View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by University of Tennessee, Knoxville: Trace

University of Tennessee, Knoxville
Trace: Tennessee Research and Creative Exchange

Doctoral Dissertations

Graduate School

5-2007

Scale and Contingency in Plant Demography: Quantitative Approaches and Inference

Sean Maurice McMahon
University of Tennessee - Knoxville

Recommended Citation

McMahon, Sean Maurice, "Scale and Contingency in Plant Demography: Quantitative Approaches and Inference. " PhD diss., University of Tennessee, 2007.
https://trace.tennessee.edu/utk_graddiss/241

This Dissertation is brought to you for free and open access by the Graduate School at Trace: Tennessee Research and Creative Exchange. It has been accepted for inclusion in Doctoral Dissertations by an authorized administrator of Trace: Tennessee Research and Creative Exchange. For more information, please contact trace@utk.edu.

To the Graduate Council:

I am submitting herewith a dissertation written by Sean Maurice McMahon entitled "Scale and Contingency in Plant Demography: Quantitative Approaches and Inference." I have examined the final electronic copy of this dissertation for form and content and recommend that it be accepted in partial fulfillment of the requirements for the degree of Doctor of Philosophy, with a major in Ecology and Evolutionary Biology.

James A. Drake, Major Professor

We have read this dissertation and recommend its acceptance:

Dan Simberloff, Nathan Sanders, Halima Bensmail

Accepted for the Council:
Dixie L. Thompson

Vice Provost and Dean of the Graduate School

(Original signatures are on file with official student records.)

To the Graduate Council:
I am submitting herewith a dissertation written by Sean Maurice McMahon entitled "Scale and contingency in plant demography: quantitative approaches and inference". I have examined the final electronic copy of this dissertation for form and content and recommend that it be accepted in partial fulfillment of the requirements for the degree of Doctor of Philosophy, with a major in Ecology and Evolutionary Biology.

James A. Drake
James A. Drake Major Professor

We have read this dissertation
and recommend its acceptance:

Dan Simberloff

Nathan Sanders

Halima Bensmail

Accepted for the Council:

Linda Painter
Interim Dean of
Graduate Studies

(Original signatures are on file with official student records)

Scale and contingency in plant demography: quantitative approaches and inference

A Dissertation
Presented for the
Doctor of Philosophy
Degree
The University of Tennessee, Knoxville

Sean Maurice McMahon
May 2007

Copyright © 2006 by Sean M. McMahon.
All rights reserved.

Dedication

This dissertation is dedicated to my parents and debo.

Acknowledgments

I would like to thank most importantly, my family. My parents, Sarah Lynne and Michal, my brother and sister-in-law, Jeff and Lizanne, and Deb supported me in every way a family could. My friends both at UT and elsewhere also formed an essential network that provided emotional support, intellectual stimulation, and important connections and reminders of the world outside of graduate school. The research for this thesis was aided by my adviser, Jim Drake, who offered me all manner of support, guidance, mentorship, and friendship. My committee, as they stand now (in Nate Sanders, Dan Simberloff, and Halima Bensmail) and have stood (Michael Huston, Adrian Mayer, and Aaron King), also offered critical guidance and assistance. The members of my stats masters team, Monica and Justin, not only made learning statistics fun; they made it possible. I thank the many people that assisted in field work: Matt, Matt, Justin, Deb, and others. The National Park Service, the All Taxa Biodiversity Inventory, especially Keith Langdon, Chuck Parker, and Jeanie Hilton, provided material support, permits, and housing for the field component of this thesis. The Department of Ecology and Evolutionary Biology at Tennessee offered financial support and an excellent overall learning environment. I'd like to thank Phyllis and Cheryl for their fantastic administrative work for the department and myself.

Abstract

Ecologists have long recognized that patterns measured in nature often depend upon the context in which they are observed and the scale at which they are observed. When studying plant populations, the role of scale and contingency becomes crucial. Thinking about a plant community as a system is essential as populations of plants are centered within a network that influences their dynamics in direct and indirect ways. Plant populations are inherently scale-dependent because they have properties as a group that can be independent of their properties as individual stems. Although the challenge of interpreting population patterns in the face of contingency and scale has been addressed conceptually, there has been less success in applying those concepts to observational and experimental studies. This dissertation addresses the challenges of modeling the demographic dynamics of a forest understory herb, *Eurybia chlorolepis* (Asteraceae) or mountain aster. The study population consisted of twenty patches containing between 20 and 70 individual stems in each patch. These patches spanned three sites within the Indian Camp Creek watershed in the Cosby Ranger district of Great Smoky Mountains National Park. Plants in the forest understory in this dense old-growth forest are influenced by a myriad of biotic and abiotic components of the community: light, soil characteristics, other plant species, herbivores, pollinators, seed predators, and the feet of bears. This dissertation shows that the mechanisms that influence sexual reproduction of this plant are structured almost entirely on the stem-to-stem scale, indicating little coarse-scale influence of the environment over sexual reproduction. The use of a Bayesian learning network showed that the environmental influences (soil in particular) operated most importantly in the transition from juvenile stage to adult stage. Taken together, these analyses indicate that the coarse-environtment (such as gaps, soil profiles, soil moisture, and the presence of other plants) dictates where *E. chlorolepis* becomes reproductive, while the success of that reproduction is dictated by mechanisms operating between individual stems.

Contents

List of Tables

List of Figures

Part I

Introduction

Understanding the distribution of organisms in space and time has been a core theme of ecology. Experiments conducted under controlled conditions as well as theoretical models have suggested that fundamental mechanisms hypothesized to describe species distributions, such as niche theory (Chase and Leibold, 2003; Pianka and Huey, 1978; Silvertown, 2004), competition (Pacala and Tilman, 1994; Tilman, 1982), dispersal, and disturbance (Connell, 1978; Watt, 1947), can in fact structure populations. Efforts to map these mechanisms onto natural populations, however, have failed to produce a universal set of rules that ecologists can apply to any particular community. This may be because most natural populations show the product of multiple mechanisms operating simultaneously. Furthermore, these mechanisms operate at different scales, disturbing their straightforward measurement. My dissertation research has focused on detecting the mechanisms that structure a natural plant population. To do this, I measured physical and biological variables of a forest understory herb, *Eurybia chlorolepis* (mountain aster) in Great Smoky Mountains National Park and applied two novel statistical methods. The second and third parts of this thesis introduce these methods and the fourth and fifth parts apply them.

Hierarchical linear models (HLM) consist of nested regression equations. These models can be estimated using maximum likelihood or hierarchical Bayesian techniques (Raudenbush and Bryk, 2002). They are designed to quantify the role of scale in structuring a response variable and estimating covariates that can explain that structure. HLM evolved from classical statistical models like nested ANOVA and mixed models. In the past two decades, however, new computationally intensive estimation techniques have allowed these models to address more complex problems. In Part II, I describe how these models can be estimated from hierarchically structured data. I focus on the use of unconditional models to measure the scale at which a

response variable shows the most variation. Following this diagnostic model, conditional models are built using covariates to explain relevant variation. I show two example applications of these models. The first example builds a linear two-level model estimated with maximum likelihood techniques to explain leaf damage from herbivores on a clonal plant. The second uses Bayesian estimation techniques to build a three-level logistic model to determine the influences on the probability of flowering of another forest herb structured over three spatial scales.

Bayesian learning networks (BLN) were developed in the field of artificial intelligence to quantify networks of interacting variables. In Part III, I explain how BLN can find structure in a collection of correlated variables, quantify that structure, and draw inference from observed or hypothesized changes in that structure. I focus on the benefits of applying this method to the direct and indirect effects of invasive organisms. Research in invasion biology has struggled to identify simple mechanisms that determine the probability or results of invasive organisms. BLN, I argue, is an ideal method to use to tease apart important pathways in invaded ecological systems.

In Part IV, I apply HLM to a field study of *Eurybia chlorolepis*, focusing on the biotic and abiotic variables that influence seed fate. The fate of seeds is important to measure because sexually reproductive plants invest in producing flowers, pollen, ovaries and eventually seeds at the expense of other tissues important to resource acquisition, vegetative reproduction, and survival (Ehrlen, 1991; Kneitel and Chase, 2004; Silvertown, 2004; Wright and Meagher, 2003). The production of floral parts, however, does not guarantee successful seed production and dispersal. Flowers must be pollinated. Pollenated ovaries must avoid predation. Further, in many plant species, ovaries are aborted before viable seeds are produced. There are a number of hypotheses that explain the determinants of flower production, pollination, seed predation and abortion. By modeling the scale at which these processes occur in

a natural population of *Eurybia chlorolepis*, I was able to show that almost all of the variation in these processes was at the scale of the stem and not the patch. This has important implications for determining which of many biotic and abiotic components of an ecological community actually determine the spatial pattern of sexual reproduction in a clonal plant species. I discuss how the prevalence of stem-level variation conflicts with the assumptions of several possible mechanisms that have been hypothesized to determine seed fate.

Having shown that the determinants of seed fate are idiosyncratic with respect to the patch-level distribution of flowering stems in *E. chlorolepis*, in Part V, I apply BLN to measure which features of the forest understory environment determine whether a patch of *E. chlorolepis* has reproductive individuals or not. Patch-level heterogeneity in soil, light, and biotic features of the understory can be heavily influenced by the overstory tree population. Eastern hemlocks (*Tsuga canadensis*) are an important tree in the cove forest where *E. chlorolepis* is found. The hemlock woolly adelgid (*Adelges tsugae*) , recently discovered in Southern Appalachian forests, can kill up to 95% of the hemlock trees it infests (Orwig and Foster, 1998). By analyzing the network of patch-variables that can create source-populations of *E. chlorolepis* using BLN, I show that soil features related to hemlock presence are important to *E. chlorolepis* populations. This analysis, though preliminary with regards to the multiple and long-term potential effects of hemlock mortality, demonstrates the value of quantifying a network of correlated components of the forest when asking questions about the direct and indirect influences of an invader on a community.

Bibliography

Chase, J. M. and M. A. Leibold. 2003. Ecological niches. Linking classical and contemporary approaches. Univ. of Chicago Press., Chicago.

Connell, J. H. 1978. Diversity in tropical rain forests and coral reefs. Science **199**:1302–1310.

Ehrlen, J. 1991. Why do plants produce surplus flowers? A reserve-ovary model. American Naturalist **138**:918–933.

Kneitel, J. M. and J. M. Chase. 2004. Trade-offs in community ecology: linking spatial scales and species coexistence. Ecology Letters **7**:69–80.

Orwig, D. A. and D. R. Foster. 1998. Forest response to the introduced hemlock woolly adelgid in southern new england, usa. Journal of the Torrey Botanical Society **125**:60–73.

Pacala, S. W. and D. Tilman. 1994. Limiting similarity in mechanistic and spatial models of plant competition in heterogeneous environments. American Naturalist **143**:222–257.

Pianka, E. R. and R. B. Huey. 1978. Comparative ecology, resource utilization and niche segregation among gekkonid lizards in the southern kalahari. Copeia page 691.

Raudenbush, S. W. and A. S. Bryk. 2002. Hierarchical linear models: applications and data analysis methods. Advanced quantitative techniques in the social sciences, Sage Publications, Thousand Oaks, CA.

Silvertown, J. 2004. Plant coexistence and the niche. Trends in Ecology and Evolution **19**:605–611.

Tilman, D. 1982. Resource Competition and Community Structure. Monographs in population biology, Princeton University Press.

Watt, A. S. 1947. Pattern and process in the plant community. Journal-of-Ecology. **35**:1–22.

Wright, J. W. and T. R. Meagher. 2003. Pollination and seed predation drive flowering phenology in *Silene latifolia* (Caryophyllaceae). Ecology **84**:2062–2073.

Part II

Scales of association: hierarchical linear models and the measurement of ecological systems

This part was co-written with Jeff Diez, University of Massachusetts. Sean McMahon is first author.

Abstract

A fundamental challenge to understanding patterns in ecological systems lies in employing methods that can analyze, test, and draw inference from measured associations between variables across scales. Hierarchical linear models (HLM) use advanced estimation algorithms to measure regression relationships and variance-covariance parameters in hierarchically structured data. Although hierarchical models have occasionally been used in the analysis of ecological data, its full potential to describe scales of association, diagnose variance explained, and to partition uncertainty has not been employed. In this paper we argue that full utilization of the HLM framework will enable significantly improved inference about ecological processes across levels of organization. We suggest that ecologists must begin adopting a hierarchical framework if advances are to be made in our field. After briefly describing the principals behind HLM, we give two illustrations to highlight its power to simultaneously describe relationships between variables at multiple scales. The first example employs maximum likelihood methods to construct a two-level linear model predicting herbivore damage to a perennial plant at the individual and patch scale; the second example uses Bayesian estimation techniques to develop a three-level logistic model of plant flowering probability across individual plants, microsites, and populations. HLM model development and diagnostics illustrate the importance of incorporating scale when modeling associations in ecological systems and offers a sophisticated yet accessible method for studies of populations, communities, and ecosystems.

Introduction

Scale is essential to the analysis of ecological systems. The relationship between two variables in a natural system can be obscured by other variables at other scales (Maurer, 1999; Wiens, 1991), and the inferences drawn from an observed relationship can be distorted or even reversed depending on the scale at which that relationship is measured (Cadotte and Fukami, 2005; Denny et al., 2004; Wiens, 1991). For this reason, there have long been calls to incorporate scale explicitly in designing, analyzing, and drawing inference from ecological studies (Allen and Starr, 1982; O'Neill et al., 1986; Levin, 1992, 2000; Holling, 1992; Wiens, 1991; Rahel, 1990). Although a great deal of work has addressed quantitative methods for measuring scale (Borcard et al., 2004, 1992; Dale, 1999; Dungan et al., 2002; Harte et al., 2005; He and Legendre, 2002; Thrush et al., 1997), it is remarkable that so few ecological studies incorporate scale in the analysis of observed natural patterns or experiments. As the importance of scale in determining ecological patterns has become more apparent (Harte et al., 2005; Levin, 1992, 2000) techniques explicitly designed to measure and interpret interactions and associations at different scales will better enable the generalization of these analyses to other systems and the predictive application of the results to future system behaviors (Noda, 2004; Underwood and Chapman, 1996).

Scale in general and hierarchical approaches to scale in particular have rich histories in ecological theory, observation, and experimentation. Ecological data are often hierarchically structured; a fact that arises both from common sampling designs as well as biological truth (e.g., quadrats on transects, species within genera, clonal stems attached to rhizomes, behaviors over time, fish in watersheds). Hierarchical structure in ecology has over the years inspired treatises on proper experimental design (Hurlbert, 1984; Oksanen, 2001), statistical analysis (Clark and Gelfand, 2006;

Raudenbush and Bryk, 2002; Wu and David, 2002), and broader theoretical and philosophical explorations (Allen and Starr, 1982; O'Neill et al., 1986; Rahel, 1990; Whittaker et al., 2001; Levin, 1992, 2000; Noda, 2004). Thus, over the years, ecologists generally agree scale is important, have offered methods to quantify scale, and have implemented a number of studies that show scale to be important. Unfortunately, most ecological research that is not specifically focused on the issue of scale fails to account for scale in analysis and inference. It is this oversight, the vast gap between the agreed importance of scale and the failure to include scale in analysis, that we hope to address in this paper.

There seems to be no consensus approach to quantifying scale in ecological studies. The methods that are applied generally fall into three categories. The first consists of methods that primarily determine the scale at which a pattern is evident (e.g., principle coordinates of neighbor matrices (Borcard et al., 2004), wavelet analysis (Keitt and Urban, 2005), fractal dimensions, (Keitt et al., 1997; Sugihara and May, 1990), canonical correspondence analysis (Cushman and McGarigal, 2003)). Although these methods effectively designate the scale at which a response variable shows distinct patterns and are quite effective at capturing scale-dependent patterns along a continuously-scaled variable (e.g., a fine-grained time-series, or detailed spatial measurements such as GIS), they require an indexing (ordination) of the variable of interest and generally cannot take into account correlation between the same measurement of predictor variables across scales (Keitt and Urban, 2005). These methods also require specialized conceptual knowledge of the techniques and math for application, inference, and communication of the results.

The second category of methods includes classic design of experiments. Nested analysis of variance (ANOVA) and mixed models estimated with ordinary least squares fall into this category (Benedetti-Cecchi et al., 2005; Benedetti-Cecchi, 2001; Cadotte

and Fukami, 2005; Chase and Leibold, 2002). We include for organizational purposes here studies that use basic statistical techniques to find effects (any technique) but apply them at two scales to perform a qualitative analysis (Tolimieri, 1995; Gotelli and Ellison, 2002). These methods have the benefit of ease of use, ease of interpretability, and clarity of result. The fail, however, to be flexible in design (unbalanced data, more complex model constructions, and missing data are difficult to contend with). Further, results from models that require reduction of the system (specifically nested ANOVA designs) frustrate the generalization of results, or the application of inference in prediction (Moran, 2003; Clark, 2003*b*, 2005).

The third category of approaches include a suite of statistical methods built from the base of classical statistical approaches but ones that have advanced because of new computational power that allows the estimation of more complex, flexible, and robust models. In this category we would place modeling variance components (Edwards, 2004; Searle et al., 1992), multi-level models (Buckley et al., 2003), and hierarchical Bayesian models (Clark et al., 2005; Clark, 2003*a*; Gelman et al., 2004; Helser and Lai, 2004; Hooten et al., 2003). Hierarchical linear models (HLM), the focus of this paper, relate to all three of these methods as they offer a specific model structure within the hierarchical Bayesian context, a generalization of the mixed models, and specialize in estimating variance components. Although HLM can be estimated using maximum likelihood or Bayesian approaches, iterative computational techniques are required for those estimates (EM algorithm (Dempster et al., 1977) or Gibbs sampler (Gelfand and Smith, 1990) respectively). Further, although estimated with sophisticated algorithms, the structure, lexicon, and analysis of HLM use the common language of regression analysis. Results and predictions can be communicated across systems and research programs. HLM has been applied to ecological problems related to community interactions (Vazquez and Simberloff, 2004), species-area relationships

(Storch et al., 2005), spatial covariance (Gering and Crist, 2002). The focus of these applications has primarily been to account for nested observations and not to explore explicitly the role of scale in determining the associations between variables (but see Buckley et al. (2003)). Applications of HLM in ecology, however, would benefit from a protocol of analysis that can build models that develop a clear concept of the role of scale in a system, and extend its analysis to important diagnostics measures of variation and association.

In this paper we demonstrate how HLM both identifies important scales of information and measures associations that explain the information at those scales, developing this in the conceptual and mathematical framework of linear and generalized linear regression, and then demonstrate the variety of models that can be built within this framework. This method is conceptually accessible to a wide range of ecologists with a wide range of statistical experience. Our goal, therefore, is not merely to offer another approach to ecologists interested in scale issues. We hope to show that ecological studies with no explicit interest in scale must justify that omission. We also hope to offer a method that is readily applicable to studies with classic sampling and experimental designs. This method can be applied to any nested question, whether phylogenetic, geographical, experimental design, or physiological. In the following sections, we introduce the basic mathematical structure of hierarchical models and use two examples to show how HLM can be fully exploited to draw inference beyond that possible using other approaches. To firmly establish the core application of HLM, we apply maximum likelihood methods to estimate the parameters in a two-level linear model that describes the association between the amount of herbivore damage to plant leaves, plant size, and the species richness of the plots in which the plants grow. In order demonstrate the more flexible approaches of generalized HLM (HGLM) as well as some of the strengths of using Bayesian techniques

to estimate variance-covariance parameters, we present a three-level model exploring how biotic and abiotic factors at the individual plant-level, the microsite level, and the population-level influence the probability that an individual plant will flower. We conclude with a discussion of these analyses focusing on the role of scale in inference and a call for expanded incorporation of scale into quantitative analyses of natural systems.

Hierarchical linear models

Hierarchical linear models (HLM) use nested regression equations to investigate associations between variables at different scales and account for the fact that observations are related through the groups within a hierarchy. HLM can apply hypothesis tests and diagnostic reports that address not only the significance of the relationships between variables at different scales, but the strength of those relationships and their explanatory power across scales. Although the equations describing HLMs can be generalized to contain multiple predictors and link functions, a basic two-level linear model serves to demonstrate the core structure of HLM, the parameters that need to be estimated, and the inferences that can be drawn from an estimated model. Further, Raudenbush and Bryk (2002) demonstrate a protocol for model building that effectively incorporates variation at different scales into the analysis. We begin with a description of HLM model structure and then explain the protocol for model building in our first example.

Fundamentally, HLM comprises nested regression models that describe distinct levels of hierarchical data and explain how relationships within the dataset can be explained by other variables at other scales. Data can be modeled at the level they were collected, or any higher level, which will be explained in more detail below. The

first level (the individual-level in our examples) of an HLM in its linear form is the simple regression equation (all notation in this paper follows Raudenbush and Bryk (2002)),

$$Y_{ij} = \beta_{0j} + \beta_{1j} X_{ij} + r_{ij}, \tag{1}$$

where Y_{ij} is a measured response variable, which has a group-level intercept β_{0j} and is related to an individual-level predictor variable X_{ij} by the group-level regression coefficient β_{1j}. Why are these β terms 'group' variables? This stems from the fact that the residual error of the estimated relationships between the Y_{ij} response variable to the X_{ij} predictor variables, r_{ij}, is assumed in a simple linear regression model to be distributed normally with a mean of zero and variance σ^2. Because the response variable Y is associated not only with the individual $_i$ observations, but is nested within the $_j$ groups, the residuals can not be assumed to be normally distributed (to assume so would constitute pseudo-replication Hurlbert (1984)). To correct this aggregation in HLM, the first level relationships are modeled not around an overall intercept and slope, but around the intercept and slope of each of the $j = 1, \ldots, J$ level-2 groups. This corrects for the clustering of the error term and 're-normalizes' the residual error. Doing this however, results not in a single regression, but effectively in J different regression equations. To obtain an overall estimate of the relationships between the response variable and the predictors, we then use the J regressions to form two, higher-level regressions:

$$\beta_{0j} = \gamma_{00} + \gamma_{01} W_j + u_{0j}, \tag{2a}$$

and

$$\beta_{1j} = \gamma_{10} + \gamma_{11} W_j + u_{1j}, , \tag{2b}$$

where γ_{00} and γ_{01} are the level-2 coefficients for the intercept and slope respectively of these level-2 regression models (in other words, the γ parameters are group-level equivalents of the β parameters at the individual level). W_j is a level-2 predictor, and behaves as the X_{ij} does in equation (16). The level-2 random effects u_{0j} and u_{1j} are assumed to be distributed normally with means of zero and variances of τ_{00} and τ_{11} respectively. The covariance between these random effects is τ_{01}. If we substitute equations (17a) and (17b) into equation (16) we get the combined model that simultaneously describes the relationships between all predictors and response variables including their error terms at the two levels (see Table 1 for a detailed explanation of this model)(all tables and figures referred to in this part are in an appendix at the end of the part):

$$Y_{ij} = \gamma_{00} + \gamma_{01}W_j + \gamma_{10}X_{ij} + \gamma_{11}W_jX_{ij} + u_{0j} + u_{1j}X_{ij} + r_{ij}. \tag{3}$$

The fundamental difference between this combined model and models typical of single-level models is that instead of having independent random errors with constant variance (the r_{ij} term), the error term of equation (18) takes the form $u_{0j} + u_{1j}X_{ij} + r_{ij}$. We assume now that $r_{ij} \sim N(0, \sigma^2)$ and that $u_{.j} \sim N(0, \tau_{..})$, where τ is the variance-covariance matrix of the $u_{.j}$ terms, whose diagonal elements describe the variance of each u parameter. The τ variance-covariance matrix of the second level models becomes an important set of parameters as it describes between-group variance and determines whether higher-level relationships between variables are needed, significant, or explanatory (this will be clarified in the examples below).

