in how we treat n_0 across scales, and in our use of a threshold for n_0.

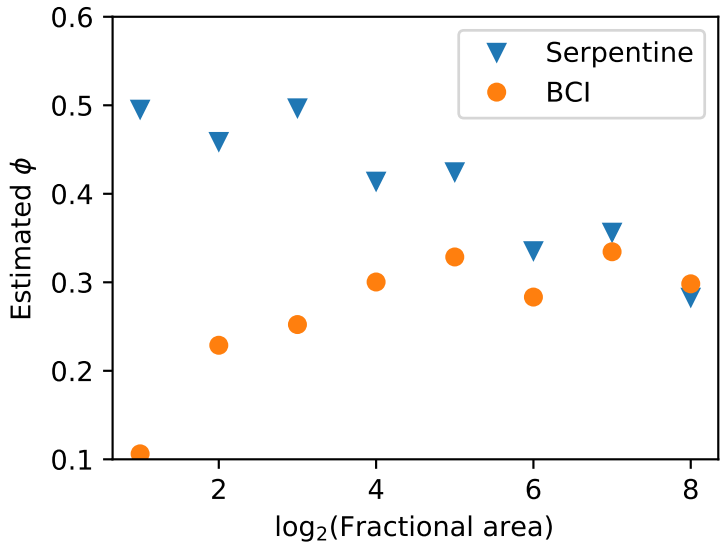


Figure B.4: The community α scaling relationship, as Fig. 2.4, where α has been transformed to ϕ using Eq. B.11. At each scale, the median n_0 across plots and species was used for this relationship.

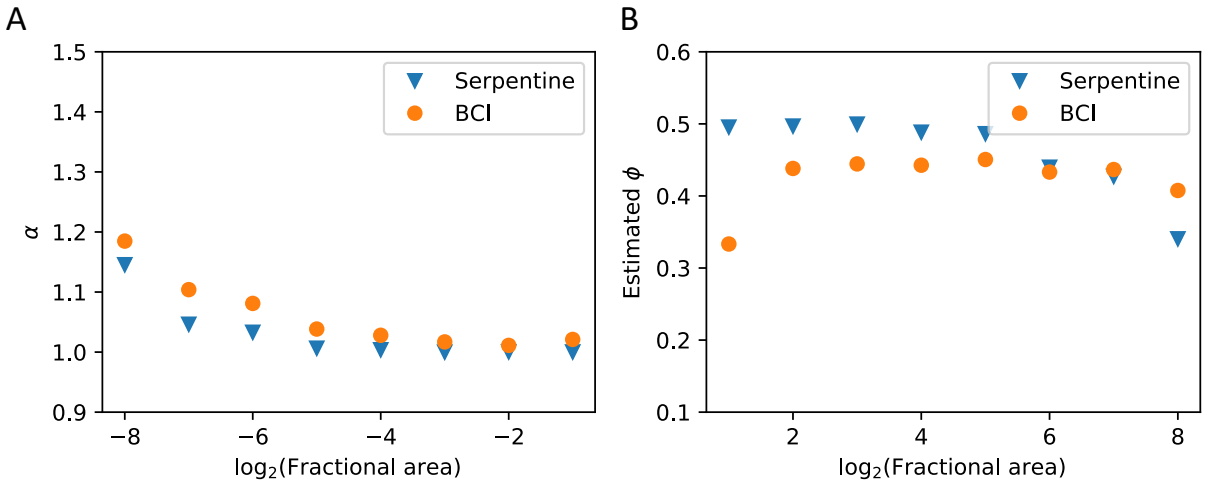


Figure B.5: The community α scaling relationship, as Fig. 2.4, but we have included species of all abundance (including $n_0 < 128$), and species of all measured dbh (dbh $>$ 10mm). **(A)** shows this relationship for α directly, and **(B)** shows this relationship after α has been transformed to ϕ using Eq. B.11. At each scale, the median n_0 across plots and species was used for this relationship.

B.7 Sampling effect

As noted in the main text, at small spatial scales it becomes difficult to distinguish different patterns of aggregation. More specifically, it is difficult to distinguish aggregation and randomness when the number of individuals is small. This creates a sampling effect where patterns are more likely to look random at small spatial scale, particularly when $n_0 A/A_0 << 1$ (Harte 2011, pg 63). In our analysis, this sampling effect should not have too large of an impact. At each spatial scale, we bisect each quadrat, which means $A/A_0 = 1/2$ for all data points. Additionally, as noted in Methods, we used a threshold of $n_0 > 128$ to avoid too many low abundance cells, as $\Pi_\alpha(n)$ does not depend on α for $n_0 \leq 2$. Thus, as long as there are not too many cells with small n_0, this sampling effect should be small.