The uncertainty ascribed to this modeled system contains random error at the individual level and the group level. The error estimation that partitions uncertainty across groups for both the mean and slope of the level-one model does so by estimating

the group level variance of the mean (u_{0j}) and slope (u_{1j}). This model provides a great deal of information about the relationships between predictor variables and the response variable and the scales at which those relationships are found. These error terms mark an essential difference between the ordinary least squares (OLS) approach which requires the deviations from the grand mean to be independent, normally distributed, and with constant variance. Because the terms u_{0j} and u_{1j} can differ between groups, their variances are not assumed equal. When these terms are null (there is no group-error variance), this model reduces to an analogue of the OLS regression model. To estimate whether these terms are null, however, we must employ iterative maximum likelihood methods, such as the EM algorithm. Furthermore, by setting various parameters of this combined model to zero, a variety of more specific questions that incorporate the scale components of the system can be tested (Table 1).

Examples

The great advantage of HLM lies in its ability to estimate complex models that incorporate scale explicitly in the analysis. Understanding the influences of biotic and abiotic factors on populations presents a challenge ideally suited for analysis using HLM. In our examples we explore the possible biotic and abiotic mechanisms that influence plant populations at different scales. These examples use datasets not specifically designed for the purpose of illustrating HLM, yet successfully demonstrate how both naturally occurring biological hierarchies (such as populations of clonal plants) and experimental hierarchies (such as nested sampling designs) can take advantage of HLM analysis.

The two-level and three-level models we constructed were estimated using two different approaches on two different datasets. The two-level model assumes maximum likelihood parameters which were estimated using expectation-maximization (EM) algorithms (Dempster et al., 1977; Raudenbush and Bryk, 2002) written in Matlab (The Mathworks, 2003) and employs traditional hypothesis tests for model diagnosis and interpretation (see Appendix A). In the three-level model we applied a hierarchical Bayes approach that uses a Markov chain Monte Carlo (MCMC) sampling procedure in WinBugs (Spiegelhalter and Best, 2000) to estimate model parameters (see Appendix B for code). It is important to note that Singer (1998) provides a clear tutorial on using SAS for the maximum likelihood estimation of HLM.

Two-level maximum likelihood model

The data used in this example are a portion of a larger demographic study. The models constructed here are designed only to advance an understanding of HLM, and not address issues in plant community ecology. Here we explore the possible relationship between several characteristics of an understory forest herb and its micro-environment with the interest of identifying associations with patterns of leaf herbivory. Why would this problem merit a hierarchical approach? First, a protocol that includes the collection of such fundamental environmental variables as species richness, soil moisture, and canopy openness will necessarily require sampling at a scale above the individual ramet of the plant. Second, the plant of interest has a natural hierarchical structure. An individual plant (a genet) is comprised of many stems (ramets) which themselves show variation in size, stage, herbivore damage, phenology, etc. These plants are further aggregated in populations that experience different habitat characteristics. Any questions that implicitly or explicitly address genotype, phenotype,

habitat traits, or demographics should distinguish between patterns and associations at the scale of the ramet, the genet, and the population. Reproductive stems vary significantly in size; they can grow between 10 and 50 cm. and have between 3 and 25 leaves. Herbivory on the leaves of *E. chlorolepis* during the summer by a host of arthropods (the primary one is the leaf-mining beetle, *Sumitrosis inaequalis* (Hispinae)) and mollusks can influence both reproductive output in any one year and age-class structure in the following year (unpublished data). Measuring the relationship between stem height and herbivore damage indicates two potential processes, If the relationship is positive, it would indicate that herbivores might key in on healthy, large stems. If the relationship is negative, it might indicate that plants can outgrow herbivore loads and therefore taller stems would display proportionally less herbivore damage. Soil moisture at the patch level can indicate increased mollusc activity (personal observation) and light levels can influence plant resources and, herbivore activity. The number of understory plant species in dense forests indicates overall microsite soil quality (e.g., higher pH and nitrate content) (unpublished data).

Species, study site, and protocol

Eurybia chlorolepis (Asteracea) is an understory perennial herb that grows in densely canopied forests in the southern Appalachian mountains of the United States. Ramets emerge from rhizomes in the early spring as either a juvenile form (rosettes) or reproductive form (with internodal stems). The reproductive ramets grow through the summer and, if conditions allow, flower, are pollinated, and set seed in the fall (September through November). This study tagged 10 stems in 20 plots and measured the herbivore damage to leaves on each stem by visually estimating a percent damage to every leaf on every stem. Damage was then averaged for each stem. Stem height is a good proxy for plant biomass as plant allometry is similar across stems.

Soil moisture was measured using gravimetric water content methods. Canopy openness was measured using hemispherical photographs and Gap Light Analyzer software Frazer et al. (1999). Species richness was calculated for all patches.

Model building, parameter estimation, and hypothesis testing

The percentage of leaf herbivory in September was chosen as a response variable to see if damage to leaves differed between patches and was predicted by plant characteristics or environmental variables. Constructing a hierarchical model, unlike a linear model, explains variation in the response variable differently at different scales and therefore requires an assessment of the scale at which variation in the response variable occurs. By fitting what Raudenbush and Bryk (2002) term an 'unconditional model,' which is effectively a one-way ANOVA model where the levels of the data hierarchy as the treatments of the single factor we can establish this baseline of variation. The 'combined model,' the analogue of Equation (18) is:

$$HERBij = \gamma_{00} + u_{0j} + r_{ij}. \tag{4}$$

Here, the percent of herbivore damage for an individual plant i in a specific patch j can be modeled as an overall average of the damage to every plant in every patch (the 'grand mean' γ_{00}) plus some difference between the average herbivore damage to plants in that jth patch from that overall mean (u_{0j}) plus the difference between the damage to that individual plant and its patch mean (r_{ij}). Thus, the variance component of every plant has two parts, the individual variance (taking into account group-variance) and group variance (taking into account individual variance). This simple formulation offers a base understanding of variation in a hierarchical system. Although rudimentary in the context of this problem, this basic understanding of the

variation of a simple response variable is almost universally overlooked in ecological studies of hierarchical design.

After assessing the scale of variation in the response, a model can be built to explain that variation in the two scales of the response variable (in this case, that of the individual stems and that of the quadrat 'patches'). How this is done depends directly on the distribution of the variance components discovered in the first model. The two distinct but not mutually exclusive additions to this 'unconditional model' then include a 'random effects' formulation (Table]1) which includes covariates at the individual level and a 'means-as-outcome' formulation, which includes covariates at the higher level to explain the intercepts among the groups. We begin with the random effects model. Using plant height as a predictor of leaf herbivore damage, and remembering that modeling the relationship between plant height and herbivore damage within groups ($\beta_{1j} * X_{ij}$) becomes $\beta_{1j} = \gamma_{10} + u_{1j}$, we have the complete random effects model:

$$HERBij = \gamma_{00} + \gamma_{10}(HEIGHT_{ij}) + u_{0j} + u_{1j}HEIGHT_{ij} + r_{ij}. \tag{5}$$

This model posits a series of relationships that combine to describe the herbivore damage to an individual stem given the height of that stem. Across-patch characteristics of herbivore damage are captured by the overall average damage to all plants γ_{00} (given the new regression relationship included in the model) and the deviation of this plant's patch intercept from that grand intercept (u_{0j}). The relationship between an individual's height and the amount of herbivore damage exhibited is partitioned into two components: first is the across-patch slope relating herbivore damage to plant height multiplied by the individual's height (($\gamma_{10} * HEIGHT_{ij}$), and second the difference between the slope of the within-patch relationship between plant height

and herbivore damage multiplied by that plant's height ($u_{1j} * HEIGHT_{ij}$). Finally r_{ij}, which is now the residual error, takes into account deviation from the expectation of this individual given all of the above model components.

The second basic expansion on the unconditional model estimates predictor variables at the second level (the patch level). It is termed the 'means-as-outcomes' model (Table 1) because the explanatory variables are set up to explain variation in group means of the response variable and not variation in individual observations of the response. If we regress herbivore damage of individual plants on patch-level soil moisture, for example, we get this combined model:

$$HERBij = \gamma_{00} + \gamma_{01} SM_j + u_{0j} + r_{ij}. \tag{6}$$

Here, we again have an individual's herbivore damage explained first by the grand intercept of herbivore damage (γ_{00}). The regression term ($\gamma_{01} SM_j$) contains the relationship between patch-average herbivore damage and the soil moisture at each site. The u_{0j} term is the residual difference between average site herbivore damage and the grand across-site damage, taking into account site soil moisture. The r_{ij} term is the difference between the herbivore damage to an individual stem and the average damage within that stem's site. Its variance should not have changed from the unconditional model, as no predictors were set up to explain that variance. Another potential expansion of the unconditional model is, which focuses on deviations of the group-level responses around the grand mean. As variation can exist at both of these levels, far more complex models are readily built.

These two models, the random effects and means-as-outcome models, are easily combined or expanded to construct more sophisticated models (Table 1). Indirect effects can be modeled as second-level predictors of first-level slopes (in other words,

second-level predictors can be used to predict the relationship between first-level variables and not the response group averages as in the means-as-outcome variables). Variance components can also be modeled with covariances.

For this example, we applied a straightforward model-building design and to estimate parameters in all models, an expectation-maximization algorithm was run in Matlab (The Mathworks, 2003). For the unconditional model, the parameters of interest were σ^2 and τ_{00}, the first- and second-level variance components respectively, as these describe how herbivore damage Y_{ij} varies both from plant to plant and patch to patch. After estimating the unconditional model, it was determined that herbivore damage did show differences across the patch (quadrat) scale (see results below). In order to explain those differences, two separate models were estimated. First, to determine whether the size of the plant (a level-one variable) predicted differences in herbivore damage, a random-coefficients regression model (Table 1) was estimated regressing mid-summer plant height against herbivore damage. The main parameter of interest in this level-1 model was β_{1j}, which describes the relationship between plant height and herbivore damage in each group j. At the group level, the parameter γ_{10} describes the average slope of this relationship and $u_{1j} * X_{ij}$) describes how the slope of this relationship varies from group to group around that average (the residuals after accounting for the overall average slope). The level-2 predictors were plant species richness, canopy openness, and soil moisture. Because there was no significant relationship between the plant height and herbivore damage (see results below), plant height was removed from the subsequent models, and the level-2 predictors were included in an means-as-outcome model (Table 1).

Results and interpretation

In every model, the EM algorithm successfully converged in under 200 iterations. The unconditional model estimated the variance at the first level, σ^2, to be 656.6. The variance at the second level τ_{00}, was 192.12, which was significantly different from 0 ($\chi^2 = 478, df = 19, P < 0.001$). This indicates patch level variation in herbivore damage to plants. To better quantify this variation, we determined the proportion of variance in the system that is described by the patch level as the interclass correlation coefficient: $\rho = \tau_{00}/(\tau_{00} + \sigma^2)$. In this model $\rho = 0.226$, indicating that 23% of the total variation in herbivory exists between patches of plants. From this starting point, we can try to explain this variation at each level.

Mid-summer plant height was not related to herbivore damage (95% CI for γ_{10} was -0.87, 0.39, overlapping 0). In the means-as-outcome model, canopy openness and soil moisture were not associated with patch level herbivore damage (95% CI forγ_{01} and γ_{02} were -1.03,1.68 and -35,06, 62.35 respectively, both overlapping 0). Species richness, however, did show a negative association with patch-level herbivory (95% CI for γ_{03} was -9.34, -3.40 with a point estimate of -6.37). By using the calculations of $\hat{\tau_{00}}$ done in the ANOVA design and in the means-as-outcome design, we can determine the between patch variation in herbivore damage accounted for by species richness. We do so by calculating $(\widehat{\tau_{00}}(\text{ANOVA}) - \widehat{\tau_{00}}(\text{Species richness}))/\widehat{\tau_{00}}(\text{ANOVA})$ or $192.12 - 65.68/192.12 = 0.6581$. Over 65% of inter-patch variation in herbivore damage is explained by the sampled species richness of the patch.

What this hierarchical approach to these community relationships offers that simple linear models to don is the explicit partitioning of variation and its explanation across scales. Every relationship identified takes into account the scale of the pattern explained. In this example, we find that the majority of variation in herbivore

damage (77%) occurs among stems. Although we did not test whether plants were genetically related, because these plants are clonal and grow in patches, this variation indicates that a selective response to herbivore damage will be muted by the distribution of damage within and not between genets (patches). It also indicates that most environmental variables, which generally influence herbivore damage at scales larger than the individual stem, likely have little influence on damage in this system. It is more likely that herbivores forage according to cues (or convenience) at the stem-level, and see patches of plants as being more or less equally accessible and edible. The smaller amount of patch-level variation (23%) indicates some association between herbivore damage and patch-scale biotic or abiotic characteristics, but not strong. Because richness was found to be negatively associated with this patch-level variation, community components could provide insight into this scale of damage that simple environmental variables cannot. We now look at how more complex models can be organized using HLM.

Three-level hierarchical Bayes model

The same computational power that has made the maximum likelihood estimation of complex HLM analyses possible has also led to the advance of Bayesian methods of model analysis. The Bayesian framework for analysis has recently been re-introduced to ecologists as a useful approach to a range of questions, most significantly for understanding complex, hierarchical ecological problems (Ellison, 2004; Clark, 2005). Among the appealing characteristics of these models is an ability to incorporate multiple experimental and observational datasets, and provide realistic parameter estimation and prediction that incorporate all model uncertainties and process variation (Wikle, 2003; Clark et al., 2005; Clark and LaDeau, 2006). Hierarchical Bayesian

methods, where parameters of interest are treated as random variables and therefore have hyperparameters that describe their distributions, are easily able to model nested data structures (Clark and Gelfand, 2006; Gelman et al., 2004). However, by designing models with the structure of HLM, a clear interpretation of nested data structures can be employed. Thus, the advantages of using Bayesian methods in an HLM framework are similar to that of the maximum likelihood estimation. The key difference is that Bayesian methods treat variance-covariance components as random variables with distributions instead of point estimates (see below), and therefore offer more realistic descriptions of these critical parameters (Raudenbush and Bryk, 2002). We choose not to debate the relative merits of Bayesian and frequentist methods in this paper, although the distinction can be important, especially as the hierarchical Bayesian approach accurately distinguishes between error in models and biological variation in models (Raudenbush and Bryk, 2002). We refer the interested reader to work that explicitly and effectively tackles this issue in statistical ecology (Ellison, 1996, 2004; Clark, 2005). This paper instead focuses on the importance of analyzing hierarchically structured data in general. We include both estimation methods to show how either approach offers insight into the scale of ecological processes, while acknowledging a growing interest in development and estimation of ecological data with Bayesian approaches.

While sharing the same basic multi-level structure as the 2-level example, the 3-level model builds on the 2-level in important ways. First, as a Bayesian model, all parameters of the model are considered random variables to be estimated (Gelman et al., 2004), and as such they are given prior distributions that are updated by the data to yield full posterior probability distributions (see Appendix A for details). Second, as with many ecological datasets, flowering is a discrete response benefiting from a generalized linear model framework. It is important to be clear that the term

'hierarchical' in 'hierarchical Bayes' refers to the use of hyper-parameters, meaning that the parameters in a model (any model) have their own distributions with other parameters that describe that distribution. The 'hierarchical' in 'hierarchical linear models' refers not to the structure of the data used in the model. These distinctions can be seen clearly in Figure II, a conceptual description of the three-level model developed in this example.

Species, study site, and protocol *Tipularia discolor* is a wintergreen terrestrial orchid found in mixed deciduous forests of eastern North America. In *T. discolor*'s southern range, a plant's single green leaf emerges above ground in late fall (end of September) and remains until spring (March-April). Flowers, if produced, are found on a single flowering stalk that that emerges in August, before the leaf emerges. Study populations range from 250 to 480 m^2 in size, each divided into 4 m^2 cells within which all plants were individually marked. Thus, the levels of the experiment were the individual plants (flowering or not), cells that reflect microsite characteristics, and the population-level grids that contain a range of cells. For this study, we use floral surveys from the late fall of 2004. Predictor variables include plant size (measured by the width of the emerging leaf), understory light levels and soil moisture at the cell level, and soil pH and soil texture at the grid level (see Appendix A for a detailed explanation of sampling protocol).

Model building and parameter estimation The 3-level Bayesian models were fit using an MCMC sampling method run in WinBugs 1.4 (Spiegelhalter and Best, 2000), and we used the R computing package (R Development Core Team, 2005) for calculating R^2 following Gelman and Pardoe (2005) (see Appendix A for details about the estimation algorithms and Appendix B for WinBugs code). Because we lack prior

information to inform the likely values of the model parameters, each parameter was given a non-informative, diffuse prior that allows the data to dominate the posterior inference (Gelman et al., 2004) (see Appendix A for details).

Bayesian results and inference

We fit these models in a step-wise fashion starting with an unconditional model and adding explanatory variables at the individual, cell, and population levels. All subsequent inference from the model comes from the posterior distributions for the parameters and the variance diagnostics. The posterior parameter distributions are summarized in Table 4 for the unconditional and fully conditional models. The size of individual plants had a positive effect on flowering probability, with a mean effect of 2.1 and over 95% of the mass of the posterior of β_1 located above 0. Note that given the logit link function, this is an effect on the log-odds of flowering. At the cell level, light availability had a positive effect on flowering, with a mean of 4.6 and its 95% probability interval slightly overlapping zero. At the population level, soil pH was significantly negatively related to flowering probability, with a mean of -2.94 and 95% interval from -5.61 to -1.12. The 95% intervals of all other explanatory variables overlap zero and thus are not considered likely to differ from zero (Figure II). Although we built these models based on parameter estimates, scoring models for selection may also be used. The deviance information criteria was developed to estimate the penalty term in hierarchical models, but its implementation proves challenging and remains somewhat controversial (see Spiegelhalter et al., 2002 and subsequent discussion).

The variance analysis of the unconditional model showed 34% of the variation among individuals, 16% among cells, and 50% at the population level. Using the

R^2 approach of Gelman and Pardoe (2005) for estimating the explanatory power of the covariates, we find that light levels explain 31% of the variation at the cell level and pH explains 78% of the variation at the grid level (see Appendix A on describing variance). Leaf width is clearly an important explanatory variable at the individual level, given its posterior distribution significantly different from zero, but because individual level variance is the constant $\pi^2/3$ (Snijders and Bosker, 1999), the percent variation explained by individual-level covariates in GLM models cannot be well estimated. Note, however, that because we have partitioned variation across scales, we can now estimate higher-level explanations of variance.

Model interpretation. As with the 2-level model, our inference about the questions of ecological interest in the 3-level model benefit from a hierarchical approach. Factors at multiple scales can influence plant flowering, from individual level traits, to microsite variation in abiotic resources and biotic impacts, to larger population level canopy, soil, or topographic effects. Indeed, although there is often significant variation from plant to plant in the likelihood of flowering, flowering synchronicity at different scales can be observed in many plant populations, suggesting the need to explicitly explore possible mechanisms at a range of scales (Satake 2004, Crone and Lesica 2004).

From our initial calculations of variance partitioning (Tables 3 and 4) we learn that as much of the variability in flowering resides among populations as within (50% of the variation in flowering probability exists at the population level), and most of the variation within populations is found among individuals (34%). At this individual level, plant leaf size has a strong influence on the probability of flowering. Variation in light availability within populations helps explain 31% of the variation from cell to cell, but differences in light availability among populations do not contribute much

to explaining overall differences in flowering. This supports the ecological hypothesis that within-population variation in canopy light transmittance, such as light gaps, is more influential than average light transmittance differences among populations, which is plausible since these populations were all located in full canopy forest sites. However, soil pH differences among the populations do help explain 71% of the variation that we see in flowering at this scale. Given the correlative nature of this study, we cannot attribute the effects to soil pH in any mechanistic sense, but it is an interesting finding nonetheless and suggestive that edaphic factors that vary on the scale of populations are important in determining reproductive behavior of these orchids. It is also important to recognize that without the explicit incorporation of scale into this analysis of flowering probability, a researcher measuring light and flowering probability from a microsite perspective (quadrat to quadrat) might over-emphasize the importance of light availability for flowering probability.

These inferences, taken together, illustrate how important it is to measure the scale at which a life history trait varies and further record the scale at which biotic and abiotic components of the system explain that variation. One might suspect that with greater future deployment of micro sensor technologies for measuring the environment at finer spatial and temporal scales, the ability to explore a range of scales may increase dramatically, and using a statistical framework that can accommodate multi-level analysis will be critical. In this study, for example, our exploration of light effects were constrained to the patch level at which we could make measurements. Were we able to measure light availability at the scale of individual plants, it would be interesting to explore its fine-scale importance relative to leaf size.

Towards a broader study of scale in ecology

Two paradigms have traditionally guided the discussion of scale and ecological systems: one describes ecological patterns as fundamentally scale-invariant (Harte et al., 2005; Marquet et al., 2005), while the other focuses on ecological patterns as hierarchical and distinct among scales (Leibold et al., 2004; Noda, 2004; Takada and Miyashita, 2004; Wu and David, 2002). Clearly both of these conceptualizations of ecological systems are appropriate depending on the question being asked (O'Neill et al., 1986). Regardless of the paradigm, however, ecologists need to explain mechanisms that influence patterns at different scales (Huston, 1999). In the case of scale-invariant systems, power law relationships, whether derived from a single process (Marquet et al., 2005) or multiple processes across scales (Allen et al., 2001), will remain only an intriguing mathematical artifact until specific mechanisms can be identified that explain why the association between two variables does not differ across scales. Investigating patterns of species diversity, for example, will entail measuring associations with species richness at different scales in order to develop a common description of the number of species observed and the area considered and potential explanatory variables (Gotelli and Ellison, 2002; Lyons and Willig, 2002). Because HLM can explore the same variable at different scales, interactions between variables, and describe the uncertainty in these relationships, it is well suited for such an inquiry.

The hierarchical paradigm of ecological systems requires even closer attention to associations across scales, as not only might the mechanisms change with scale, the inferences drawn from those relationships might change (Cadotte and Fukami, 2005; Fukami, 2004; Menge, 1992). HLM offers a powerful tool with which ecologists can explore the associations between environmental and biotic variables at different scales,

the strengths of those associations, the covariance between those associations, and the propagation of uncertainty in those relationships across scales. Although experimental designs often structures data in a hierarchical manner, many sub-disciplines are explicitly interested in biological hierarchies. Population ecology, ecological genetics, and demography inherently deal with associations among individuals, within and among populations, and the scale of inference about key variables can be crucial to the ecological and evolutionary inferences (Buckley et al., 2003; Doak et al., 1992; Scott et al., 2002). The processes driving species distributions unfold across environmental gradients at a range of spatial and temporal scales, from individual generations within microsites and populations to longer-term community level shifts over the course of decades. Accounting for scale in such analysis will be essential to any fundamental understanding of the role of ecological niches in structuring the biodiversity patterns (Chase and Leibold, 2003; Menge and Olson, 1990; Pulliam, 2000). The study of metapopulations and metacommunities, as well, are based fundamentally on a hierarchical approach to populations. Studies in these fields could benefit greatly from a more explicit incoporporation of predictive relationships between variables at the sub- and meta-population scales.

The development of ecological theory, inference drawn from empirical studies, and the confrontation of one by the other will be well served by a more expanded use of tools for explicitly analyzing scale. As computational power increases and data collection begins to reflect the potential for high-dimensional models, HLM can serve to integrate sub-disciplines, which are often focused around specific levels of organization.

Acknowledgements

The authors would like to thank Aaron Ellison for helpful comments on the manuscript. The Department of Ecology and Evolutionary Biology at the University of Tennessee helped support S.M.'s work. The *Tipularia* research was supported by NSF grant DEB-0235371 to H. Ronald Pulliam. JMD would like to acknowledge critical training in hierarchical Bayesian methods attained through the NSF-funded Summer Institute on Ecological Forecasting at Duke University.

Bibliography

Allen, A. P., B. L. Li, and E. L. Charnov. 2001. Population fluctuations, power laws and mixtures of lognormal distributions. Ecology Letters **4**:1–3.

Allen, T. F. H. and T. B. Starr. 1982. Hierarchy: perspectives for ecological complexity. The University of Chicago Press, Chicago.

Benedetti-Cecchi, L. 2001. Variability in abundance of algae and invertebrates at different spatial scales on rocky sea shores. Marine Ecology-Progress Series **215**:79–92.

Benedetti-Cecchi, L., L. Bertocci, S. Vaselli, and E. Maggi. 2005. Determinants of spatial pattern at different scales in two populations of the marine alga rissoella verruculosa. Marine Ecology-Progress Series **293**:37–47.

Borcard, D., P. Legendre, C. Avois-Jacquet, and H. Tuomisto. 2004. Dissecting the spatial structure of ecological data at multiple scales. Ecology **85**:1826–1832.

Borcard, D., P. Legendre, and P. Drapeau. 1992. Partialling out the spatial component of ecological variation. Ecology **73**:1045–1055.

Buckley, Y. M., D. T. Briese, and M. Rees. 2003. Demography and management of the invasive plant species hypericum perforatum. i. using multi-level mixed-effects models for characterizing growth, survival and fecundity in a long-term data set. Journal of Applied Ecology **40**:481–493.

Cadotte, M. W. and T. Fukami. 2005. Dispersal, spatial scale, and species diversity in a hierarchically structured experimental landscape. Ecology Letters **8**:548–557.

Chase, J. M. and M. A. Leibold. 2002. Spatial scale dictates the productivity-biodiversity relationship. Nature **416**:427–430.

Chase, J. M. and M. A. Leibold. 2003. Ecological niches: linking classical and contemporary approaches. Interspecific interactions, The University of Chicago Press, Chicago.

Clark, J. S. 2003*a*. Uncertainty and variability in demography and population growth: A hierarchical approach. Ecology **84**:1370–1381.

Clark, J. S. 2003*b*. Uncertainty in ecological inference and forecasting. Ecology **84**:1349–1350.

Clark, J. S. 2005. Why environmental scientists are becoming bayesians. Ecology Letters **8**:2–14.

Clark, J. S., G. A. Ferraz, N. Oguge, H. Hays, and J. DiCostanzo. 2005. Hierarchical bayes for structured, variable populations: From recapture data to life-history prediction. Ecology **86**:2232–2244.

Clark, J. S. and A. E. Gelfand. 2006. Hierarchical modeling for the environmental sciences: statistical methods and applications. Oxford University Press.

Clark, J. S. and S. L. LaDeau, 2006. Synthesizing ecological experiments and observational data with hierarchical bayes. Pages 41–57 *in* J. S. Clark and A. Gelfand, editors. Hierarchical modeling for the environmental sciences: statistical methods and applications. Oxford University Press.

Cushman, S. A. and K. McGarigal. 2003. Landscape-level patterns of avian diversity in the oregon coast range. Ecological Monographs **73**:259–281.

Dale, M. R. T. 1999. Spatial pattern analysis in plant ecology. Cambridge University Press.

Dempster, A. P., N. M. Laird, and D. B. Rubin. 1977. Maximum likelihood from incomplete data via em algorithm. Journal of the Royal Statistical Society Series B-Methodological **39**:1–38.

Denny, M. W., B. Helmuth, G. H. Leonard, C. D. G. Harley, L. J. H. Hunt, and E. K. Nelson. 2004. Quantifying scale in ecology: Lessons from a wave-swept shore. Ecological Monographs **74**:513–532.

Doak, D. F., P. C. Marino, and P. M. Kareiva. 1992. Spatial scale mediates the influence of habitat fragmentation on dispersal success implications for conservation. Theoretical Population Biology **41**:315–336.

Dungan, J. L., J. N. Perry, M. R. T. Dale, P. Legendre, S. Citron-Pousty, M. J. Fortin, A. Jakomulska, M. Miriti, and M. S. Rosenberg. 2002. A balanced view of scale in spatial statistical analysis. Ecography **25**:626–640.

Edwards, M. S. 2004. Estimating scale-dependency in disturbance impacts: El ninos and giant kelp forests in the northeast pacific. Oecologia **138**:436–447.

Ellison, A. M. 1996. An introduction to bayesian inference for ecological research and environmental decision-making. Ecological Applications **6**:1036–1046.

Ellison, A. M. 2004. Bayesian inference in ecology. Ecology Letters **7**:509–520.

Frazer, G., C. Canham, and K. Lertzman, 1999. Gap light analyzer (gla).

Fukami, T. 2004. Assembly history interacts with ecosystem size to influence species diversity. Ecology **85**:3234–3242.

Gelfand, A. E. and A. F. M. Smith. 1990. Sampling-based approaches to calculating marginal densities. Journal of the American Statistical Association **85**:398–409.

Gelman, A., J. B. Carlin, and H. S. S. B. Rubin. 2004. Bayesian Data Analysis, 2nd ed. Chapman and Hall CRC, New York.

Gelman, A. and I. Pardoe. 2005. Bayesian measures of explained variance and pooling in multilevel (hierarchical) models. to appear in Technometrics .

Gering, J. C. and T. O. Crist. 2002. The alpha-beta-regional relationship: providing new insights into local-regional patterns of species richness and scale dependence of diversity components. Ecology Letters **5**:433–444.

Gotelli, N. J. a. and A. M. Ellison. 2002. Biogeography at a regional scale: determinants of ant species density in new england bogs and forests. Ecology **83**:1604–1609.

Harte, J., E. Conlisk, A. Ostling, J. L. Green, and A. B. Smith. 2005. A theory of spatial structure in ecological communities at multiple spatial scales. Ecological Monographs **75**:179–197.

He, F. L. and P. Legendre. 2002. Species diversity patterns derived from species-area models. Ecology **83**:1185–1198.

Helser, T. E. and H. L. Lai. 2004. A bayesian hierarchical meta-analysis of fish growth: with an example for north american largemouth bass, micropterus salmoides. Ecological Modelling **178**:399–416.

Holling, C. S. 1992. Cross-scale morphology, geometry, and dynamics of ecosystems. Ecological Monographs **62**:447–502.

Hooten, M. B., D. R. Larsen, and C. K. Wikle. 2003. Predicting the spatial distribution of ground flora on large domains using a hierarchical bayesian model. Landscape Ecology **18**:487–502.

Hurlbert, S. H. 1984. Ecological soc amer pseudoreplication and the design of ecological field experiments. Ecological Monographs **54**:187–211.

Huston, M. A. 1999. Munksgaard int publ ltd local processes and regional patterns: appropriate scales for understanding variation in the diversity of plants and animals. Oikos **86**:393–401.

Keitt, T. H. and D. L. Urban. 2005. Scale-specific inference using wavelets. Ecology **86**:2497–2504.

Keitt, T. H., D. L. Urban, and B. T. Milne. 1997. Detecting critical scales in fragmented landscapes. Ecology and Society **1**:4.

Leibold, M. A., M. Holyoak, N. Mouquet, P. Amarasekare, J. M. Chase, M. F. Hoopes, R. D. Holt, J. B. Shurin, R. Law, D. Tilman, M. Loreau, and A. Gonzalez. 2004. The metacommunity concept: a framework for multi-scale community ecology. Ecology Letters **7**:601–613.

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

Levin, S. A. 2000. Multiple scales and maintenance of biodiversity. Ecosystems **3**:498–506.

Lyons, S. K. and M. R. Willig. 2002. Species richness, latitude, and scale-sensitivity. Ecology **83**:47–58.

Marquet, P. A., R. A. Quinones, S. Abades, F. Labra, M. F. Tognelli, M. Arim, and M. Rivadeneira. 2005. Scaling and power-laws in ecological systems. Journal of Experimental Biology **208**:1749–1769.

Maurer, B. 1999. Untangling ecological complexity: the macroscopic perspective. University of Chicago Press, Chicago.

Menge, B. A. 1992. Community regulation - under what conditions are bottom-up factors important on rocky shores. Ecology **73**:755–765.

Menge, B. A. and A. M. Olson. 1990. Role of scale and environmental-factors in regulation of community structure. Trends in Ecology and Evolution **5**:52–57.

Moran, M. D. 2003. Arguments for rejecting the sequential bonferroni in ecological studies. OIKOS **100**:403–405.

Noda, T. 2004. Spatial hierarchical approach in community ecology: a way beyond high context-dependency and low predictability in local phenomena. Population Ecology **46**:105–117.

Oksanen, L. 2001. Logic of experiments in ecology: is pseudoreplication a pseudoissue? Oikos **94**:27–38.

O'Neill, R. V., D. L. Deangelis, J. B. Waide, and T. F. H. Allen. 1986. A hierarchical concept of ecosystems. Princeton University Press, Princeton, NJ.

Pulliam, H. R. 2000. On the relationship between niche and distribution. Ecology Letters **3**:349–361.

R Development Core Team, 2005. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. URL `http://www.R-project.org`. ISBN 3-900051-07-0.

Rahel, F. J. 1990. The hierarchical nature of community persistence:a problem of scale. American Naturalist **136**:328–344.

Raudenbush, S. W. and A. S. Bryk. 2002. Hierarchical linear models: applications and data analysis methods. Advanced quantitative techniques in the social sciences, Sage Publications, Thousand Oaks, CA.

Scott, M. J., P. J. Heglund, M. L. Morrison, J. B. Haufler, M. G. Raphael, W. A. Wall, and F. B. Samson. 2002. Predicting species occurrences: issues of accuracy and scale. Island Press, Washington.

Searle, S. R., G. Casella, and C. E. McCulloch. 1992. Variance components. John Wiley and Sons, New York.

Singer, J. D. 1998. Using sas proc mixed to fit multilevel models, hierarchical models, and individual growth models. Journal of Educational and Behavioral Statistics **23**:323–355.

Snijders, T. and R. Bosker. 1999. Multilevel Analysis: An introduction to basic and advanced multilevel modeling. Sage Publications, London.

Spiegelhalter, D. J., N. G. Best, B. R. Carlin, and A. van der Linde. 2002. Bayesian measures of model complexity and fit. Journal of the Royal Statistical Society Series B-Statistical Methodology **64**:583–616.

Spiegelhalter, D. T. A. and N. Best, 2000. WinBUGS User Manual. MRC Biostatistics Unit: Cambridge.

Storch, D., K. L. Evans, and K. J. Gaston. 2005. The species-area-energy relationship. Ecology Letters **8**:487–492.

Sugihara, G. and R. M. May. 1990. Applicatoin of fractals in ecology. Trends in Ecology and Evolution **5**:79–86.

Takada, M. and T. Miyashita. 2004. Additive and non-additive effects from a larger spatial scale determine small-scale densities in a web spider neriene brongersmai. Population Ecology **46**:129–135.

The Mathworks, I., 2003. Matlab.

Thrush, S. F., V. J. Cummings, P. K. Dayton, R. Ford, J. Grant, J. E. Hewitt, A. H. Hines, S. M. Lawrie, R. D. Pridmore, P. Legendre, B. H. McArdle, D. C. Schneider, S. J. Turner, R. B. Whitlatch, and M. R. Wilkinson. 1997. Matching the outcome of small-scale density manipulation experiments with larger scale patterns an example of bivalve adult/juvenile interactions. Journal of Experimental Marine Biology and Ecology **216**:153–169.

Tolimieri, N. 1995. Efects of microhabitat characteristics on the settlement and recruitment of a coral reef fish at two spatial scales. Oecologia **102**:52–63.

Underwood, A. J. and M. G. Chapman. 1996. Scales of spatial patterns of distribution of intertidal invertebrates. Oecologia **107**:212–224.

Vazquez, D. P. and D. Simberloff. 2004. Indirect effects of an introduced ungulate on pollination and plant reproduction. Ecological Monographs **74**:281–308.

Whittaker, R. J., K. J. Willis, and R. Field. 2001. Scale and species richness: towaards a general, hierarchical theory of species diversity. Journal of Biogeography **28**:453–470.

Wiens, J. A. 1991. Ecological similarity of shrub-desert avifaunas of australia and north-america. Ecology **72**:479–495.

Wikle, C. K. 2003. Hierarchical models in environmental science. International Statistical Review **71**:181–199.

Wu, J. G. and J. L. David. 2002. A spatially explicit hierarchical approach to modeling complex ecological systems: theory and applications. Ecological Modelling **153**:7–26.

Appendix: Tables and Figures

Table 1: Parameters in the combined multi-level model.

Parameter	Description
The model: $Y_{ij} = \gamma_{00} + \gamma_{01}W_j + \gamma_{10}X_{ij} + \gamma_{11}W_jX_{ij} + u_{0j} + u_{1j}X_{ij} + r_{ij}.$	
Y_{ij}	The estimated percentage of leaf damage for individual plant i in patch. j.
X_{ij}	Initial height of individual plant i in patch j.
W_j	Species richness in each patch j.
γ_{00}	The grand mean of leaf herbivory.
γ_{01}	The mean effect of patch species richness on leaf herbivory.
γ_{10}	The average slope of the relationship between initial plant height and herbivore damage.
γ_{11}	The average effect of patch species richness on the relationship between plant height and herbivore damage.
u_{0j}	The effect of patch j on leaf herbivory, holding species richness (W) constant.
u_{1j}	The effect of patch j on the relationship between herbivore damage and plant size, holding species richness (W) constant.
$r_{ij}.$	The random effects on individual leaf damage.

Table 2: Various models described by hierarchical equations.

Model	Description
Full regression model:	
$Y_{ij} = \gamma_{00} + \gamma_{01}W_j + \gamma_{10}X_{ij} + \gamma_{11}W_jX_{ij} + u_{0j} + u_{1j}X_{ij} + r_{ij}$.	Describes relationship between the individual leaf herbivory, initial plant height and patch-level species richness.
One-way ANOVA with random effects:	
$Y_{ij} = \gamma_{00} + u_{0j} + r_{ij}$.	Describes the grand mean of leaf herbivory (γ_{00}), the effects of patch on individual leaf herbivory (u_{0j}), taking into account individual variation in leaf herbivory (r_{ij}).
Means-as-outcomes regression:	
$Y_{ij} = \gamma_{00} + \gamma_{01}W_j + u_{0j} + r_{ij}$.	Estimates how the mean leaf herbivory for each patch of plants can be predicted by species richness (W_j) taking into account the difference between patch variation in leaf herbivory (u_{0j}) and individual variation in leaf herbivory (r_{ij}).
One-way ANCOVA with random effects:	
$Y_{ij} = \gamma_{00} + \gamma_{10}(X_{ij} - \bar{X}_{..}) + u_{0j} + r_{ij}$.	Estimates the average patch leaf herbivory, accounting for how the level-1 covariate (initial plant height (X_{ij})) influences herbivore damage within each patch.
Random-coefficients regression model:	
$Y_{ij} = \gamma_{00} + \gamma_{10}(X_{ij} - \bar{X}_{..}) + u_{0j} + u_{1j}(X_{ij} - \bar{X}_{..}) + r_{ij}$.	Describes leaf herbivory as a function of the average slope of the regression between leaf herbivory and initial plant size ($\gamma_{00} + \gamma_{10}(X_{ij} - \bar{X}_{..})$) with estimates of three error terms: the effect of patch j on the mean level of leaf herbivory (u_{0j}), the effect of patch j on the slope of the regression relationship between leaf herbivory and initial plant size β_{1j} ($u_{1j}(X_{ij} - \bar{X}_{..})$, and the individual variation in leaf herbivory (r_{ij}).

Table 3: Variance Components of 3-level model.

Calculation	Description
$\frac{\pi^2/3}{(\pi^2/3+\tau_\beta+\tau_\gamma)}$	proportion of variance at level-1
$\frac{\tau_\beta}{(\pi^2/3+\tau_\beta+\tau_\gamma)}$	proportion of variance at level-2
$\frac{\tau_\gamma}{(\pi^2/3+\tau_\beta+\tau_\gamma)}$	proportion of variance at level-3

Table 4: Parameter estimates for three-level logistic models.

Unconditional model			
Parameter	Mean estimate	Lower interval	Upper interval
$\sigma^2_{individual}$	$\pi^2/3$	constant	
σ^2_{cell}	1.58	0.37	3.86
σ^2_{grid}	4.80	0.64	23.32
$\rho_{individual}$	0.34		
ρ_{cell}	0.16		
ρ_{grid}	0.50		
Conditional Model: leaf width, light, pH			
Parameter	Mean estimate	Lower interval	Upper interval
$\sigma^2_{individual}$	$\pi^2/3$	constant	
σ^2_{cell}	15.16	4.37	45.34
σ^2_{grid}	2.77	.0042	16.58
Regression Coefficients			
β_{lw}	1.55	1.11	2.06
γ_{par}	1.71	-.636	4.74
π_{ph}	-2.94	-5.61	-1.12
Percent Variation Explained (R^2)			
Cell-level	.31		
Grid-level	.78		

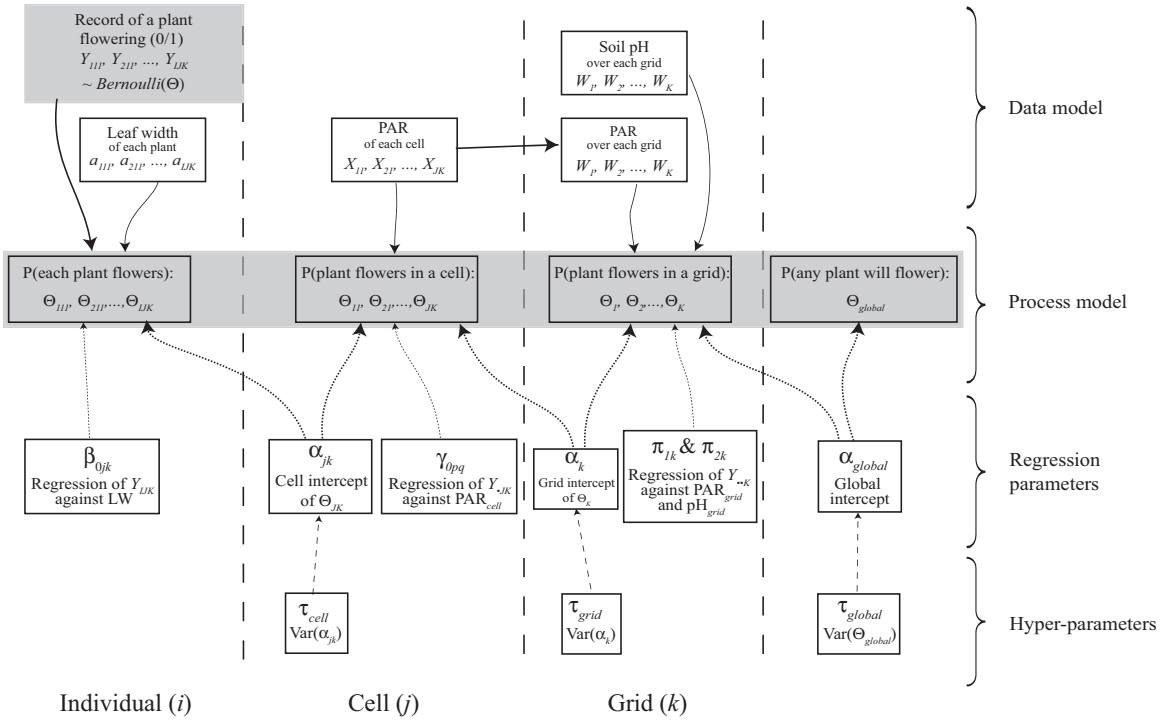


Figure 1: A conceptual map of the three-level model.