Figure B.6 shows the distribution of $n_0 A/A_0$ at each scale for each dataset. Since for each quadrat $A/A_0 = 1/2$, this is the distribution of $n_0/2$ where each data point represents one species in one qudrat. Note that even at the smallest scale, the median $n_0 A/A_0$ is greater than 1 for both datasets.

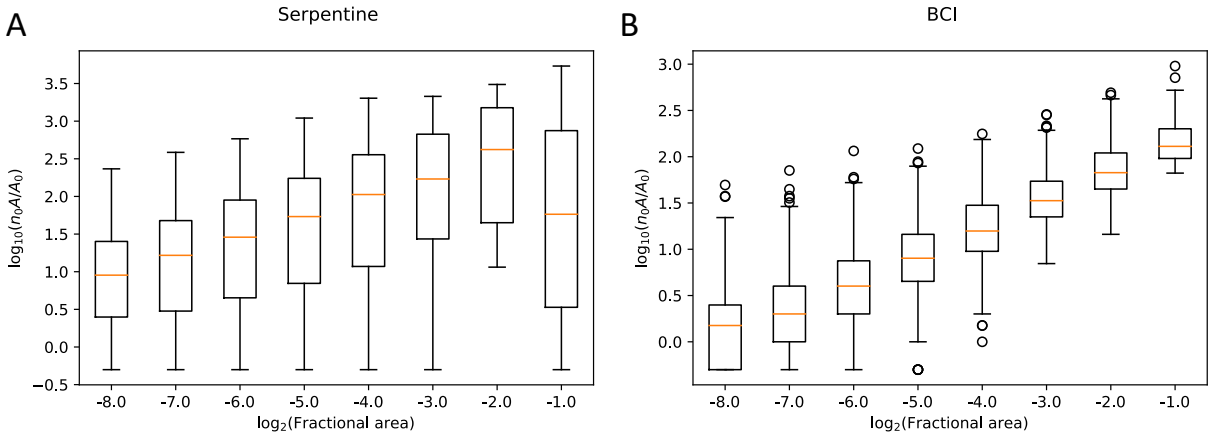


Figure B.6: The distribution of $n_0 A/A_0$ at each scale for **(A)** Serpentine and **(B)** BCI. Note that since $A/A_0 = 1/2$, this is equivalently the distribution of $n_0/2$ at each scale, where each n_0 value is the abundance of a single species in a single qudrat at that scale. The boxplots show boxes from quartile 1 (Q1) to quartile 3 (Q3) with a line at the median. The whiskers extend to 1.5×(Q3-Q1). The remaining points are plotted as individual circles.

B.8 Trends in diameter and abundance for BCI data

Each figure in this section displays relationships between α and species characteristics that may correlate with α. Overall we find that there is a small but significant relationship of α with the log of abundance, where α and its variance decrease with increasing abundance. There are no other significant relationships among those tested. The significance of the regression lines is displayed as a two-sided p-value calculated using a Wald Test with the null hypothesis that the slope is zero.

Figure B.7 shows the relationship between α and log of species abundance at different scales, using the BCI data. At smaller scales, the distribution of α is broader and generally higher, as we found in Fig. 2.2. Note that there aren't any high abundance high α values. This relationship is verified by the displayed p-values, where at all scales but the two smallest $p < 0.05$. The correlation coefficient r, as well as the slope and its standard error, are shown in Table B.2.

Figure B.8 shows the relationship between α and the mean dbh for each species at different scales, again using the BCI data. The correlation coefficients r, the slopes, and the standard error of the slopes are shown in Table B.3. We still see the trend that at smaller scales the α distribution is broader, but at no scale is there an apparent relationship between α and the mean dbh. The relationship at every scale except the finest is very flat. We might expect that larger species have higher α values at larger scales than smaller species, as they are competing at a different scale, however we don't see that effect. It is possible that bisections are not sufficient to pick up on this difference, or that intraspecies size variation obscures this trend. Additionally, the range of species' mean dbh here is only from about 100 mm to about 500 mm, so it is also possible that we need to consider a much larger range of dbh before we see this effect. Finally, as mentioned in the main text, it is possible we did not go to scales small enough to see the difference in aggregation from size variation.

Finally, we can test if the lack of high α high n_0 species is a consequence of energy equivalence. That is, the most abundant species are also the smallest. We didn't see a trend in dbh alone, but we may see a trend in species total metabolic rate. For trees, metabolic rate scales approximately as dbh^2. Figure B.9 shows how α scales with the log of the total metabolic rate of all individuals in a species (which scales as $n_0 \times \text{dbh}^2$). Here we see only one significant relationship at the scale of 64 cells, and at other scales the relationship is not significant and the slope is small. The correlation coefficients r, the slopes, and the standard error of the slopes are shown in Table B.4.

Scale	r	Slope	Std. Error
256 Cells	-0.14	-0.22	0.25
128 Cells	-0.25	-0.41	0.26
64 Cells	-0.34	-0.51	0.22
32 Cells	-0.32	-0.30	0.22
16 Cells	-0.37	-0.32	0.13
8 Cells	-0.33	-0.67	0.31

Table B.2: The correlation coefficients r, as well as the slopes and their standard errors for the regression lines in Fig. B.7.

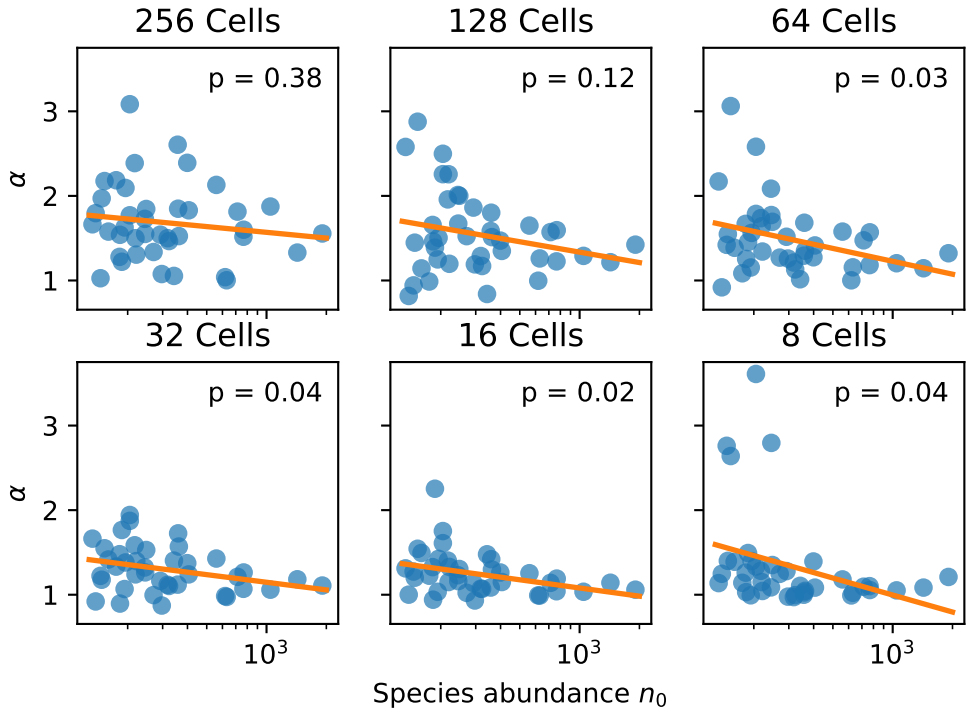


Figure B.7: Species abundance and their corresponding maximum likelihood α value using the BCI data at a range of spatial scales from 3 bisections to 8 bisections. The absolute scale is 50 ha divided by the number of cells. The displayed p-value is calculated with the null hypothesis that the slope is zero using the Wald test.

Scale	r	Slope	Std. Error
256 Cells	0.29	0.0016	0.0008
128 Cells	-0.05	-0.0003	0.0009
64 Cells	-0.08	-0.0004	0.0008
32 Cells	0.09	0.0003	0.0005
16 Cells	0.04	0.0001	0.0005
8 Cells	-0.01	-0.0001	0.0011

Table B.3: The correlation coefficients r, as well as the slopes and their standard errors for the regression lines in Fig. B.8.