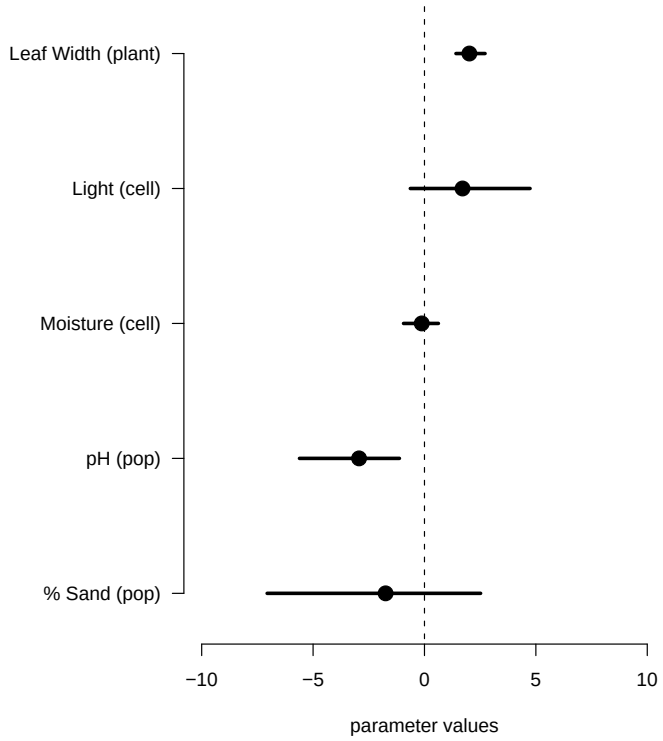


Figure 2: Scale-explicit coefficient estimates. Solid lines represent 95% posterior credible intervals for estimated effects of variables at three levels of the model. Those intervals not overlapping the zero line may be considered significantly different from zero. Leaf width of individual plants, microsite availability of light and moisture, and population level soil pH and % sand content are considered. Light refers to winter PAR readings.

Part III

Quantifying the community: using Bayesian learning networks to find structure and conduct inference in invasions biology

This part was published in Biological Invasions, vol. 7, 2005, pp. 833-844.

Abstract

One of the key obstacles to better understanding, anticipating, and managing biological invasions is the difficulty researchers face when trying quantify the many important aspects of the communities that affect and are affected by non-indigenous species (NIS). Bayesian Learning Networks (BLNs) combine graphical models with multivariate Bayesian statistics to provide an analytical tool for the quantification of communities. BLNs can determine which components of a natural system influence which others, quantify this influence, and provide inferential analysis of parameter changes when changes in network variables are hypothesized or observed. After a brief explanation of these three functions of BLNs, a simulated network is analyzed for structure, parameter estimation, and inference. Discussion of this approach to invasions biology is explored and expanded applications for BLNs are then offered.

Introduction

Since Elton (1958) recognized the potential magnitude of the threat of biological invasions, researchers have sought to determine the characteristics of invasive populations (Baker, 1965) and their non-native habitats (Howard et al., 2004; Levine and D' Antonio, 1999) that permit, catalyze or cause invasions. Intertwined in the problem of how we can understand and manage the interface between a natural system and its invader, however, lies the problem of how to understand the natural systems through which invading populations spread (Parker et al., 1999).

Ecological communities, the biotic and abiotic context of invading organisms, defy easy description, qualitatively and quantitatively (Peters, 1991). Lawton (1999, p. 178) claims that one reason community ecology has struggled to derive general laws that govern its components is that community components are highly interrelated and inferences from any one community are 'contingent' on conditions in that community. In order to quantitatively describe the way in which components of a community interact with a non-indigenous species (NIS), therefore, researchers must better quantify these 'contingencies' of the community into which it invades.

Bayesian Learning Networks (BLNs) (also called Bayesian Belief Networks) were developed in the fields of artificial intelligence and machine learning (Pearl, 1988) and have been applied in such diverse fields as medical research Sakellaropoulos and Nikiforidis (1999), structural engineering (Castillo et al., 1997), and genomics (Bockhorst et al., 2003). BLNs have not been applied to community ecology or population biology, yet are well suited to the study of combinations of direct and indirect processes that fuel these systems (Wootton, 1994).

BLNs are graphical models, which represent variables as nodes in a network and dependencies between these variables as arrows connecting the nodes (Figure 3)(all

tables and figures referred to in this part are in an appendix at the end of the part). There are other approaches to multivariate data with explicitly structured dependencies, such as path analysis and structural equation modeling (SEM) that offer methods of analyzing influence diagrams such as Figure 3 quantitatively. These methods successfully test whether data fit a given graphical model using ordinary least squares or other regression techniques (Shipley, 2000a), however, as will be discussed later, BLNs provide three approaches to graphical models that no other method offers. BLNs can use data associated with variables and quantify how those components influence one another. Having estimated a likely network of dependencies between variables, BLNs estimate the parameters that quantitatively describe the nature and strength of the nodes and connections. Finally, BLNs can be used to anticipate changes in system structure by re-estimating parameter values based on hypothesized changes in system variables. Thus, structure learning, parameter estimation, and inference all offer important methods towards a better understanding of the intricate connections of a natural community and the populations within it.

This paper describes BLNs, paying special attention to the Gaussian network model (modeling a network of continuous as opposed to discrete variables). The paper then details the components and methods required to implement structure learning, parameter estimation, and inference. Simulated data from a simple four-variable network then uses these techniques in a hypothesized scenario in which an invasive insect is defoliating indigenous trees of a forest canopy and an invasive understory herb is responding to that defoliation. The paper concludes with a discussion of the results of the simulation and a description of extended applications.

What is a Bayesian Learning Network?

A Bayesian Learning Network is a multivariate distribution that contains within its parameters the dependencies (and independencies) of the marginal variables it contains (Geiger and Heckerman, 1994). Thus, using probability calculus, a system of variables that have observations associated with them is described explicitly as a joint probability distribution. From this distribution, we can tease apart the specific parameters that describe the system's components. We can also describe the relationships between system components (effectively, the regression coefficients that describe dependencies between system parts) and model the way in which uncertainty exists and propagates through that network.

Because of the joint interdependence described by the multivariate distribution, hypothesizing changes to any one variable in the system forces the reassessment of the parameters of other variables in the system. In other words, by quantifying a system as an interconnected whole, an adjustment in our knowledge of one variable requires that we reassess our knowledge of all other variables. This feature of BLNs is not unlike the scientific process at its most resourceful.

Formally, we designate a BLN as comprising two components, a network structure B_S that describes the dependencies among the variables (Figure 3), and a set of probabilities density functions, B_P, that describe the relationships in the network graph as a multivariate density function Geiger and Heckerman (1994). The graphical representation of the network contains nodes that represent variables for which data can or has been collected. The arrows connecting those variables reflect conditional dependencies of the variables lower on the graph ('children') on those above ('parents'). A variable with no parents is called a 'root' node.

Two assumptions must hold for the graphical representation of the variables to be a Bayesian Learning Network. The graph must be a directed acyclic graph (DAG); that is, it must contain no cycles and have arrows which delimit direction. The graph must also obey the Markov condition, whereby each node in the graph is independent of all other nodes given the value of that node's parents. From probabilitiy theory, we know that the joint probability distribution of a set of variables (the essential quantitative description of a system that a BLN provides) equals the product of the marginal probability distributions of those variables if the marginal variables are independent. That is the joint probability $p(X_1, X_3)$ equals the product of the marginals, $p(X_1) \times p(X_3)$ if $X_1 \perp X_3$. In a BLN, however, we are assuming a great deal of dependence between variables in the network (in Figure 3, X_3 depends upon X_1 and X_2, and so cannot fit into this simple description of a joint distribution). That is, this joint probability that we want to derive will not equal the product of the marginals. Remembering the Markov condition from above, however, we can make the marginal probabilities of each variable in the network independent of all other variables by incorporating the probabilities of the parents of each variable. That is from (Geiger and Heckerman, 1994) we can see in Figure 3 that,

$$\rho(x_1, x_2, x_3|\xi) = \prod_{i=1}^{3} \rho(x_i|\Pi_i, \xi), \tag{7}$$

where ξ is the prior information about the system and Π_i are the parents of variable x_i. Equation (7) dictates that to formulate our BLN, we need a graph that conveys which variables are parents Π_i of every variable x_i. We need data from those variables with which we can construct density functions of the variables and their parents. Finally, we need prior information ξ on the distribution of these variables, which, as with

any Bayesian analysis, can be modeled from expert information or as prior ignorance (Bernardo and Smith, 1994).

Given, then, a graph B_S, a probability distribution that describes that graph B_P, and data D, we can find the structure of a group of variables, learn the parameters of that structure, and perform inference using both the parameters and the graph structure.

Structure Learning

Learning the structure of a certain system of variables requires that we hypothesize a structure (propose a DAG), find the posterior probability of that DAG given prior information and the data, and compare that probability to the probabilities of other possible DAGs. The number of potential DAGs that can describe a given set of variables increases super-exponentially with the number of nodes (Neapolitan, 2001), so for more than five variables, a search algorithm must be used to explore the space of potential graphs. To score the graphs, moreover, we need first to score a complete network B_{S_C}, which is any network that has arrows connecting every variable to every other. After scoring this complete DAG, we use the variance-covariance matrix obtained in its scoring to score every other DAG.

Probabilities of the data fitting a complete graph. To construct a multivariate density function for any graphical representation of a system, we need to describe a graphical relationship between the variables and then force the data in to a multivariate distribution described by that graph. From (Geiger and Heckerman, 1994) we have a multivariate distribution described by continuous variables being multivariate normal, with mean vector, $\vec{m}$ and the precision matrix T, which is the inverse of the variance-covariance matrix Σ, or Σ^{-1}; that is, $\vec{x} \sim N(\vec{m}, T)$. This multivariate

distribution can also be written as the product of conditional independent normal distributions, where each conditional distribution in the product is of the form

$$\rho(x_i|x_1,\cdots,x_{i-1},\xi) = n(m_i + \sum_{j=1}^{i-1} b_{ij}(x_j - m_j), 1/v_i), \tag{8}$$

and where m_i is the unconditional mean of x_i, v_i is the conditional variance of x_i, given values of all x_j that precede x_i in the graph,, and b_{ij} is a linear regression coefficient denoting the influence variable x_j has on x_i (akin to Yule's partial regression coefficients (Yule, 1907). If $b_{ij} = 0$, then no arrow connects x_j to x_i in the graph B_S. Because of Equation (8), a multivariate normal distribution describes any Gaussian learning network and vice versa Geiger and Heckerman (1994).

An important step in translating a DAG into a multivariate normal distribution, is building the conditional precision matrix. Shachter and Kenley (1989) offer an algorithm to transform the variance vector, $\vec{v}$, and the dependence parameters $(b_{ij}|i < j)$ into the precision matrix T. From (Shachter and Kenley, 1989), we define $T(i)$ as the $i \times i$ upper left submatrix of T, $\vec{b_i}$ as the column vector $(b_{1,i},\cdots,b_{i-1,i})$, and $\vec{b_i'}$ as its transpose. $T(1)$ is simply $1/v_1$, or the precision of the first variable. From there we iteratively build the precision matrix as

$$T_{i+1} = \begin{pmatrix} T(i) + \frac{\vec{b_{i+1}}\vec{b'_{i+1}}}{v_{i+1}} & -\frac{\vec{b_{i+1}}}{v_{i+1}} \\ -\frac{\vec{b_{i+1}}}{v_{i+1}} & -\frac{1}{v_{i+1}} \end{pmatrix}. \tag{9}$$

This precision matrix is important not only to the description of a system as a multivariate normal distribution, but for parameter learning and inference as will be seen later.

Following from this multivariate normal description of the system, scoring a probability that this representation matches the data requires that we posit prior values

for the mean vector ($\vec{m}$) and variance vector ($\vec{v}$). Although beyond the scope of this paper, the Bayesian updating of the graph parameters requires several other steps described in detail in Geiger and Heckerman (1994) and Neapolitan (2001). In short, the prior distribution of the joint density function for $\vec{m}$ and T is the normal-Wishart distribution. When this distribution is updated, using the data and precision matrix from the algorithm in Equation (19) we end up with a multivariate t distribution. From Box and Tiao (1973), we use a variation of the traditional multivariate t distribution to score the complete DAG, B_{S_C}:

$$p(D|B_{S_C},\xi) = (2\pi)^{-nm/2}(\frac{v}{v+m})^{n/2}\frac{c(n,\alpha)}{c(n,\alpha+m)}|T_0|^{\frac{\alpha}{2}}|T_m|^{-\frac{\alpha+m}{2}}, \tag{10}$$

where n is the number of variables in the network, m is the number of observations associated with those variables, v and α roughly represent the number of observations used to determine prior estimates for other parameters ('equivalent sample sizes'). $|T_0|$ and $|T_m|$ are the determinate of the prior and posterior multivariate precision matrices respectively. Further,

$$c(n,\alpha) = \left[2^{\alpha n/2}\pi^{n(n-1)/4}\prod_{i=1}^{n}\Gamma(\frac{\alpha+1-i}{2})\right]^{-1}$$

This is our final posterior probability of the data given a posited complete Bayesian network and some prior knowledge, or $\rho(D|B_S,\xi)$.

Scoring a DAG set. A DAG describing the relationship between n variables can be represented by an n by n matrix where zeros represent no connection between variables and ones represent the conditional dependence of a child on a parent. Equation

(11) is the coded matrix for the DAG displayed in Figure 3

$$\begin{pmatrix} 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix}. \tag{11}$$

By changing the off-diagonal elements of this matrix, taking care not to create a cycle within the graph, we can posit all 25 DAGs that can describe the relationship between these three variables. Each DAG describes the conditional dependencies between variables in a system. Having found the posterior probability of a complete DAG (such as when the upper triangle of the DAG code is all ones), we then use the prior and posterior precision matrices from the final derivation (10) (T_0 and T_m) to find the posterior probability of DAGs that are not complete (or assume independence between some or all variables). To create a final posterior probability of an arbitrary DAG, we apply the following equation extended from DeGroot (1970):

$$p(D|B_S, \xi) = \prod_{i=1}^{n} \frac{p(D^{x_i \Pi_i}|B_{S_C}, \xi)}{p(D^{\Pi_i}|B_{S_C}, \xi)}, \tag{12}$$

where, as in (7), Π_i are the parents of X_i. From this, we need to find the various joint and marginal probabilities as dictated by the structure of the DAG. Each probability denoted in (12) is found from the multivariate t score from (10). This is achieved by scoring the partial DAG as in Equation (10), eliminating in the calculation columns of data and rows and columns of (T_0 and T_m) that correspond to the variables not used in the calculation of each component of (12). When all potential DAGs have been scored (or a selection algorithm has been run on a subset of all possible DAGs), we select the highest scoring DAG (or select the similar high-scoring DAGs) for analysis.

Parameter estimation

Parameter estimation for Gaussian networks advances directly from the precision matrix, T, built with the algorithm in (19). For example, if we build the precision matrix for the three-variable network depicted in Figure 3, keeping in mind that the term $b_{21} = 0$ and will therefore drop out, we get

$$T = \begin{pmatrix} \frac{1}{v_1} + \frac{b_{31}^2}{v_3} & \frac{b_{31}b_{32}}{v_3} & -\frac{b_{31}}{v_3} \\ \frac{b_{31}b_{32}}{v_3} & \frac{1}{v_2} + \frac{b_{32}^2}{v_3} & -\frac{b_{32}}{v_3} \\ -\frac{b_{31}}{v_3} & -\frac{b_{32}}{v_3} & \frac{1}{v_3} \end{pmatrix}. \tag{13}$$

During the course of scoring a complete DAG, this matrix is filled with posterior values based on prior estimates and the data. When any DAG is selected, the posterior values of this matrix can be matched with the symbolic representations shown above and then solved for the various parameters. For example, if the posterior value of $T(3, 3)$ is 27.5, then the posterior variance of variable 3 is equivalent to 1/27.5. Inserting that number into other cells that contain v_3 the entire matrix can be solved. These posterior parameters then define the updated DAG model.

Inference

The application of inference calculation in BLNs consists of dictating a subset A of the variables V in graph G that have been instantiated for particular hypothesized or observed parameter values. The variable or variables in A are considered 'evidence'. Similarly, a latent variable or set of latent variables, A, that are not part of V and that extend graph, G, can be created, and these exogenous influences calculated. Inference algorithms, in their various forms, reconstruct the multivariate distribution of the graph and its parameter values in light of the evidence, A. Calculating inference

in a BLN is NP-hard (Neapolitan, 2001), and thus requires simulation algorithms for graphs with many variables. In the simulated example for this paper, only four variables are used and the symbolic inference algorithm of Castillo et al. (1997) is used.

Simulation methods and results

In order to simulate data that could intuitively be applied to the context of invasions biology, four variables were created to represent a simple network of a forest community. Two major components of a forest floral community are the canopy tree species, which can dictate light levels reaching the understory as well as soil and litter quality, and the understory plant community, which can contain seedlings and saplings of the canopy tree populations, understory herbs, ferns, and shrubs, that provide resources for a number of vertebrate and invertebrate herbivores, pollinators, and their predators. Canopy openness (and the increased light available to understory plants) has been shown to correlate with numbers of NIS (Charbonneau and Fahrig, 2004). Further, NIS have been responsible for increasing canopy openness (Kizlinski et al., 2002). In this simulation, the impact of a non-indigenous species (NIS) can thus be modeled both directly and indirectly. Directly, an invasive plant may spread through an understory community, out-competing indigenous species (Wiser et al., 1998). Indirectly, an invasive insect pest may change the canopy structure, allowing greater light to reach the forest floor (Kizlinski et al., 2002). Therefore, the four variables in the model are canopy openness, herbivore damage to an understory invasive plant, invasive plant height, and seed production from that plant. After simulating the data, finding the structure of this network, and estimating the posterior parameters of the network, I apply inference algorithms to hypothesize increased light (decreased

canopy cover) due to radical changes in forest canopy structure, reduced herbivory on the invasive understory plant, and a combination of both of these scenarios.

Simulated data

Because these variables have very different measurement units and the structure and parameter learning algorithms both use the covariance matrix in their calculations, the data need to be standardized. This four-variable network was simulated using the copula method (Perkins and Lane, 2003), where fifty observations of these variables were generated from univariate uniform random variables that were filtered through a multivariate normal copula. The correlation values used to design the network were as follows: $\rho_{21} = 0.27, \rho_{31} = 0.55, \rho_{41} = 0.0, \rho_{32} = -0.32, \rho_{42} = -0.81$, and $\rho_{43} = 0.7$, Thus, in this simulation, light through the canopy moderately influences herbivore damage, but influences invasive plant height more strongly, and is independent of seed set. Herbivory has a moderate negative association with plant height, but has a strong negative correlation with seed set. Plant height, in this model, is strongly correlated with seed production.

Structure learning

Because of the obvious physical relationships between these variables, the graph space searched using the structure learning algorithms was limited to those maintaining the designated order of these variables. That is, no graph that broke the order (hypothesized that understory plant height influences canopy openness) was included in the scoring search. This ordering is equivalent to weighting graphs with prior probabilities of occurring that is either zero or one (Geiger and Heckerman, 1994). A total of 64 graphs follow the order of simulated variables. All graphs were scored

using the structure algorithm described above. After scoring, an index of relative scoring was applied, whereby the scores were summed and each score was represented as a percentage of that total probability.

From the 64 possible graphical representations of the data, the complete network, where every variable is connected in order, scored the highest, with 61.3% of the total score. A second network, similar to the complete network excepting a missing edge from X_1 to X_4 scored second best with 32.6% of the total scoring. Thus, all other 62 hypothesized graphical representations of the data combined to share the remaining 6.1% of the cumulative score. Although the complete network scored highest and was selected for parameter learning, a network with no edge between 'light' and 'seed production' scored a solid second, with no other DAGs scoring close to these.

Parameter estimation

The top scoring DAG from the structure learning algorithm was used for parameter estimation. This DAG was entered into the algorithm in (19). Instead of using prior estimates to build the variance-covariance matrix quantitatively, the symbolic Maple kernel in Matlab (The Mathworks, 2003) was used to build the variance-covariance matrix symbolically. Each cell in the four by four covariance matrix contains the symbolic representation of the conditional variance and covariance of the variables. For example, cell 1,1 contains 'σ_1^2', and cell 2,2 contains $b'_{32}\sigma_2^2 + b'_{32}\, b_{21}\, b'_{21}\sigma_1^2 + b'_{31}\sigma_1^2 b_{21}$. From the structure learning algorithms, we have updated (posterior) estimates of the

variance-covariance matrix, which is

$$\Sigma^* = \begin{pmatrix} 0.1344 & 0.0470 & 0.0796 & 0.0223 \\ 0.0470 & 0.1333 & 0.0154 & -0.0480 \\ 0.0796 & 0.0154 & 0.1502 & 0.0861 \\ 0.0223 & -0.0480 & 0.0861 & 0.1403 \end{pmatrix}. \quad (14)$$

With the symbolic representation of each cell in the updated variance-covariance matrix and the cell values, the Maple kernel can be used to solve for each posterior parameter value. For example, combining cell 1,1 of the solution and the symbolic representation, we know that the posterior unconditional variance for the first variable is $\sigma_1^2 = 0.1344$ (this also is its conditional variance as the variable is a root node). Proceeding through the two matrices, we collect all of the posterior parameter values of the selected network. The results of this parameter estimation are shown in Table 5 and Table 6. In Table 5, the univariate (prior) regression parameters are shown with their updated multivariate (posterior) estimates. In Table 6, the univariate sample estimates for the variance are given, and the posterior unconditional and conditional estimates for the variances of the variables are given. The mean vector, $\vec{m}$ is not shown, as it is unconditional and therefore calculated simply as the vector of univariate sample means, or $\overline{x}$.

Inference

Inference using BLNs has in two basic forms. BLNs can be used to hypothesize changes in a network supposing a change in the distribution of one or more variables inside (or outside) of that network. BLNs can also predict parameter values of the network in the face of actually observed values in one or more variables. The nodes

which have new distributions or values assigned tothem are termed the 'evidential nodes', and the values of these nodes are termed 'evidence'. Therefore, the goal of inference is to determine parameters in one or more of the network's nodes in the face of evidence.

There are a number of methods to conduct inference with BLNs. With large networks, a Markov chain Monte Carlo approach, such as Gibb's sampling, is efficient (Neapolitan, 2001). In this example, because the network contains only four variables, I applied the symbolic inference approach of Castillo et al. (1997) . This approach consists updating the network's parameters given evidence using known properties of multivariate distributions:

$$f_i(x_i|x_1, \cdots, x_{i-1}) \sim N\left(\mu_i + \sum_{j=1}^{i-1} b_{ij}(x_j - \mu_j), v_i\right), \tag{15}$$

We can use this relationship to incorporate hypothesized changes to a variable's (or variables') parameter values or observed changes to the parameters in the network. Specifics are detailed in Castillo et al. (1997).

In this simulation, three inference scenarios were applied to the network (Table 7). The first scenario models the effects of an outbreak of a canopy pest on other network variables assuming that canopy tree mortality would lead to an overall increase in canopy openness by a standard deviation (from a mean of .5 to .8) and an increase in the variation in canopy openness due to the patchiness of a particular, susceptible tree species (σ_1^2) from 0.1344 to .3. The second scenario models a decrease in average herbivore damage to the invasive understory plant ($\mu_2 = 0.2$) combined with a reduced variation in that damage across sites (from $\sigma_2^2 = 0.11$ to 0.03). The third scenario models the combined influence of the insect pest and herbivore damage. Unlike the first scenario, however, while the mean canopy openness increases, the variance in

canopy openness decreases from 0.1344 to 0.03. Results from these scenarios are presented in Table 7.

Discussion

Structure and parameters

Of the three applications of BLNs to the simulated community, structure learning is the least informative in the present example, as the network modeled consisted of only four, highly correlated variables. The only two variables that were modeled with independence (X_1 and X_4) were linked in the highest scoring (complete) network (Figure 4). Structure learning, clearly, would be a more important application of BLNs if more variables are explored and many more intuitive networks could reflect the data. Networks with more than four variables, however, require search and score methods that sample the space of possible graphs instead of employing comprehensive searches. Successful methods that have been used in BLNs include genetic algorithm (Larrañaga et al., 1996), neighborhood search (Cooper and Herskovits, 1992), and others.

Once the structure of a network has been established, the parameters that describe and respond to that structure become the focus of the investigation. Because BLNs combine information related to multivariate distributions, Bayesian statistics, graphical representations of variable dependence, and inference, attending the many ways parameter values and estimates operate in BLNs is both important and powerful. Although, at base, a multivariate normal distribution contains simply a vector of means and a variance covariance matrix, because this distribution incorporates various levels of dependencies between the variables, the parameters are effectively

split between forms that are unconditional, which have no relation to values of the other variables, and conditional forms that depend on the structure of the network and the values of other variables in that network. When inference is conducted, the conditional parameters change even as the unconditional parameters are unaffected.

Dependence parameters Because the construction and analysis of a BLN incorporates Bayesian calculations, the parameters have prior values and posterior estimates. This simulated network assumed a mean vector of zeros and a vector of univariate variances of one. The univariate regression relationships between the variables were used as prior values to describe the dependencies within the network. In Table 5, the basic trend and magnitude of the univariate regressions are mirrored in the posterior estimates of b_{ij}. The difference between the prior and posterior values can be thought of as the difference between a calculated correlation in the observations of two variables, versus the relative (conditional) relationships between the variables when they are described by a single, multivariate distribution. Although the structure search included nearly independent variables in the network, parameter estimation allowed the quantification of that structure, which showed the relationship between X_1 and X_4 to be trivial ($b_{41} = -0.0885$). In the terms of artificial intelligence, the posterior values of the b_{ij} reflect our relative belief in the network structure as indicated by our graph, G (Pearl, 1988). Our posterior belief that seed production depends on canopy openness is very low. The sign of these regression parameters also indicate the direction of the effect.

Variance components Variance parameters play an important role in BLNs. First, the variance parameters can be conditional or unconditional, a difference which dictates what we infer from the value of the parameter. Second, the variance components

of a network effectively model our uncertainty in the role of the different variables in the network and the way in which our uncertainty is structured by the network. These two concepts of variance will be discussed here and in reference to conducting inference.

Univariate sample variance is a straightforward calculation (Table 6). As these variables are scaled to identical units by the copula method, their univariate sample variances are very similar. The posterior estimates of variance, however, provide two complimentary components of the multivariate parameters. The unconditional variance is extracted from the symbolic form of the precision matrix constructed from (19). This variance describes the posterior estimate of the variance of each variable in the network independent of the variance communicated by the other nodes. In a sense, it is not a completely unconditional parameter estimate because it is still a part of a multivariate distribution (the truly independent variance parameter is simply the sample estimate). Instead, the unconditional variance is the variance that is independent of the *variances* of the other components. It is not, however, independent of the dependencies between components. Because variables that occur earlier in the network explain variables further down the network, when these variables are removed from our calculation of later variables, later variables have more certainty (or lower variance) because means of these variables are described, in part, by the means of earlier variables (see Equation (7)). Therefore, variation around the mean is incorporated into the dependence parameters, which leaves the unconditional variance smaller the further down the network we go. The conditional variance is essentially the opposite. With conditional variance, the variance of each variable contributes to the variance of later variables proportional to the dependence parameters.

In Table 6 we see the unconditional variance and conditional variance of the first node is the same. The parameters of this node depend on no others, and therefore our

uncertainty in its mean value is the same when we take into account other parameters or not. The conditional variance in the network gets larger, as uncertainty in early variables propagates through the system proportional to the dependence parameters. Conversely, in the unconditional variance, uncertainty in the shape of the parameters decreases as uncertainty is attributed not to the variance, but to the dependencies dictated by the network structure.

Means The primary parameter of interest is often the mean. This is the parameter that gives us an understanding of biological characteristics of the system, such as reproductive capacity, density, or light. Our uncertainty about the mean parameter is intertwined in the network, and therefore, there is no use looking at how the mean parameters behave without assessing the conditional variance of the variables. The conditional variance is used, because, as will be explained in the discussion on inference, our belief in the mean of any one variable depends on the certainty of our belief in the means of the variables that influence it. Therefore, just as we see the mean values changing in response to changes in other variables, our certainty in the value of the later variables depends on our certainty in the values of earlier variables.

In scenario one, the mean of the first variable (canopy openness) is increased from 0.49 to 0.8. The variance of this variable is also increased. This change models what we might see if an invasive pest defoliates a forest tree: a significant increase in light reaching the canopy floor occurs, but it is highly patchy, as some areas of canopy remain intact and others are completely opened. In this scenario, we see in Table (7) that the average values for herbivore damage to the understory invader increases, but not as much as the growth of that plant, which is responding favorably to the increase in light. As seed output depends both on height and herbivory, this increase is moderate. Very importantly, however, because of the patchiness of the canopy

change, variance in (or uncertainty about) the change in network variables' means increases dramatically. Plant height, especially, varies widely under the disparate light regimes.

Scenario two models a different type of inference. Instead of hypothesizing a change in a variable's distribution, we might actually measure a change in a variable and want to predict how other variables may behave under this observed condition. In this example, we find that herbivory on the understory invader has declined remarkably across the forest, and that this measurement is fairly tight (we could model complete certainty by making the variance of the evidential node approach zero). Because the variance of the evidential node does depend on the uncertainty in the first node (in this example, again, we could lock in a certain value), the modeled change is highly influential, but not exactly mirrored in the posterior estimate of σ_2^2 (0.0465 instead of 0.03). In this scenario, not only do our estimates of the expected values for X_3 and X_4 go down, as would be expected, our uncertainty in these estimates declines also, as can be seen in the lower conditional variances of these variables. That is, we expect a drop in reproductive output in the understory invader, and are more certain that this decline will be observed.

The final scenario combines a measured increase in canopy openness and herbivory. This evidence is fairly clear (the variance in both of these evidential nodes is low), and this certainty is supported in the lower conditional variance around the mean values of the non-evidential nodes. Both invasive plant height and reproductive output is predicted to increase significantly under this scenario, in roughly equal amounts.

Potential and expanded applications

The problem of understanding the impact an NIS can have on an ecological community is two-fold. First, the threat that an NIS poses results from the features of the invaded community that lead to changes in the population dynamics of the non-native species. Second are the features of the non-native species that affect components of the invaded community. The difficulty in discovering either of these processes emerges from the inherent complexity of the invaded community. Populations of organisms (native and non-native) respond simultaneously to multiple biotic and abiotic factors at multiple spatial and temporal scales.

One strategy for trying to make sense of dynamic community systems has been to isolate mechanisms or processes that are thought to be most important to the population dynamic of interest. In controlled experiments, one to several components of the system hypothesized to influence the populations of the organisms within it are altered, while all other components are kept as invariant as possible. Controlled experiments, however, take as preliminary assumptions that few processes dictate population dynamics, and that these processes are known *a priori*.

Although some sample systems and long term research programs can provide unique insights into the role of invasives (Turner et al., 2003), many important invasions are researched after they have become destructive, and the potential system components that enabled the invasion are extensive, highly correlated, and potentially no longer present in the disturbed community. Every population of NIS at every point in its non-native range is, in effect, an ecological experiment. While a researcher must suppose that a few of many possible components of the system are important to the population dynamics of the NIS, the NIS can explore many more components over a longer time period. If these 'experiments' do not work, we do not

identify the NIS as important or potentially dangerous. When the 'experiment' does work, and the NIS population grows, we are at a disadvantage in trying to tease apart the important pathways that led to the establishment and growth of the population. In effect, we have little data on why invasions fail, and many possible reasons why they succeed.

In the face of such complex dynamics, this paper advocates the application of statistical and experimental approaches to NIS populations that explicitly incorporate multiple system components and pathways into the exploration and modeling of invaded communities. Although a number of methods exist for these applications Shipley (2000b), BLNs not only allow such exploration of system structure, and allow the quantification of that structure, BLNs allow us to then model how our understanding of the system may change in the face of new or hypothesized evidence. These applications recommend BLNs for analyses that tackle the difficult but important range of ecological data analysis from exploratory analysis to quantification to prediction. Estimating the structure of complex networks using data can give invasions biologists important information about which parts of complex and complicated natural systems may influence or be influenced, directly and indirectly, by non-native populations.

Beyond these applications, however, BLNs can be used in conjunction with other modeling techniques used in invasions biology, such as population matrix models, ODE models, individual based and spatially explicit models. Including variables in a BLN that can be incorporated into these other models (such as growth rates, competition coefficients, or dispersal ability) can offer a more powerful understanding of model components. BLNs can offer estimates of parameter values, show potential direct and indirect influences on these values, model uncertainty in these parameters, and predict changes in model parameters due to changes in the system that influences

them. It is also possible to incorporate BLNs explicitly in an iterative model, whereby output from one model can act as evidence in a BLN which then produces a change in another network variable that can then be reincorporated into the model. In this sense, BLNs can act as a powerful bridge between field observations, mathematical models, and predictive applications. Because BLNs combine a DAG with parameter estimates, directed graph theory can be applied to BLNs in order to determine causal pathways in the system (Geiger and Heckerman, 1994; Shipley, 2000b). Within the context of invasions biology, there should be no limits on the types of methods applied to this complex and critical ecological problem.

Bibliography

Baker, H. G., 1965. Characteristics and modes of origin of weeds. Pages 154–184 *in* H. G. Baker and G. L. Stebbins, editors. The genetics of colonizing species. Academic Press, New York.

Bernardo, J. M. and A. F. Smith. 1994. Bayesian theory. Wiley series in probability and mathematical statistics, Wiley, New York.

Bockhorst, J., M. Craven, D. Page, J. Shavlik, and J. Glasner. 2003. A bayesian network approach to operon prediction. Bioinformatics **19**:1227.

Box, G. and G. Tiao. 1973. Bayesian inference in statistical analysis. Addison-Wesley series in behavioral science, Addison-Wesley Publishers, Reading, Mass.

Castillo, E., J. M. Gutierrez, A. S. Hadi, and C. Solares. 1997. Symbolic propagation and sensitivity analysis in gaussian bayesian networks with application to damage assessment. Artificial Intelligence in Engineering **11**:173–181.

Charbonneau, N. C. and L. Fahrig. 2004. Influence of canopy cover and amount of open habitat in the surrounding landscape on proportion of alien plant species in forest sites. Ecoscience **11**:278–281.

Cooper, G. F. and E. H. Herskovits. 1992. A bayesian method for the induction of probabilistic networks from data. Machine Learning **9**:309–347.

DeGroot, M. H. 1970. Optimal Statistical Decisions. McGraw-Hill, New York.

Elton, C. S. 1958. The ecology of invasions by animals and plants. The University of Chicago Press, Chicago.

Geiger, D. and D. Heckerman. 1994. Learning gaussian networks. Proceedings of the Tenth Conference on Uncertainty in Artificial Intelligence pages 235–243.

Howard, T. G., J. Gurevitch, L. Hyatt, and M. Carreiro. 2004. Forest invasibility in communities in southeastern new york. Biological Invasions **6**:393–410.

Kizlinski, M. L., D. A. Orwig, R. C. Cobb, and D. R. Foster. 2002. Direct and indirect ecosystem consequences of an invasive pest on forests dominated by eastern hemlock. Journal of Biogeography **29**:1489–1503.

Larrañaga, P., M. Poza, Y. Yurramendi, R. H. Murga, and C. M. H. Kuijpers. 1996. Structure learning of bayesian networks by genetic algorithms: A performance analysis of control parameters. IEEE Transactions on Pattern Analysis and Machine Intelligence **18**:912.

Lawton, J. H. 1999. Are there general laws in ecology? Oikos **84**:177–192.

Levine, J. M. and C. M. D' Antonio. 1999. Elton revisited: A review of evidence linking diversity and invasibility. Oikos **87**:15–26.

Neapolitan, R. E. 2001. Learning Bayesian Networks. Pearson, Prentice Hall, Upper Saddle River, NJ.

Parker, I. M., D. Simberloff, W. M. Lonsdale, K. Goodell, M. Wonham, P. M. Kareiva, M. H. Williamson, B. Von Holle, P. Moyle, J. E. Byers, and L. Goldwasser. 1999. Impact: toward a framework for understanding the ecological effects of invaders. Biological Invasions **1**.

Pearl, J. 1988. Probabilistic reasoning in intelligent systems: networks of plausible inference. Morgan Kaufmann, San Mateo, California.

Perkins, P. and T. Lane, 2003. Monte-carlo simulation in matlab using copulas.

Peters, R. H. 1991. A critique for ecology. Cambridge University Press, Cambridge, UK.

Sakellaropoulos, G. C. and G. C. Nikiforidis. 1999. Development of a bayesian network for the prognosis of head injuries using graphical model selection techniques. Methods of Information in Medicine **38**:37.

Shachter, R. D. and C. R. Kenley. 1989. Gaussian influence diagrams. Management Science **35**:527–550.

Shipley, B. 2000a. A new inferential test for path models based on directed acyclic graphs. Structural Equation Modeling **7**:206–218.

Shipley, B. 2000b. Cause and correlation in biology: a user's guide to path analysis, structural equations and causal inference. Cambridge University Press, Cambridge.

The Mathworks, I., 2003. Matlab.

Turner, M. G., S. L. Collins, A. E. Lugo, J. J. Magnuson, T. S. Rupp, and F. J. Swanson. 2003. Disturbance dynamics and ecological response: The contribution of long-term ecological research. Bioscience **53**:46–56.

Wiser, S. K., P. W. Clinton, and K. H. Platt. 1998. Community structure and forest invasion by an exotic herb over 23 years. Ecology **79**:2071–2081.

Wootton, J. T. 1994. Putting the pieces together: testing the independence of interactions among organisms. Ecology **75**:1544–1551.

Yule, G. U. 1907. On the theory of correlation for any number of variables, treated by a new system of notation. Proceedings of the Royal Society Series A **79**:182–193.

Appendix: Tables and Figures

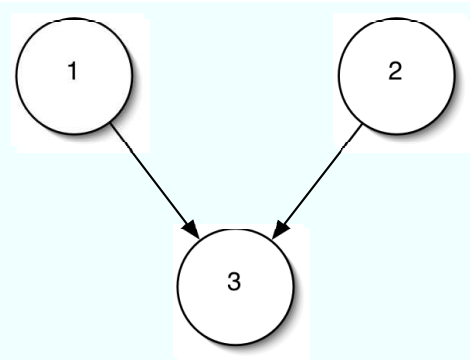


Figure 3: A three-node Bayesian Learning Network, with independence modeled between variables one and two. The dependencies between variables X_1, X_2, and X_3 are designated by b_{ij}, where i is the node lower on the graph and j is the variable that influences that node. μ_i and σ_i denote the unconditional mean and conditional variance of the variables.

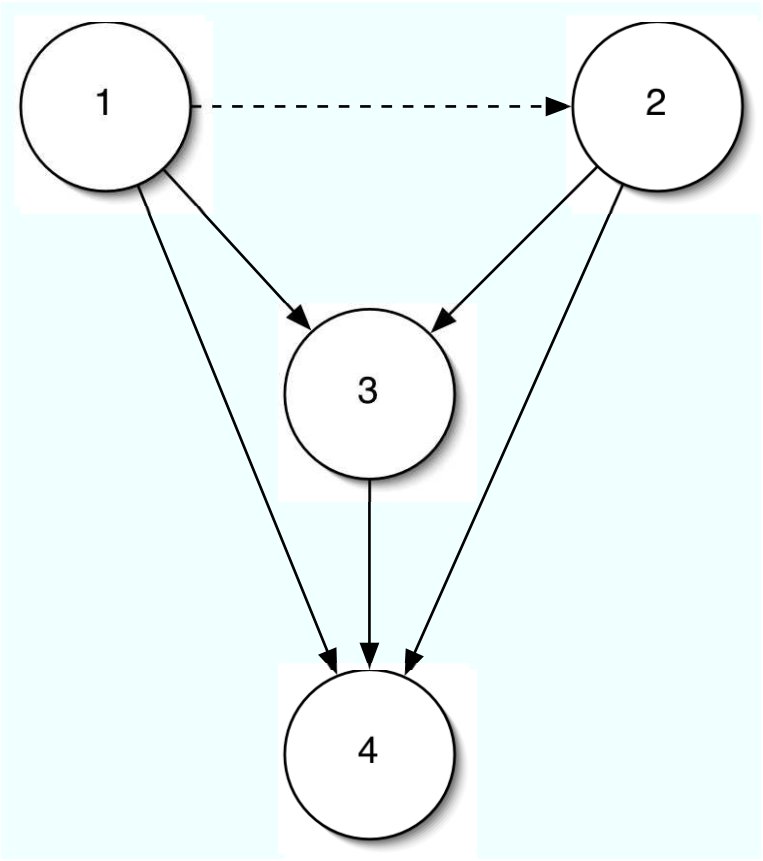


Figure 4: The highest scoring DAG. Nodes reflect simulated components of a forest system: 1) canopy openness, 2) herbivore damage to an invasive understory plant, 3) invasive plant height, and 4) invasive plant seed set. The dashed line indicates an included edge that was not generated from the copula, but reflects a very weak dependence in the parameter value of the connection.

Table 5: Dependence parameters. Prior regression coefficients and posterior estimates of the dependence between variables in the selected network.

Parameter	Prior values	Posterior estimates
b_{21}	0.2834	0.3967
b_{31}	0.5174	0.6967
b_{41}	-0.0203	-0.0885
b_{32}	-0.2448	-0.2781
b_{42}	-0.7279	-0.4311
b_{43}	0.5894	0.7139

Table 6: Variance components of the network. Posterior unconditional and conditional variances estimates. Prior values consisted of a vector of ones.

Parameter	Univariate	Posterior estimates	
	sample estimate	Unconditional	Conditional
σ_1^2	0.0842	0.1344	0.1344
σ_2^2	0.0824	0.1145	0.1333
σ_3^2	0.0864	0.0928	0.1502
σ_4^2	0.0740	0.0664	0.1403

Table 7: Inference. Posterior parameter values and updated parameter estimations based on inference from evidence. A star (*) denotes an evidential node.

	Starting parameters	Inference scenarios		
		1	2	3
σ_1^2	0.1344	*0.3	0.1344	*0.03
σ_2^2	0.1333	0.1535	*0.0465(0.03)	*0.0337(0.03)
σ_3^2	0.1502	0.2084	0.1492	0.1126
σ_4^2	0.1403	0.1449	0.1206	0.1177
μ_1	0.4851	*0.8	0.4851	*0.8
μ_2	0.4901	0.6003	*0.2	*0.2
μ_3	0.518	0.7046	0.3491	0.6754
μ_4	0.5045	0.5566	0.4573	0.6782

Part IV

The scale of seed fate in a perennial herb

Introduction

The production of viable seeds by a plant is often reported as a measure of fitness (Mothershead and Marquis, 2000; Mena-Ali and Rocha, 2005; Louda and Potvin, 1995; Maron, 1998; Primack and Kang, 1989). Even among plants able to reproduce vegetatively, dispersing viable seeds can constitute an essential component of overall plant population dynamics (Kanno and Seiwa, 2004; Verburg and During, 1998), determine spatial population structure (Rautiainen et al., 2004), allow escape from competition (Nishitani et al., 1999), rescue locally extinct sub-populations in a metapopulation (Stocklin and Winkler, 2004; Willi and Fischer, 2005; Dupre and Ehrlen, 2002), maintain genetic diversity (Chung and Chung, 1999; Gabrielsen and Brochmann, 1998), and propagate advantageous traits through a population (Slatkin, 1985). A detailed understanding of patterns of flower production, pollination, and viable seed survival to dispersal is therefore essential to interpreting ecological and evolutionary questions related to sexual reproduction in plants.

When an individual stem is sexually reproductive, flowers represent the result of long-term (or potentially life-time in monocarps) investment in resource acquisition, allocation, and growth (Metcalf et al., 2003), but this investment that does not guarantee successful sexual reproduction. During the interval between bud development and the successful dispersal of viable seeds, a number of sub-processes can positively and negatively affect the plant and drive the fate of seed set. The relative number of flowers that a plant produces and the pollination of the those flowers contribute positively to overall viable seed dispersal. Flower and seed abortion and predation reflect negative influences on final viable seed dispersal. Myriad biotic and abiotic influences on these sub-processes create a network of potential interactions that ultimately dictates the reproductive contribution of a mature stem. The environmental forces that

affect the production of seeds can therefore exert a strong influence on strategies of floral production, resource allocation, predation defense, and nectar production.

Plant species that exhibit both clonal and sexual reproduction (Washitani et al., 1996) tend to grow as patches of clones that are produced vegetatively. Within these 'genets,' specific stems can become sexually reproductive (i.e., produce flowers and set seeds). These two reproductive strategies, vegetative and sexual, result in a hierarchically structured plant community comprised of genetically similar or identical stems that exist in patches (Chung and Chung, 1999). The stems within patches can vary in size, life history stage, and experience herbivores, pollinators, seed predators, and pathogens that their clonal neighbors do not. Yet, environmental variables such as soil moisture, soil nutrients, light, and temperature may remain relatively homogenous across the patch. Thus the scale at which patterns of seed fate are measured can play a large role in determining the importance of which processes ultimately dictate the successful production of seed. Many hypotheses about the pattern of and influences on these sub-components of seed production imply a focal scale at which the process occurs. The implicit assumption of a particular scale at which seed fate is determined has not been tested in one study, yet holds important consequences for how inference can draw from an analysis. this study simultaneously measures how seed set is influenced by production, pollination, abortion and predation at the stem scale and the patch scale. The following subsection and Table 8 reviews hypotheses that explain these components of seed fate at those two scales (Figure 5 displays the relationship between these components of seed fate) (all tables and figures referred to in this part are in an appendix at the end of the part).