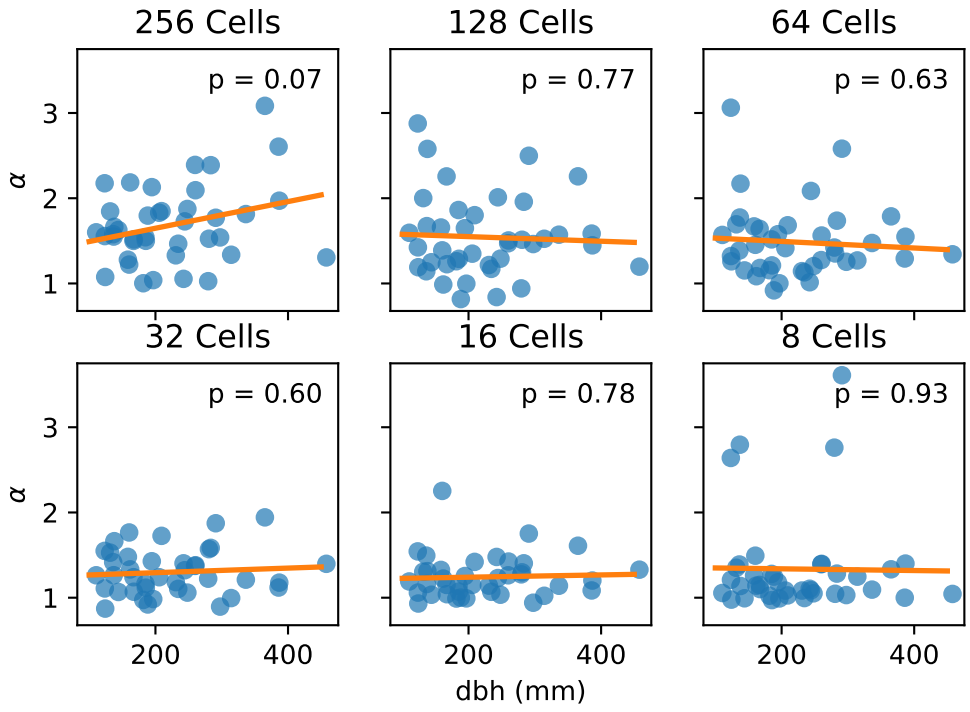


Figure B.8: Mean dbh in mm for each species and their corresponding maximum likelihood α value using the BCI data at a range of spatial scales from 3 bisections to 8 bisections. The absolute scale is 50 ha divided by the number of cells. The displayed p-value is calculated with the null hypothesis that the slope is zero using the Wald test.

Scale	r	Slope	Std. Error
256 Cells	0.11	0.13	0.18
128 Cells	-0.24	-0.27	0.18
64 Cells	-0.33	-0.34	0.16
32 Cells	-0.17	-0.11	0.10
16 Cells	-0.23	-0.14	0.10
8 Cells	-0.25	-0.34	0.22

Table B.4: The correlation coefficients r, as well as the slopes and their standard errors for the regression lines in Fig. B.9.

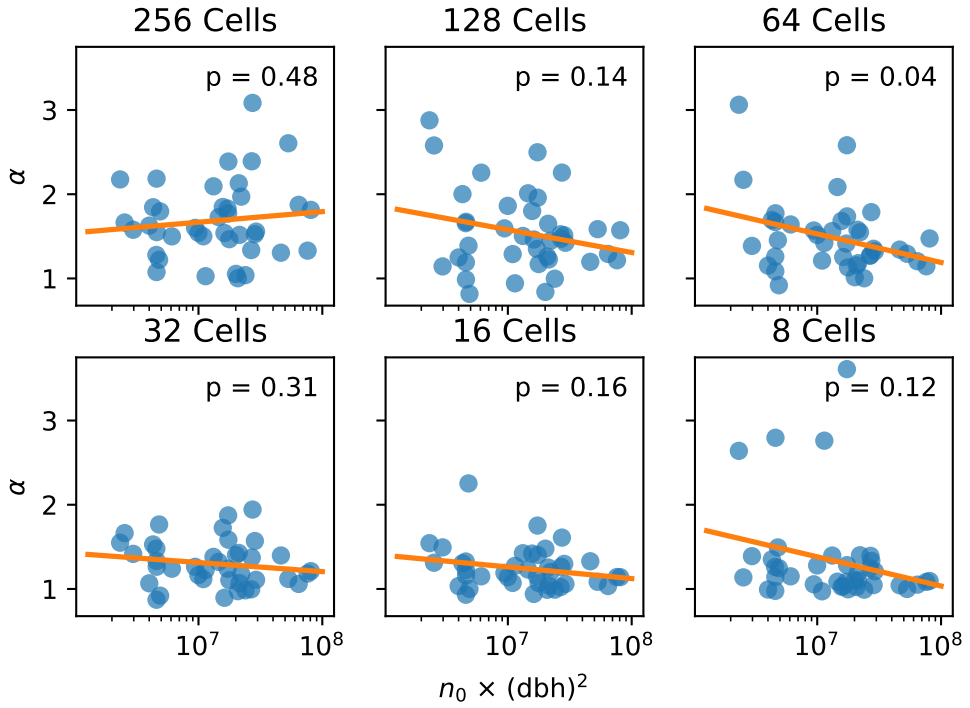


Figure B.9: Each species' total metabolic rate scales as $n_0 \times \text{dbh}^2$, which is plotted against their corresponding maximum likelihood α value using the BCI data at a range of spatial scales from 3 bisections to 8 bisections. The absolute scale is 50 ha divided by the number of cells. The displayed p-value is calculated with the null hypothesis that the slope is zero using the Wald test.