Inflorescence and floret production. The number of inflorescences produced by a plant represents its total potential reproductive output and reflects an explicit

allocation of resources to those structures. Production of flowers is therefore viewed as an important component of possible trade-offs, as investment in floral structures may come at the expense of shoot growth, root growth, or vegetative reproduction. Variation in the number of inflorescences produced by a stem or a genet may reflect plant responses to various exogenous or endogenous processes at the scale of a patch of plants sharing a similar micro–habitat, or at the stem-scale, where the plant responds to finer scale physiological and biotic stimuli. Exogenous influences on inflorescence production can include patch-level variables such as light levels (Gehring and Delph, 2006), soil nutrients (Munoz et al., 2005), or stem-level variables such as plant height or biomass, leaf herbivore damage (Ehrlen, 2002, 1997; Hersch, 2006; Berjano et al., 2006) or finer-scaled soil characteristics. Endogenous influences on inflorescence production can include genotype (Karkkainen et al., 1999) (which is patch-level in clonal plants), and stem-level components such as stem age (Ehlers and Olesen, 2004).

Pollination. Among plants that exploit animals as pollen vectors, pollination efficacy can be highly variable and depend on biotic and environmental variables (Herrera, 1995). Pollinator flight temperatures, floral recognition, nectar acquisition, flower damage, and the color and size of flower corollas can all potentially influence floral visitation and successful pollination (Proctor et al., 1996; Herrera, 1995; Lehtila and Strauss, 1997; Louda and Potvin, 1995; Strauss, 1997). These mechanisms to which the plant responds can operate at different scales. Environmental variables, such as temperature and light are generally important at the scale of the patch, as multiple stems share roughly the same micro-environment (and can even average the environment using rhizomes (Salzman and Parker, 1985), but patches that are in gaps or forest understory, or have different flowering densities can show different levels of pollination due to light and temperature (Herrera, 1995). Mechanisms that influence

pollination such as herbivore damage, nectar production, and stem height can show both intra- and inter-patch variation.

Floret and seed abortion. After inflorescences are produced, individual florets and seeds or entire inflorescences may be aborted by the plant. This can occur before or after pollination. Floret or seed abortion occurs when, after inflorescence formation, resources are withdrawn from inflorescence development, stymying growth and leading to the desiccation or decay of the entire inflorescence or individual seeds within the inflorescence. Given that floral structures reflect allocation of resources, an important question in the evolution of flower production revolves around why many plant species produce far more flowers than the seeds that result from their pollination (Burd, 1998; Ehrlen, 1991; Stephenson, 1981). Exogenous and endogenous mechanisms can influence seed abortion patterns. Exogenous conditions that trigger abortion include herbivore or floret damage (Balestri and Cinelli, 2003; Wise and Cummins, 2006), extreme temperatures (Young et al., 2004), and limited resources (Melser and Klinkhamer, 2001). Endogenous causes of seed abortion could include 'poor' genotypes of the seeds or the maternal plant (Karkkainen et al., 1999) or amplification of male flower function (pollen production) which can then be balanced with resource availability for seed development through selective abortion (Burd, 1998). Exogenous causes would more likely affect multiple stems, while endogenous causes would more likely structure seed abortion at the stem (except the maternal genotype hypothesis, whereby clones would be equally poor at producing viable seeds).

Floret and seed predation. Pre-dispersal seed predation has been shown to adversely influence final seed production in a number of plant systems (Louda, 1982; Louda and Potvin, 1995; Guretzky and Louda, 1997; Albrectsen, 2000; Pilson, 2000).

Seed predators range from mammals, such as mice and deer (Herrera et al., 2002) to a variety of insect larvae (Albrectsen, 2000). The pattern of larval seed predators may be dictated by plant resources (Niesenbaum, 1996) or oviposition choices made by the adult female flies or by the developmental status of seeds in the inflorescences. Factors such as floral density and sunlight can all influence oviposition choice larval insect seed predators (Brody, 1992; Ehrlen, 1996; Garcia-Robledo et al., 2005; Sheppard et al., 1994) at the intra-patch scale. Intra-patch variation in seed predation could be due to difference in the number of mature, pollinated seeds that a stem produces.

This study does not seek to prove the causes of seed fate, but to organize the many possible influences by measuring the scale at which these processes occur. Observational studies are conducted in situations where controls cannot be assigned by a researcher and are instead incorporated not in the experimental design but in the statistical analysis. Because the processes that influence seed fate operate in concert in nature, and are themselves influenced by the environment, this study was conducted by taking measurements in the field and then using statistical models to tease apart effects. A full Bayesian estimation of hierarchical linear models (HLM) was used to estimate variation floral production, pollination, seed abortion and predation in the understory herb *Eurybia chlorolepis* (Asteraceae) at two scales: that of the ramet (the reproductive stem in a clonal or rhizomous plant) and that of the patch (the cluster of ramets, often genetically identical stems that share common microhabitat features).

An important part of this study was designating the unconditional (unexplained) variation in the response variables at the two scales. HLM effectively estimates the variance components of these variables. Employing a hierarchical Bayesian approach to the estimation of HLM model parameters enables a level of analysis that maximum likelihood methods cannot achieve. Bayesian HLM accurately describes the

distribution of regression and variance components across scales, uses knowledge of parameters at the stem scale to support the estimation of components at the patch scale (and vise versa), and models biological variation as well as statistical uncertainty (Clark and Gelfand, 2006; Clark, 2005). Furthermore, the very nature of the variables of interest (such as inflorescence and seed abortion and seed predation) lead to significant non-ignorable missing data. A computational approach to estimating the Bayesian model parameters, a Gibbs sampler (Gelfand et al., 1990) constructs models every iteration for hundreds of iterations. This is an ideal tool for estimating these missing data (Gelman et al., 2004).

Site, species, and methods

Site

Data were collected from plots in the Cosby Ranger District, Cocke County, Tennessee in Great Smoky Mountains National Park (GSMNP). Twenty quadrats (1m×1 m) were established in three extant populations of *E. chlorolepis* within a single watershed. Within these quadrats, all ramets of *E. chlorolepis* were labeled in 2004 (total ramets = 877). A subset of these patches were used in this analysis of seed fate. Of the 20 patches, eight of them contained enough reproductive stems with inflorescence heads to use in data analysis, and these eight quadrats contained a total of 136 reproductive ramets with at least one inflorescence head. Of the twelve patches excluded from this analysis, four patches had no reproductive stems, and eight patches had four or fewer reproductive stems with flowers, which provides too few samples to estimate the parameters for the conditional models at the inter-patch scale. All eight patches included in this analysis were located in a five hectare area of old-growth hemlock

and mixed hardwood forest. A minimum inter-patch distance of 10m was maintained, and each patch was set within a larger colony, separated from other colonies by at least 10m. Thus, each patch was separated from other patches not just by space but by forest understory with no stems of *E. chlorolepis*.

Species

Mountain aster, *Eurybia chlorolepis* (Burgess) Nesom, is an understory perennial herb found in the middle elevations of the Southern Appalachians. *E. chlorolepis* reproduces vegetatively through the propagation of new ramets from the base of old ramets and from ramets growing from a rhizome. The rhizome that connects ramets of *E. chlorolepis* is found just below the soil surface. Roots from the rhizome and ramets are generally shallow. Flowering stems produce composite inflorescences that are hermaphroditic and generalist pollinated.

The life-cycle of *E. chlorolepis* has a typical perennial stage structure (Figure 6) (Horvitz and Schemske, 1995). Sprouts new growth in the winter in the form of small ($1-3$cm leaf length) two- to four-leaved ramets. These remain small until the spring, when they grow as either juvenile plants in the form of rosettes (Figure 7a) or as reproductive individuals with inter-nodal stems Figure 7b). Reproductive forms can reach $0.5m$ in height. In August and September, reproductive ramets produce inflorescences consisting of many small disc and ray florets Figure 7c). These inflorescences have yellow disc flowers when producing pollen or are receptive of pollen. The disc flowers become white, red, or pale red when no longer actively reproductive (Figure 7c). During October, pollinated ovaries mature to seeds that are gravity and water dispersed. Seeds germinate during the winter and produce seedlings that emerge in spring (Figure 7d). Seedlings grow the entire next year as small juveniles

(they do not become directly reproductive). Thus, most patches of plants connected by rhizomes are at least several years old.

Although the data from this study were collected in one year, these populations had been monitored for four. Plants in the two years previous to this study produced virtually no seeds. The first of those years was unusually moist and mollusk predators devastated the patches broadly and evenly. The second year was unique in experiencing the remnants of three hurricanes within a several weeks with concomitant record rainfall. All reproductive stems wilted during this period because of root saturation. Reproductive stems and inflorescences in 2005 were abundant and similar to levels in 2002, the first year of monitoring, but not of seed collection. This study, therefore, was conducted in a year when all patches were experiencing positive conditions for seed production, and thus the differences in seed output between stems and patches would highlight potential differences in site and plant qualities.

Methods

Collection and measurement methods

All inflorescence heads were collected after seed set in September and October, 2005. Because of the remote nature of the patches, daily collection trips were not feasible, however, wholesale collection might have underrepresented seed predation (as the seeds would have been removed from the natural patches before seed predators had done damage). This balance between realistic development to dispersal and the comprehensive collection of seeds was maintained as best possible, resulting in a significant, but incomplete collection (see 'Missing data' below). Inflorescence heads were returned to the lab and refrigerated at 4°C until analysis.

A suite of stem variables was collected. Stem height and leaf damage were measured after plants reached their maximum annual height in late August, 2005. Herbivore damage to leaves may continue, but late August marks the peak pre-reproductive plant state. Herbivore damage was estimated by assigning a visual percent leaf-damaged to each leaf of every stem. Individual leaf damage was then averaged across a stem to give an estimate of total stem-level leaf damage. Most damage was done by the larval form of a leaf mining beetle (*Sumitrosis inaequalis* (Hispinae)), mollusks, and, more rarely, an assortment of Lepidopteran larvae, Orthoptera, and adult beetles (including the adult form of *S. inaequalis*). Stem age was estimated by the variable 'Old nodes' (ON) refers to nodes on the caudal root of plant stems resulting from the die-back of previous reproductive stems. These nodes represent a rough estimate of stem age (juveniles can follow reproductive stems and leave no node, therefore complicating the direct estimation of ramet age). Plant biomass was estimated from the regression of twenty dried and weighed *E. chlorolepis* stems (minus inflorescence heads -removed so that estimated stem biomass could be used to predict inflorescence production) against plant height and leaf herbivory. The regression was significant ($P < .01$), and these two predictor variables explained over %80 of plant biomass. Inflorescence number reflects the number of inflorescence heads produced by any one stem.

Patch-level variables collected consisted of soil moisture, canopy light transmittance, richness, and percent ground cover. Soil moisture was measured as gravimetric water content (GWC), the percent of water in the soil found from dividing the dry weight of the mix of five samples from immediately around a patch subtracted from the wet weight, divided by the wet weight. Light was measured using hemispherical photos taken from 1m above the ground at the center of every patch. These photos

were then filtered into canopy cover and openness. Using Gap Light Analyzer software (Frazer et al., 1999), estimates of direct, diffuse, and total light transmittance through the canopy at each patch were calculated. Direct light transmittance was used in analyses as it is most likely to be important in the dark forest understory. Plant species richness was determined at each patch in mid-July. Percent ground cover was calculated from digital photos of patches taken 1.6m above ground. These photos were filtered using Adobe Photoshop to a threshold of white and black reflecting vegetative cover and bare ground respectively. Ground cover, as a variable, is the percentage of pixels in the photograph of a patch that are white.

Determining seed fates

Seeds were removed from the receptacle and counted using a dissecting microscope and classified as follows. *Aborted seeds* (Figure 8d) are distinguished by their black color, often diminished length, and brittle texture. Inflorescence heads that never open are found to share these qualities of aborted seeds and assigned the class 'aborted'. The primary source of *seed predation* on *E. chlorolepis* is a tephritid fly (insect larva are difficult to identify without adult forms), a family of common seed predators of composite flowers (Albrectsen, 2000). Adult female flies lay a single egg in an inflorescence. Based on the pattern of seed predation around a ramet's inflorescence heads, one or two larvae move around the ramet and eat seeds, molting through instars, until forming a casing in which they metamorphose into adult flies. Over the course of development, larvae shift from feeding within seeds, as is evidenced by entrance holes (Figure 8e) to consuming entire seeds(minus the pappus). Additional damage was associated with an unidentified lepidopteran larva (on three inflorescence heads) (the larvae shared similar form, but were withered and unidentifiable). In

investigating *pollination*, there are clear morphological differences between pollinated ('viable seed' (VS)) and unpollinated ('inviable seed' (IS)) seeds. Inviable seeds are long and thin, wrinkled, soft, and usually range in color from light green to dark brown (Figure 8b). Viable seeds are long and cylindrical, hard-shelled, and deep brown (Figure 8c). Seeds termed 'viable', in this study, are further distinguished into those that have been predated ('eaten viables' (EV). 'Pollinated seeds' therefore include VS and EV seeds.

Statistical Methods

Hierarchical models and the Gibbs sampler

The method employed to measure associations across scales was hierarchical linear models (HLM) (Raudenbush and Bryk, 2002). HLM consists of nested regression models that describe distinct levels of hierarchical data and explain how relationships within the dataset can be explained by other variables at the same or other scales. The first level (the 'stem-level' in this study) of an HLM in its linear form is the simple regression equation (all notation follows Raudenbush and Bryk (2002)),

$$Y_{ij} = \beta_{0j} + \beta_{1j} X_{ij} + r_{ij}. \tag{16}$$

Here Y_{ij} is a measured response variable, which has a group-level ('patch-level in this study) intercept β_{0j} and is related to an individual-level predictor variable X_{ij} by the group-level regression coefficient β_{1j}. These β parameters are termed 'group' variables because the residual error of the estimated relationships between the Y_{ij} response variable to the X_{ij} predictor variables, r_{ij}, is assumed in a simple linear regression model to be distributed normally with a mean of zero and variance σ^2. Because

the response variable Y is associated not only with the individual i observations, but is nested within the j groups, the residuals can not be assumed to be normally distributed (violating this assumption would constitute pseudo-replication (Hurlbert, 1984)). To correct this aggregation in HLM, the first level relationships are modeled not around an overall intercept and slope, but around the intercept and slope of each of the $j = 1, \ldots, J$ level-2 groups. This corrects for the clustering of the error term and 're-normalizes' the residual error. Doing this however, results not in a single regression, but effectively in J different regression equations. To obtain an overall estimate of the relationships between the response variable and the predictors, the J are collected into two, higher-level regressions:

$$\beta_{0j} = \gamma_{00} + \gamma_{01} W_j + u_{0j}, \tag{17a}$$

and

$$\beta_{1j} = \gamma_{10} + \gamma_{11} W_j + u_{1j}, , \tag{17b}$$

where γ_{00} and γ_{01} are the level-2 coefficients for the intercept and slope respectively of these level-2 regression models (in other words, the γ parameters are group-level equivalents of the β parameters at the individual level). W_j is a level-2 predictor, and behaves as the X_{ij} does in equation (16). The level-2 random effects u_{0j} and u_{1j} are assumed to be distributed normally with means of zero and variances of τ_{00} and τ_{11} respectively. The covariance between these random effects is τ_{01}. Equations (17a) and (17b) are substituted into equation (16) to get the combined model that simultaneously describes the relationships between all predictors and response variables including their error terms at the two levels:

$$Y_{ij} = \gamma_{00} + \gamma_{01} W_j + \gamma_{10} X_{ij} + \gamma_{11} W_j X_{ij} + u_{0j} + u_{1j} X_{ij} + r_{ij}. \tag{18}$$

The fundamental difference between this combined model and models typical of single-level models is that instead of having independent random errors with constant variance (the r_{ij} term), the error term of equation (18) takes the form $u_{0j} + u_{1j}X_{ij} + r_{ij}$. Distributions are now $r_{ij} \sim N(0, \sigma^2)$ and that $u_{.j} \sim N(0, \tau_{..})$, where τ is the variance-covariance matrix of the $u_{.j}$ terms, whose diagonal elements describe the variance of each u parameter. The τ variance-covariance matrix of the second level models becomes an important set of parameters as it describes between-group variance and determines whether higher-level relationships between variables are needed, significant, or explanatory.

An important part of the analyses presented here is a model termed the 'unconditional model' which is effectively a one-way ANOVA whose factors are the two levels of scale. This model estimates the variance components of the above model with no predictors which describes variation at the stem and patch scales for all seed fates and predictors. This unconditional model can be evaluated and the proportion of variance calculated as ρ, the interclass correlation coefficient: $\rho = \tau_{00}/(\tau_{00} + \sigma^2)$. Percentage of variation explained by predictor variables can then be incorporated into conditional models that seek to explain this initial variance structure. As conditional models are built by subtracting estimates of the residual variance from the unconditional variance and then dividing that by the unconditional variance $((\sigma^2_{UNCON} - \sigma^2_{REGRESS})/\sigma^2_{UNCON})$. This sequence of model-building first and foremost establishes the scale structure of the data and then asks questions according to that structure.

The Bayesian estimation of the full conditional models approximates the joint posterior distribution of the intercepts and slopes of all lower-level groups (the β parameters), the higher-level intercepts and slopes (the γ parameters), and the lower- and higher-level variance components (σ and $\mathbf{T}$ respectively). Seltzer et al. (1996)

solved this joint distribution and derived conditional and marginal distributions in a Gibbs sampler (Gelfand et al., 1990) that approximates the integrations necessary for this solution. This can be done similarly with group level variation using the τ parameters.

Seltzer et al. (1996) developed a Gibbs sampler that generalizes the normal distribution for the β parameters to a Student's t-distirbution. When within-group parameters are estimated (β_j^*), the estimates draw from two sources: the estimate of each group's parameter vector ($\beta_{\mathbf{j}}$) and the overall population parameter estimates (γ). When estimated assuming a normal distribution for the β parameters, groups that have estimated values for the $\beta_{\mathbf{j}}$ parameters far from the population value tend to be pulled inward towards those γ estimates. This is termed 'Bayesian shrinkage,' as the range of group estimates 'shrink' towards the population estimate depending on sample sizes within groups (Gelman et al., 2004; Raudenbush and Bryk, 2002). This shrinkage gives less weight to outliers and can augment low sample sizes in some groups with information derived from the overall population and other groups. In this study, extreme group values may indicate ecological processes at work and not sampling deviations. Although the same principle works in this study, by using a t-distribution with low degrees of freedom (~ 8) (which has longer tails than a normal distribution) reduces the weight given to the population values.

HLM regressions All response variables and predictor variables were included in unconditional models 10. Data were transformed according to column 2 in Table 10 to meet assumptions of normality and pull mass of variable density away from extreme right-skewness, which enables better interpretation of variation (Gelman et al., 2004; Legendre and Legendre, 1997). All posterior estimates of coefficients used the median of the Gibbs sampler output because in some asymmetrical distributions (e.g.,

variance parameters), the posterior median is a better indicator of a point estimate of the parameter ($E(\theta)$) than the mean, and in symmetrical distributions they are effectively identical (Gelman et al., 2004). Diagnostics of interest from the posterior distributions of parameters included medians, upper and lower 95% probability intervals, and percentage of posterior mass above zero for regression coefficients. This last diagnostic is an important way of describing trends in the data, as probability intervals do not have the same meaning in the Bayesian context as in frequentist contexts and the distribution of mass sampling (unlike sample-size based statistics) is a straightforward way to think of the probability that a parameter differs from zero Gelman et al. (2004); Raudenbush and Bryk (2002).

After building the unconditional models, biologically meaningful predictor variables were singularly incorporated into models based on the scale at which variance was evident. Regressions were not built to explain patch-level variation if little was evident. Output diagnostics again returned mean, median, posterior probability intervals, and percentage of mass above and below zero. For regressions with significant *gamma* parameters, the median of variance parameters were used to calculate the interclass correlation coefficient (ρ) (Raudenbush and Bryk, 2002). These models and an unconditional model for viable seed number were then run against patch-level predictors in the same manner. Because there were only eight patches of plants, a hold-out bootstrap was run on all unconditional models of response variables to determine the potential influence of individual patches on regression estimates. Because initial results showed no significant influence of any one patch, the hold-out was not done for predictor variables or higher-level models.

Missing data. The iterative structure of the Gibbs sampler was used to estimate missing data. The process was conceptually similar to that employed by Gelman

et al. (2004). In the Bayesian context, missing data can be treated as unknown components of the HLM model, similar to other model parameters, and estimated by using a prior and the observed data and integrating over the rest of the unknown parameters. In this way the missing data are estimated simultaneously with the hierarchical regression parameters. The protocol for estimating missing data entailed identifying the types of missing data, assigning those data priors, and then building a posterior estimate of those data using the Gibbs sampler.

While seeds were being counted from each inflorescence, any potentially missing data were noted for that inflorescence. The missing data consisted of four 'process' categories: seeds that may have dispersed, had been aborted, had been eaten, and or those seeds that had been aborted because the inflorescence never opened. Aborted seeds were disintegrated, eaten seeds were designated when a casing, frass, or webbing was evident in the inflorescence, and dispersed seeds were indicated when there were absent seeds from mature inflorescences or seeds loosed from the receptacle (the base of the pedicel to which seeds are attached to the inflorescence (Figure 8 f)) The dispersed data are termed 'potenitally' missing because, for example, inflorescences may have seeds loosely attached to the receptacle which could indicate that some seeds were dispersed before collection. This would be noted as dispersed missing data and every iteration of the sampler would determine whether these data were missing with a certain probability built through the prior. The priors for the Gibbs sampler were derived from two functions. The first was the regression equation of total seeds per inflorescence against receptacle width (that is how many seeds would be expected from each inflorescence based on the width of the receptacle) using the regression models in (Table 9). Different regressions were used because these processes reflect different seed sizes (e.g., dispersed seeds tended to be mature and larger than aborted seeds, and therefore represented fewer missing seeds given the same receptacle width).

The regressions produced a posterior estimate of seed numbers distributed normally with a mean designated by the intercept and slope and a variance based on the residuals. The second part of the process took these prior distributions and numerically integrated them over the model using the Gibbs sampler. At each Gibbs step, each receptacle with missing data had an estimated number of seeds drawn from the process distributions. If that number was greater than the observed number of seeds, the difference between observed and estimated was filled using a draw from a multinomial distribution whose proportions of aborted, dispersed, eaten, viable and inviable seeds depended on the observed seeds in the inflorescence. If, for example, fifteen seeds were estimated to be missing due to dispersal, and of twenty-one observed seeds in that inflorescence fourteen were inviable and seven were viable, the fifteen seeds would be drawn from a multinomial distribution with 66% probability of a drawn seed being inviable and 34% chance of the seed being viable. This numerically integrates the missing data over all parameters because although in one run, a particular missing data set will result from the estimation algorithm and then used in the HLM model, the next step a different missing data set will be used. Over the 2000 Gibbs runs, the data will approach a posterior distribution, just as all parameters do.

Results

The Gibbs sampler for all models converged within 200 iterations (the 'burn-in'), and all models were run for 2000 iterations after (Figure 9). Post-hoc tests indicated that autocorrelation was minimal among MCMC runs (Gelman et al., 2004) (Figure 9). Because of skewed distributions for variance components, medians were used for all analyses (Gelman et al., 2004). The three components of output analyzed were the γ parameters, which reflect across-patch associations between variables, β parameters,

which indicate within-patch associations between variables, and the percentage of mass of the posterior distribution above zero.

Missing data were based on regressions shown in Table 9 and the 487 inflorescence heads determined to be complete. These complete heads yielded 17,731 total seeds. Missing data were normally distributed with fairly narrow variances (Figure 10). After missing data estimates, a median value of 21,544 seeds were used in regressions (Figure 10). The relative numbers of seeds determined to be part of the seed fates are in the boxes of Figure 5.

In the unconditional models. variance in both the response and predictor variables was highly skewed towards stem-level variation (Table 10). Aborted and predated seeds varied almost entirely from stem-to-stem (93.0% and 90.7% respectively). Pollination was run including and excluding aborted seeds as it is impossible to determine whether aborted seeds were in fact pollinated or if they were aborted before the florets were receptive. Predated seeds that aren't wholly eaten can be identified as either pollinated or not (EV or DS respectively). 'Early' pollination, which included seeds that may have been pollinated and then aborted, showed greater patch-level variation (21.9%) than 'late' pollination (8.3%). Only the percent of viable seeds and number of viable seeds showed greater than 25% variation at the patch level (26.4% and 29.1% respectively).

Predictor variables that could be measured at the stem-level showed more patch-level structure than seed fate variables (Table 10). Still, only plant height and leaf damage showed greater than 25% variation at the patch level (26.5% and 33.9% respectively). Biomass and leaf number showed slightly less patch-level structure (20.1% and 21.4% respectively). Old nodes varied the least at the patch level of all predictor variables (13.8%). By this estimate, patches showed relatively similar ages while stems differed in age structure.

Bootstraps of the hold-out patches (Figure IV) shows that no individual patch had undue influence on variance estimation for the seed fate variables. Because there was so little patch-level variation to begin with, this is not completely surprising. But even with variables with higher patch-level structure, such as total viable seeds and percent viable seeds, all patches fall equally around the full-model estimates.

Conditional models of stem-level variance were built for all variables. Only total seed number per plant showed significant responses from stem-level variables biomass and leaf number across all patches (γ_{01}) (Table 11, Figure 12. Plant biomass was associated with total seed number positively, explaining 40.7% of stem-level variation in total seed number. The patch-level regression coefficients for this relationship (β_j) showed differences in mass distribution between patches (Figures 13). patches 4 and 8 showed lower regression slopes than the others. This variation is explained by the τ_{11} cell in the variance covariance matrix of the β parameters (the $2x2\mathbf{T}$ matrix). The τ_{11} parameter median was 7.278, with a probability interval of 1.150 and 42.160. Although broad, this shows significant deviation from zero, meaning that the relationship between seed total and biomass differs significantly between patches. The percentage of all seeds that were pollinated and avoided predation (%viable) were predicted by no stem-level predictors (Figure 14). Although not significant across patches, inflorescence heads showed between patch variation (Figure 16). The τ_{11} parameter median for this relationship was 0.017, with a probability interval of 0.002 and 0.121. No patch-level predictors showed an association with the variation in slopes documented for total seeds against plant biomass or percent viable seeds and inflorescence number.

Only three response variables showed enough patch-level variation to warrant the construction of patch-level conditional models: percentage of seeds per plant that were viable, total number of viable seeds per plant, and percentage of seeds per plant

that were pollinated. No model was constructed for Patch-level conditional models for percent viable seeds showed no significant predictors. Direct light predicted 38.5% of patch-level variation in absolute seed number (density mass above zero was 95.1%).

Inflorescence heads per patch and patch species richness did not influence early patch-level variation in percentage of seeds pollinated. Direct transmitted light and soil moisture, however, both explained patch-level variation in the percentage of early pollinated seeds per stem (35.3% and 39.0% respectively) (density mass above zero was 90.9% and 98.3%. These were run in a multiple regression. Both variables maintained a strong positive density, with a probability mass above zero of 85.1% and 95.6% for direct light and soil moisture respectively indicating little collinearity. The patch-level variance of the multiple regression had a median of 0.0017 giving a combined reduction in patch-level variance of 62.0%. A regression was run with an interaction term, but it was not significant.

Of all response variables modeled, the log of absolute viable seed number had the highest patch-level variation (29.13%, Table 10), with a variance component of $\tau = 0.4927$ (Figure 15). Direct light transmittance from the hemispherical photo measurements showed a trend in explaining this variance. 94.3% of the posterior γ_{01} mass was above zero. The posterior median patch-level variance component given direct light $\tau_{00.LIGHT}$ was estimated as 0.276. Therefore 43.96% of the patch-level variance in viable seeds per plant was explained by direct light transmittance. Soil moisture also showed a high positive relationship with overall viable seeds produced per plant from patch to patch (98.4% of the posterior γ_{01} mass was above zero). These predictors were combined in a multiple regression and the posterior masses of light and soil moisture were 94.8% and 97.8% above zero. γ_{01} values were 0.469 and 7.887 for light and soil moisture respectively. The patch-level variance of the multiple

regression had a median of 0.087 giving a combined reduction in patch-level variance of 82.3%.

Discussion

Spatial patterns of seed production can determine ecological and evolutionary dynamics of plant populations (e.g., species presence/absence or abundance (Maron and Kauffman, 2006), source-sink dynamics (Pulliam, 2000, 1988), metapopulation structure (van Groenendael et al., 2000), genetic diversity (Hamrick et al., 1992), genotype fitness (Karkkainen et al., 1999), and Allee affects (Willi and Fischer, 2005)). This paper models floral production, pollination, seed predation, and seed abortion, to test how these processes that ultimately determine final seed set are structured across the natural hierarchical scale of a clonal plant species. The most striking result of this study is that most of the myriad processes that influence the development and dispersal of viable seeds in the studied population combine to influence seed development primarily on the scale of the stem and not the patch. The high proportion of stem-level variation relative to patch-level variation indicates that the production of inflorescence heads, their development, pollination, predation, and abortion are primarily the provenance of ramets and not patches of ramets or genets. This pattern suggests that stem-level hypotheses are most likely responsible for seed fate in the studied population (Table 8).

Total seed production in plants has been shown to be influenced by factors that would likely operate on a patch-scale. Several, as in this study, are mediated through their influence on plant growth, where patch-level variables correlate with larger plant size that in turn leads to increased flower production. Munoz et al. (2005) showed that nitrogen addition to *Chuquiraga oppositifolia* (Asteraceae), an Andean alpine shrub,

significantly increased plant growth, flower size, pollination, and ultimately final seed set compared to controls. Fertilization also increased plant size and floral production of *Ipomopsis aggregata* (Polemoniaceae) Campbell and Halama (1993). Lee and Bazazz (1982) found that in populations of *Cassia fasciculata* (Leguminosae), water addition and removal of competitors increased plant size, which was positively related to fruit production. Another patch-level predictor of seed set can be general maternal plant genotype. Some genets will naturally grow larger and produce more seeds (Campbell, 1997). In a clonal plant, large genotypes will occur in patches, as clonal stems should show similar growth patterns. In this study, nearly 86% of the variation in the production of seeds was found a the scale of the individual stem. This strong stem-scale seed production in *E. chloroepis* indicates that patch-level differences in resources and genetic differences between clonal clusters is not strong. Within patches, however, stem biomass predicted 41% of the stem-level variation in seed production. So although the resource pathway that leads to larger size and in turn greater overall seed production may exist, it is guided at a finer scale. This could indicate fine-grain heterogeneity in soil and light resources (intra-specific competition), or trade-offs within the genet. Although Salzman and Parker (1985) show that rhizome connections averaged stress (an conversely resources) across the genet.

Including viable seeds and predated viable seeds in calculating the incidence of pollination, this study found that 19.1% of all seeds were pollinated. There was a higher degree of patch-level variation (21%) than for any other variable save unpredated viable seed production (which is the product of pollination). *E. chloroepis* is pollinated by generalist insect pollinators (pers. obs.). Light has been documented to be of direct importance to pollination primarily through its influence on insect pollinators (Herrera, 1995). Insect pollinators require temperatures high enough to

support flight metabolism and in a forest understory, direct sunlight is the most common source of micro-habitat temperature changes. In a study of the effect of floral traits on pollination success, Herrera (1995) found no floral trait (e.g., corolla width, disk width, nectar production) predicted pollinator presence, but light and temperature did predict pollinator visits. In this study, light and soil moisture positively related to pollination and combined to explain 62.0% of the patch-level variation, indicating a possible link between temperature and foraging. When modeled alone, light and soil moisture explained 35.3% and 39.0% of the variance respectively. The positive relationship between soil moisture and pollination could be an indirect pathway. Holtsford (1985) showed that *Calochortus leichtlinii* (Liliaceae) produced more seeds per fruit in a water treatment. He posited that the water reduced stress and allowed the plant to mature more seeds. Because this study takes into account seed abortion, which was almost entirely a stem-to-stem process, the relationship between soil moisture and patch-level pollination variation may be more cryptic.

The predation of pre-dispersed seeds has been documented as an important influence on total seed production, with implications for population viability, plant abundance, and flowering phenology (Lee and Bazazz, 1982; Maron and Kauffman, 2006; Albrectsen and Nachman, 2001; Wright and Meagher, 2003). In this study, 19.8% of all seeds were categorized as predated (over 4,250 seeds), and 91% of the variation in seed predation occurred at the stem level. No regressed predictors created significant models (Table 11). This structure of the pattern of seed predation eliminates some patch-level environmental cues as causes of seed predation (similar to the patterns found in pollination). Because seed predation, as with pollination, reflects both the characteristics and influences on the plant as well as the predator (in this case almost always a tephritid fly), the mechanisms behind this pattern can

be multiple and complex. Search behavior of predators, temporal staggering of inflorescence development (some stems have buds earlier than others, which will be differentially exploited depending on the prevalence of predators), and the finer temporal environment that might control both (periods of rain may dampen predator oviposition). Although it is impossible in a broad study to precisely define the causes of observed seed predation, the fact that it exists on a stem-level scale eliminates maternal genotype, patch-level environmental effects, and patch-level cues (such as inflorescence density).

Seed abortion, the withholding of resources to a formed bud, inflorescence head, or ovule, showed almost no patch-level variation. An estimated 12.5% of all 21,544 seeds analyzed were aborted. 94% of the variation in the percentage of all seeds on a plant that were aborted varied from stem to stem and not from patch to patch. Although seed abortion can be influenced by herbivore damage (Krupnick et al., 1999), nutrient levels (Volis et al., 2004), and other forms of stress (Volis et al., 2004; Sun et al., 2004), this system shows that variation in seed abortion does not differ from patch to patch even though there are clear distinctions between patches in variables that could conceivably influence seed abortion patterns (Figure IV). None of the regressed stem-level independent variables predicted the variation in seed abortion (Table 11). These results point to a micro-allocation strategy by the stems, whereby individual seeds and buds are aborted due to differential pollination, production, and resource acquisition (Ehrlen, 1991; Melser and Klinkhamer, 2001). These three factors can easily confuse a simple census and would have to be experimentally tested to better determine the mechanisms responsible for seed abortions. Another potential cause of stem-level abortion is seed genotype. Wiens et al. (1987) show that in predominantly outcrossing hermaphroditic plants early genetic load from self-crossed pollen can lead to a number of genetic anomalies that lead to seed abortion. Thus, although stress

and resource variability is often cited as the reason plants produce far more fruits than seeds (Burd, 1998; Lee and Bazazz, 1982; Ehrlen, 1991), poor offspring genotype may also influence this ratio.

The processes that influence sexual reproduction in plants can have important influences on the ecological and evolutionary dynamics of a species. This study shows that when considering the many components of seed fitness, scale is essential in developing models that reflect true variation in important variables. In this study, *E. chlorolepis* seed fate dynamics were dictated almost entirely at the stem-scale. This suggests that processes such as within-genet resource allocation and poor offspring genotype are likely responsible for how many viable seeds are produced and dispersed by reproductive stems.

Bibliography

Albrectsen, B. and G. Nachman. 2001. Female-biased density-dependent dispersal of a tephritid fly in a fragmented habitat and its implications for population regulation. OIKOS **94**:263–272.

Albrectsen, B. R. 2000. Flowering phenology and seed predation by a tephritid fly: Escape of seeds in time and space. Ecoscience **7**:433.

Balestri, E. and F. Cinelli. 2003. Sexual reproductive success in *Posidonia oceanica*. Aquatic Botany **75**:21–32.

Berjano, R., C. De Vega, M. Arista, P. L. Ortiz, and S. Talavera. 2006. A multi-year study of factors affecting fruit production in aristolochia paucinervis (aristolochiaceae). American Journal of Botany **93**:599–606.

Brody, A. K. 1992. Oviposition choices by a predispersal seed predator (*Hylemya* sp.). Oecologia **91**:56–62.

Burd, M. 1998. "Excess" flower production and selective fruit abortion: a model of potential benefits. Ecology **79**:2123–2132.

Campbell, D. R. 1997. Genetic and environmental variation in life-history traits of a monocarpic perennial: a decade-long field experiment. Evolution **51**:373–382.

Campbell, D. R. and K. J. Halama. 1993. Resource and pollen limitations to lifetime seed production in a natural plant population. Ecology **74**:1043–1051.

Chung, M. G. and M. Y. Chung. 1999. Spatial genetic structure of clonal and sexual reproduction in a population of *Abeliophyllum distichum* (Oleaceae), an endangered monotypic genus. Genes and Genetic Systems **74**:9–14.

Clark, J. S. 2005. Why environmental scientists are becoming Bayesians. Ecology Letters **8**:2–14.

Clark, J. S. and A. E. Gelfand. 2006. Hierarchical modeling for the environmental sciences: statistical methods and applications. Oxford University Press.

Dupre, C. and J. Ehrlen. 2002. Habitat configuration, species traits and plant distributions. Journal of Ecology **90**:796–805.

Ehlers, B. K. and J. M. Olesen. 2004. Flower production in relation to individual plant age and leaf production among different patches of *Corydalis intermedia*. Plant Ecology **174**:71–78.

Ehrlen, J. 1991. Why do plants produce surplus flowers? A reserve-ovary model. American Naturalist **138**:918–933.

Ehrlen, J. 1996. Spatiotemporal variation in predispersal seed predation intensity. Oecologia **108**:708–713.

Ehrlen, J. 1997. Risk of grazing and flower number in a perennial plant. Oikos **80**:428–434.

Ehrlen, J. 2002. Assessing the lifetime consequences of plant-animal interactions for the perennial herb *Lathyrus vernus* (Fabaceae). Perspectives in Plant Ecology Evolution and Systematics **5**:145–163.

Frazer, G., C. Canham, and K. Lertzman, 1999. Gap light analyzer (GLA).

Gabrielsen, T. M. and C. Brochmann. 1998. Sex after all: high levels of diversity detected in the arctic clonal plant saxifraga cernua using rapd markers. Molecular Ecology **7**:1701–1708.

Garcia-Robledo, C., G. Kattan, C. Murcia, and P. Quintero-Marin. 2005. Equal and opposite effects of floral offer and spatial distribution on fruit production and predispersal seed predation in *Xanthosoma daguense* (araceae). Biotropica **37**:373–380.

Gehring, J. L. and L. F. Delph. 2006. Effects of reduced source-sink ratio on the cost of reproduction in females of *Silene latifolia*. International Journal of Plant Sciences **167**:843–851.

Gelfand, A. E., A. E., S. E. hills, A. Racinepoon, and A. F. M. Smith. 1990. Illustration of Bayesian-inference in normal data models using Gibbs sampling. Journal of the American Statistical Association **85**:972–985.

Gelman, A., J. B. Carlin, and H. S. S. B. Rubin. 2004. Bayesian Data Analysis, 2nd ed. Chapman and Hall CRC, New York.

Guretzky, J. A. and S. M. Louda. 1997. Evidence for natural biological control: Insects decrease survival and growth of a native thistle. Ecological Applications **7**:1330.

Hamrick, J. L., M. J. W. Godt, and Sherman-Broyles. 1992. Factors influencing levels of genetic diversity in woody plant species. New Forests **6**:95–124.

Herrera, C. M., M. Medrano, P. J. Rey, A. M. Sanchez-Lafuente, M. B. Garcia, J. Guitian, and A. J. Manzaneda. 2002. Interaction of pollinators and herbivores on plant fitness suggests a pathway for correlated evolution of mutualism- and antagonism-related traits. Proceedings of the National Academy of Sciences of the United States of America **99**:16823.

Herrera, C. M. a. 1995. Microclimate and invdividual variation in pollinators: flowering plants are more than their flowers. Ecology **76**:1516–1524.

Hersch, E. I. 2006. Foliar damage to parental plants interacts to influence mating success of ipomoea purpurea. Ecology **87**:2026–2036.

Holtsford, T. P. 1985. Nonfruiting hermaphroditic flowers of *Calochortus leichtlinii* (Liliaceae):potential reproductive functions. American Journal of Botany **72**:1687.

Horvitz, C. C. and D. W. Schemske. 1995. Spatiotemporal variation in demographic transitions of a tropical understory herb - projection matrix analysis. Ecological Monographs **65**:155–192.

Hurlbert, S. H. 1984. Pseudoreplication and the design of ecological field experiments. Ecological Monographs **54**:187–211.

Kanno, H. and K. Seiwa. 2004. Sexual vs. vegetative reproduction in relation to forest dynamics in the understorey shrub, *Hydrangea paniculata* (Saxifragaceae). Plant Ecology **170**:43–53.

Karkkainen, K., O. Savolainen, and V. Koski. 1999. Why do plants abort so many developing seeds: bad offspring or bad maternal genotypes? Evolutionary Ecology **13**:305–317.

Krupnick, G. A., A. E. Weis, and D. R. Campbell. 1999. The consequences of floral herbivory for pollinator service to *Isomeris arborea*. Ecology Washington D C **80**:125–134.

Lee, T. D. and F. A. Bazazz. 1982. Regulation of fruit and seed production in an annual legume, *Cassia fasciculata*. Ecology **63**:1363–1373.

Legendre, P. and L. Legendre. 1997. Numerical Ecology. Elsevier Science, Amsterdam.

Lehtila, K. and S. Y. Strauss. 1997. Leaf damage by herbivores affects attractiveness to pollinators in wild radish, *Raphanus raphanistrum*. Oecologia **111**:396–403.

Louda, S. M. 1982. Distribution ecology: variation in plant recruitment over a gradient in relation to insect seed predation. Ecological Monographs **52**:25–41.

Louda, S. M. and M. A. Potvin. 1995. Effect of inflorescence-feeding insects on the demography and lifetime fitness of a native plant. Ecology **76**:229–245.

Maron, J. L. 1998. Insect herbivory above- and belowground: Individual and joint effects on plant fitness. Ecology **79**:1281.

Maron, J. L. and M. J. Kauffman. 2006. Habitat-specific impacts of multiple consumers on plant population dynamics. Ecology **87**:113–124.

Melser, C. and P. G. L. Klinkhamer. 2001. Selective seed abortion increases offspring survival in *Cynoglossum officinale* (Boraginaceae). American Journal of Botany **88**:1033–1040.

Mena-Ali, J. I. and O. J. Rocha. 2005. Selective seed abortion affects the performance of the offspring in *Bauhinia ungulata*. Annals of Botany **95**:1017–1023.

Metcalf, J. C., K. E. Rose, and M. Rees. 2003. Evolutionary demography of monocarpic perennials. Trends in Ecology and Evolution **18**:471–480.

Mothershead, K. and R. J. Marquis. 2000. Fitness impacts of herbivory through indirect effects on plant-pollinator interactions in *Oenothera macrocarpa*. Ecology **81**:30–40.

Munoz, A., C. Celedon-Neghme, L. A. Cavieres, and M. T. K. Arroyo. 2005. Bottom-up effects of nutrient availability on flower production, pollinator visitation, and seed output in a high-andean shrub. Oecologia **143**:126–135.

Niesenbaum, R. A. 1996. Linking herbivory and pollination: Defoliation and selective fruit abortion in *Lindera benzoin*. Ecology **77**:2324–2331.

Nishitani, S., T. Takada, and N. Kachi. 1999. Optimal resource allocation to seeds and vegetative propagules under density dependent regulation in *Syneilesis plamata* (Compositae). Plant Ecology **141**:179–189.

Pilson, D. 2000. Herbivory and natural selection on flowering phenology in wild sunflower, *Helianthus annuus*. Oecologia **122**:72–82.

Primack, R. and K. S. Kang. 1989. Measuring fitness and natural selection in wild plant populations. Annual Review of Ecology and Systematics **20**:367–396.

Proctor, M., P. Yeo, and A. Lack. 1996. The natrual history of pollination. Timber Press, Portland, Oregon.

Pulliam, H. 1988. Sources, sinks, and population regulation. American Naturalist **132**:652–661.

Pulliam, H. R. 2000. On the relationship between niche and distribution. Ecology Letters **3**:349–361.

Raudenbush, S. W. and A. S. Bryk. 2002. Hierarchical linear models: applications and data analysis methods. Advanced quantitative techniques in the social sciences, Sage Publications, Thousand Oaks, CA.

Rautiainen, P., K. Koivula, and M. Hyvarinen. 2004. The effect of within-genet and between-genet competition on sexual reproduction and vegetative spread in *Potentilla anserina* ssp. *egedii*. Journal of Ecology **92**:505–511.

Salzman, A. G. and M. A. Parker. 1985. Neighbors ameliorate local salinity stress for a rhizomatous plant in a heterogeneous environment. Oecologia **65**:273–277.

Seltzer, M. H., W. H. Wong, and A. S. Bryk. 1996. Bayesian analysis in applications of hierarrchical models: issues and methods. Journal of Educational and Behavioral Statistics .

Sheppard, A. W., J. M. Cullen, and J. P. Aeschlimann. 1994. Predispersal seed predation on *Carduus-Nutans* (Asteraceae) in southern europe. Acta Oecologica-International Journal of Ecology **15**:529–541.

Slatkin, M. 1985. Gene flow in natural populations. Annual Review of Ecology and Systematics **16**:393–430.

Stephenson, A. G. 1981. Flower and fruit abortion - proximate causes and ultimate functions. Annual Review of Ecology and Systematics **12**:253–279.

Stocklin, J. and E. Winkler. 2004. Optimum reproduction and dispersal strategies of a clonal plant in a metapopulation: a simulation study with *Hieracium pilosella*. Evolutionary Ecology **18**:563–584.

Strauss, S. Y. 1997. Floral characters link herbivores, pollinators, and plant fitness. Ecology **78**:1640–1645.

Sun, K., K. Hunt, and B. A. Hauser. 2004. Ovule abortion in *Arabidopsis* triggered by stress. Plant Physiology **135**:2358–2367.

van Groenendael, J., J. Ehrlen, and B. M. Svensson. 2000. Dispersal and persistence: Population processes and community dynamics. Folia Geobotanica **35**:107.

Verburg, R. W. and H. J. During. 1998. Vegetative propagation and sexual reproduction in the woodland understorey pseudo-annual *Circaea lutetiana* L. Plant Ecology **134**:211–224.

Volis, S., K. Verhoeven, S. Mendlinger, and D. Ward. 2004. Phenotypic selection and regulation of reproduction in different environments in wild barley. Journal of Evolutionary Biology **17**:1121–1131.

Washitani, I., Y. Okayama, K. Sato, H. Takahashi, and T. Ohgushi. 1996. Spatial variation in female fertility related to interactions with flower consumers and pathogens in a forest metapopulation of *Primula sieboldii*. Researches on Population Ecology **38**:249–256.

Wiens, J. J., C. L. Calvin, C. A. Wilson, C. I. Davern, D. Frank, and S. R. Seavey. 1987. Reproductive success, spontaneous embryo abortion and genetic load in flowering plants. Oecologia **71**:501–509.

Willi, Y. and M. Fischer. 2005. Genetic rescue in interconnected populations of small and large size of the self-incompatible *Ranunculus reptans*. Heredity **95**:437–443.

Wise, M. J. and J. J. Cummins. 2006. Strategies of solanum carolinense for regulating maternal investment in response to foliar and floral herbivory. Journal of Ecology **94**:629–636.

Wright, J. W. and T. R. Meagher. 2003. Pollination and seed predation drive flowering phenology in *Silene latifolia* (Caryophyllaceae). Ecology **84**:2062–2073.

Young, L. W., R. W. Wilen, and P. C. Bonham-Smith. 2004. High temperature stress of brassica napus during flowering reduces micro- and megagametophyte fertility, induces fruit abortion, and disrupts seed production. Journal of Experimental Botany **55**:485–495.

Appendix: Tables and Figures

Table 8: Hypotheses for the patterns of seed fate at the stem scale and patch scale.

Seed fate	Scale	Hypothesed mechanisms
Floral production	Stem	Biomass, allocation within genet, insect damage or pathogens.
	Patch	Maternal genotype, resources, environmental stress, broader insect damage or pathogens.
Pollination	Stem	Ideosycratic response of pollinators to cues, staggared maturity of inflorescences, staggered temporal effects over the course of seed development.
	Patch	Environmental cues such as light, micropatch temperature, inflorescence density.
Seed abortion	Stem	Re-allocation of resources within a stem to bolster growth of already pollinated inflorescences, facilitating male flowers, response to stem-level damage or pathogens, poor seed genotype.
	Patch	Environemental stress reducing allocation of resources to all inflorescences on all patch stems, poor maternal genotype.
Seed predation	Stem	Idiosyncratic or within-stem scale cues for seed predators (number of pollinated seeds, nectar production), staggered temporal effects over the course of seed development.
	Patch	Environmental covariates fostering increased seed predation, inflorescence density.

Table 9: Missing Data Estimation Distributions. were derived from regressions on observed complete inflorescence heads that met the criteria denoted. TOT is the total number of seeds assigned to an inflorescence head with receptacle width 'RW.'

Process	Criteria	Distribution
Dispersed seeds	$DS < 5; AS < 5$	$TOT \sim N(22.52 + 5.3 * RW, 4.13^2)$
Eaten seeds	$DS < 15; AS < 5$	$TOT \sim N(16.03 + 7.38 * RW, 4.17^2)$
Aborted seeds	$DS > 15$	$TOT \sim N(11.05 + 10.75 * RW, 4.95^2)$
Aborted heads	$AS > 15$	$TOT \sim N(6.71 + 13.48 * RW, 4.72^2)$

Table 10: Posterior medians of unconditional models for response and predictor variables.

	Variable	Transformation	σ^2	τ_{00}	ρ Stem Var	ρ patch Var
Response Variables	Total seeds	$log(TOTAL)$	0.5486	0.091	0.8578	0.1422
	% pollinated	$\sqrt{\frac{VS+EV+1}{TOTAL+1}}$	0.0163	0.0046	0.7805	0.2195
	% pollinated (late)	$\sqrt{\frac{VS+EV+1}{TOTAL-AS+1}}$	0.0238	0.0021	0.9172	0.0828
	% aborted	$log(\frac{AS+1}{TOTAL+1})$	2.5869	0.1935	0.9304	0.0696
	% predated	$log(\frac{DS2+EV+1}{TOTAL-AS+1})$	2.7683	0.2831	0.9072	0.0928
	% viable	$\sqrt{VS/TOTAL}$	0.0232	0.0084	0.7356	0.2644
	Total viable	$log(VS+1)$	1.1986	0.4927	0.7087	0.2913
Predictor Variables	Height	$HEIGHT$	75.011	27.040	0.735	0.265
	Total inflor.	$log(TOTinflor)$	0.4531	0.1106	0.8038	0.1962
	Leaf damage	$\sqrt{HERB\_DAM}$	0.022	0.0113	0.6611	0.3389
	Biomass	BIOMASS	0.0277	0.0069	0.7992	0.2008
	Leaf no.	$log(LEAF\_NO+1)$	0.1105	0.03	0.7862	0.2138
	Old Nodes	ON	70.7346	0.118	0.8616	0.1384

Table 11: First-level models.

Predictor	γ_{01} CI	Mass > 0	Var. explained
Total seeds			
Biomass	$-0.021, 4.989$	0.973	0.407
Leaf number	$0.229, 2.585$	0.986	0.419
Herbivore damage	$2.140, 1.465$	0.270	–
Old nodes	$-0.242, 0.652$	0.848	–
% aborted seeds			
Herbivore damage	$-4.060, 4.883$	0.538	–
Old nodes	$-0.619, 0.894$	0.614	–
Leaf number	$-2.236, 2.327$	0.543	–
% predated seeds			
Height	$-0.059, 0.092$	0.675	–
Herbivore damage	$-3.041, 4.650$	0.696	–
Biomass	$-3.277, 4.197$	0.657	–
Inflor. number	$-0.640, 1.251$	0.744	–
% pollinated seeds (early)			
Height	$-0.010, 0.008$	0.346	–
Inflor number	$-0.083, 0.100$	0.620	–
Biomass	$-0.508, 0.293$	0.364	–
% pollinated seeds (late)			
Height	$-0.008, 0.008$	0.4585	–
Inflor. number	$-0.140, 0.080$	0.251	–
Biomass	$-0.457, 0.418$	0.450	–
% viable seeds			
Biomass	$-0.465, 0.423$	0.535	–
Inflor. number	$-0.115, 0.152$	0.622	–
Leaf number	$-0.238, 0.200$	0.370	–
Old nodes	$-0.094, 0.052$	0.2225	–

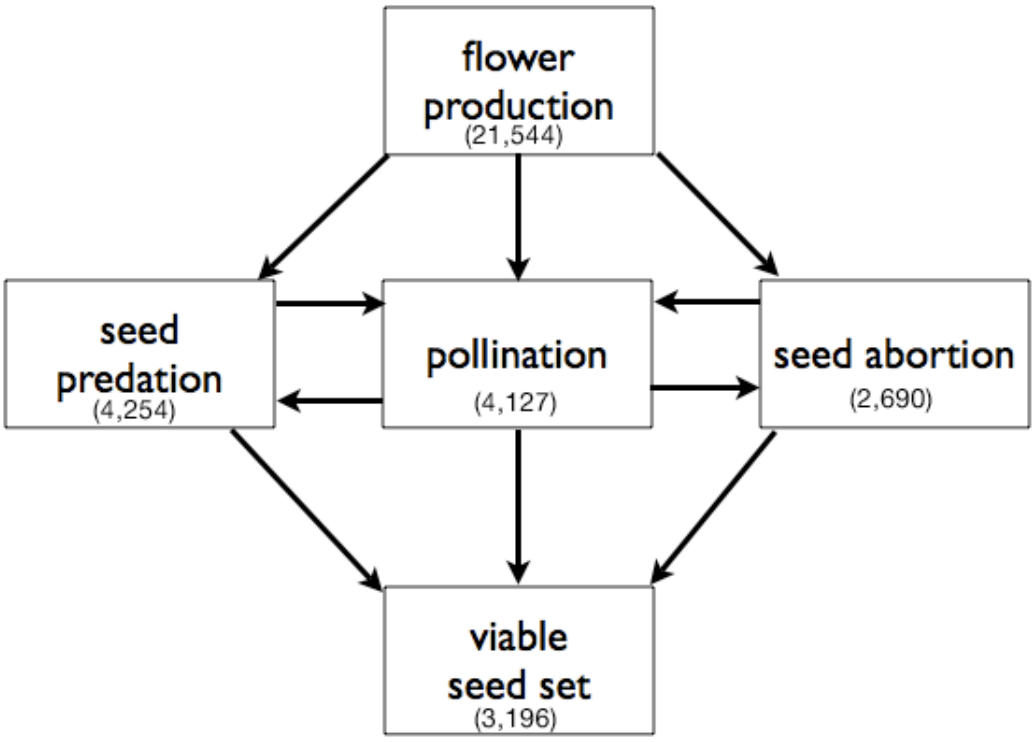


Figure 5: Diagram of the temporal process of seed production to dispersal with the observed variables in the boxes. Note that viable seeds includes eaten viable seeds and eaten seeds includes those that are viable. Pollinated includes both viable and eaten viable.

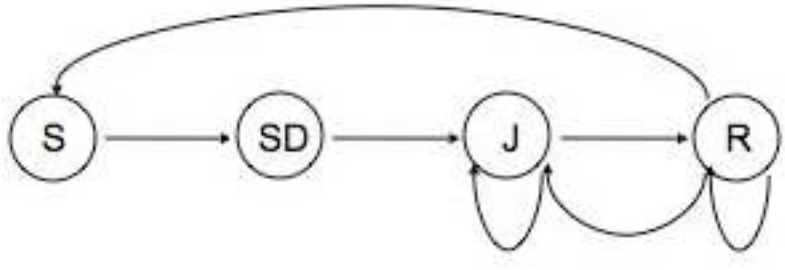


Figure 6: Stage diagram of *E. chlorolepis*. Stages are in circles (S = seed, SD = seedling, J = juvenile, and R = reproductive), and transitions, reproduction, and survival are shown as arrows.

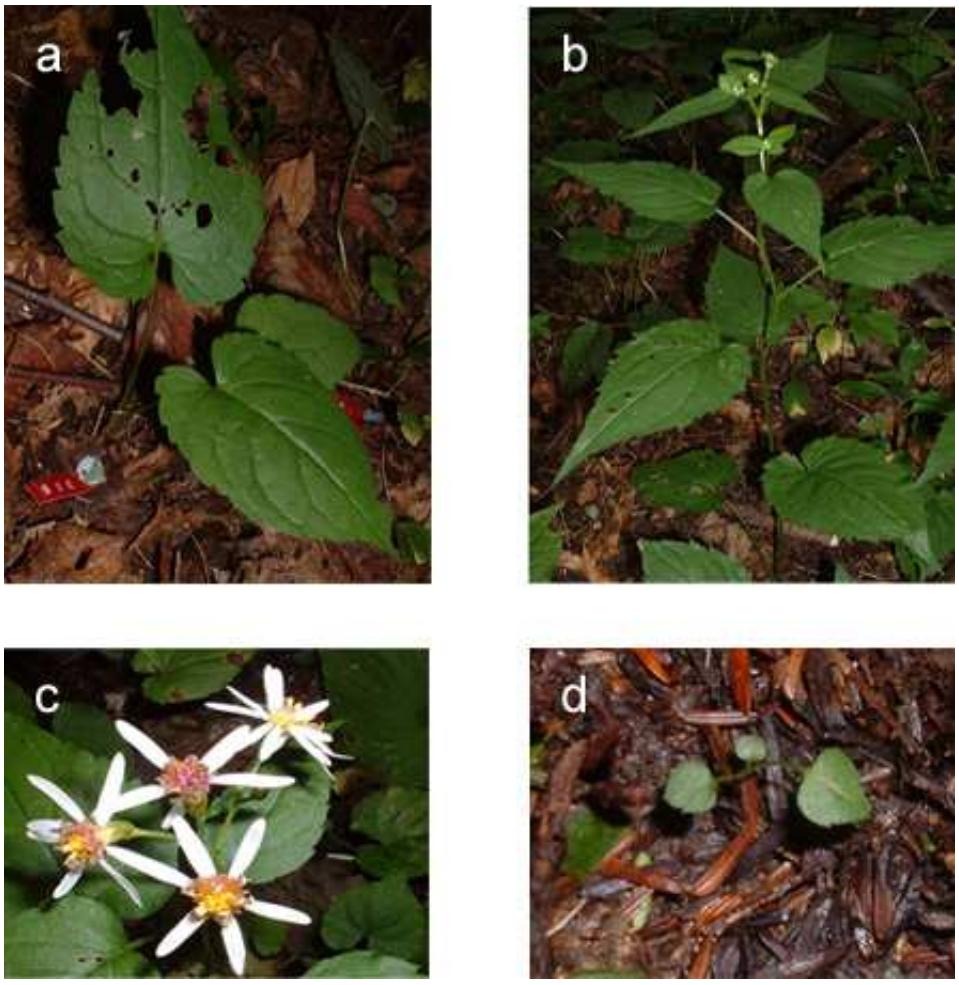


Figure 7: Photographs of *E. chlorolepis*. The juvenile stage shows leaves in a rosette form (a), while the reproductive form has inter-nodal stems (b). Inflorescences show disc and ray flowers (c). Disc flowers show various stages (colors) of reproductive receptivity. A seedling is shown with one of two cotyledons (d).

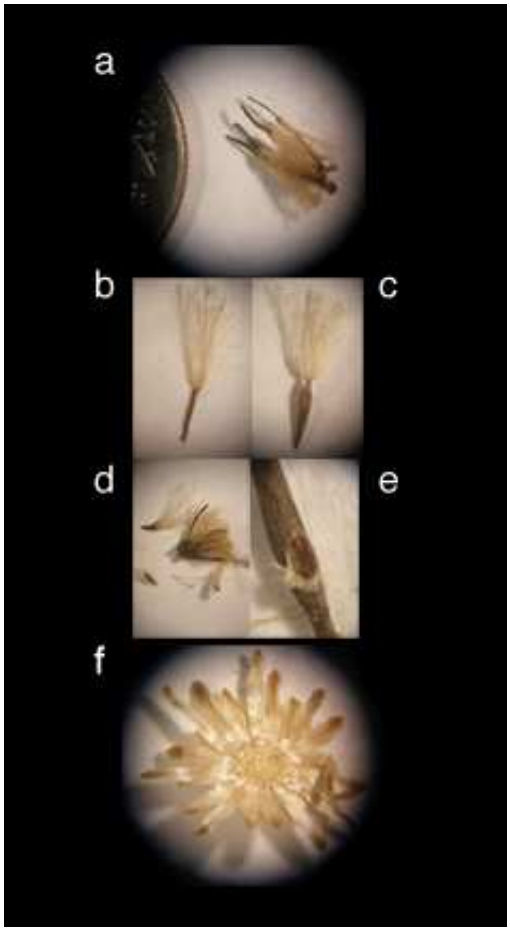


Figure 8: Photographs of *E. chlorolepis* seeds. Size of unpollinated seeds beside a U.S. dime (a), an unpollinated (inviable) seed, (b) a viable seed (c), aborted seeds (d), an eaten viable seed (e), and the receptacle (f).

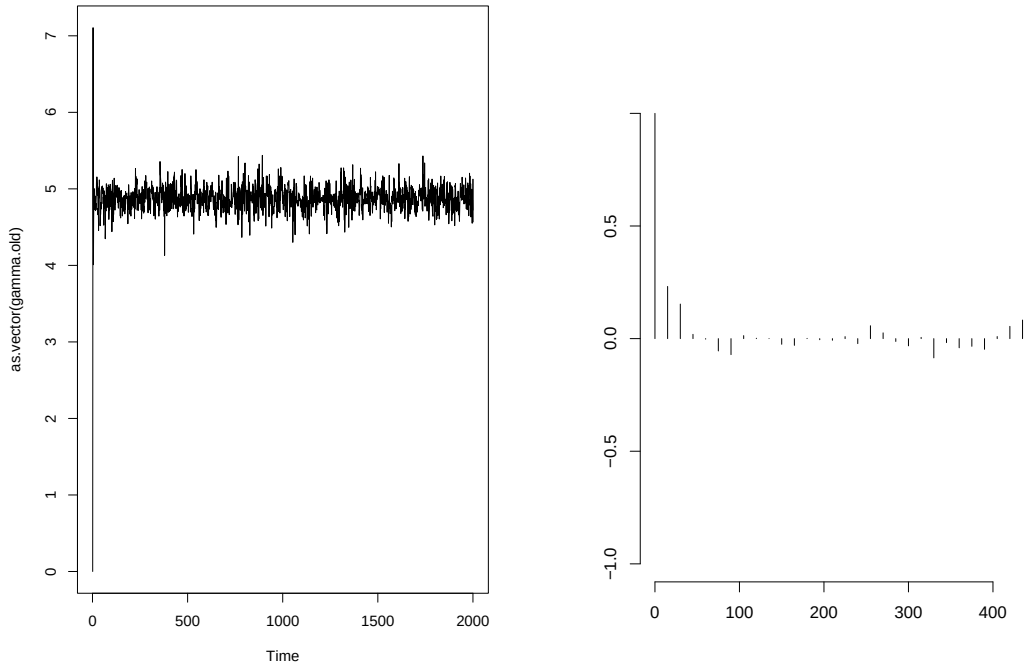


Figure 9: The γ parameter of the unconditional TOTseed model shown converging quickly on its posterior distribution.**a)** Correlation versus lag time shows that beyond immediate time-steps, there is no autocorrelation in the Gibbs sampler.

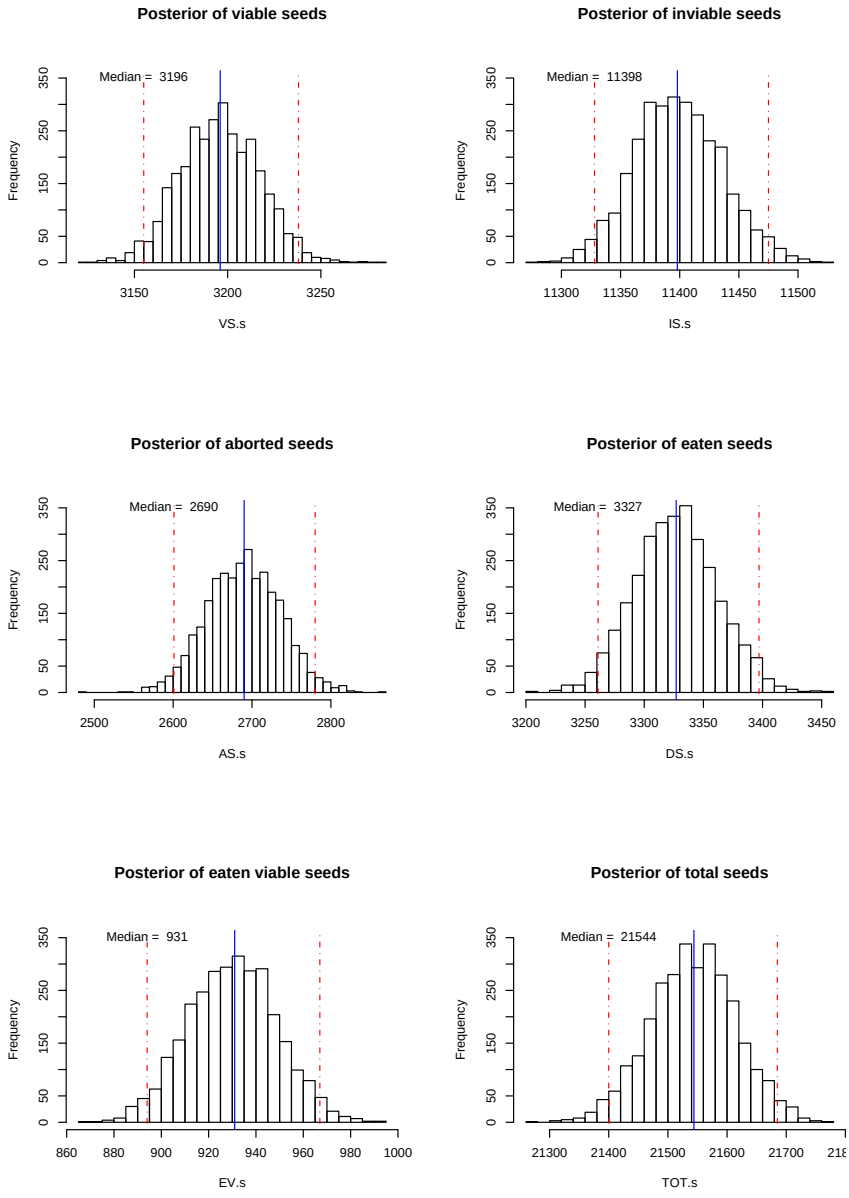


Figure 10: Histograms of posterior seed numbers by fate. Red lines indicate 2.5 and 95% quantiles. The blue line indicates the median. Note that in this figure, the designations are as in Table 10 where 'viable seeds' does not include 'eaten viables', etc.

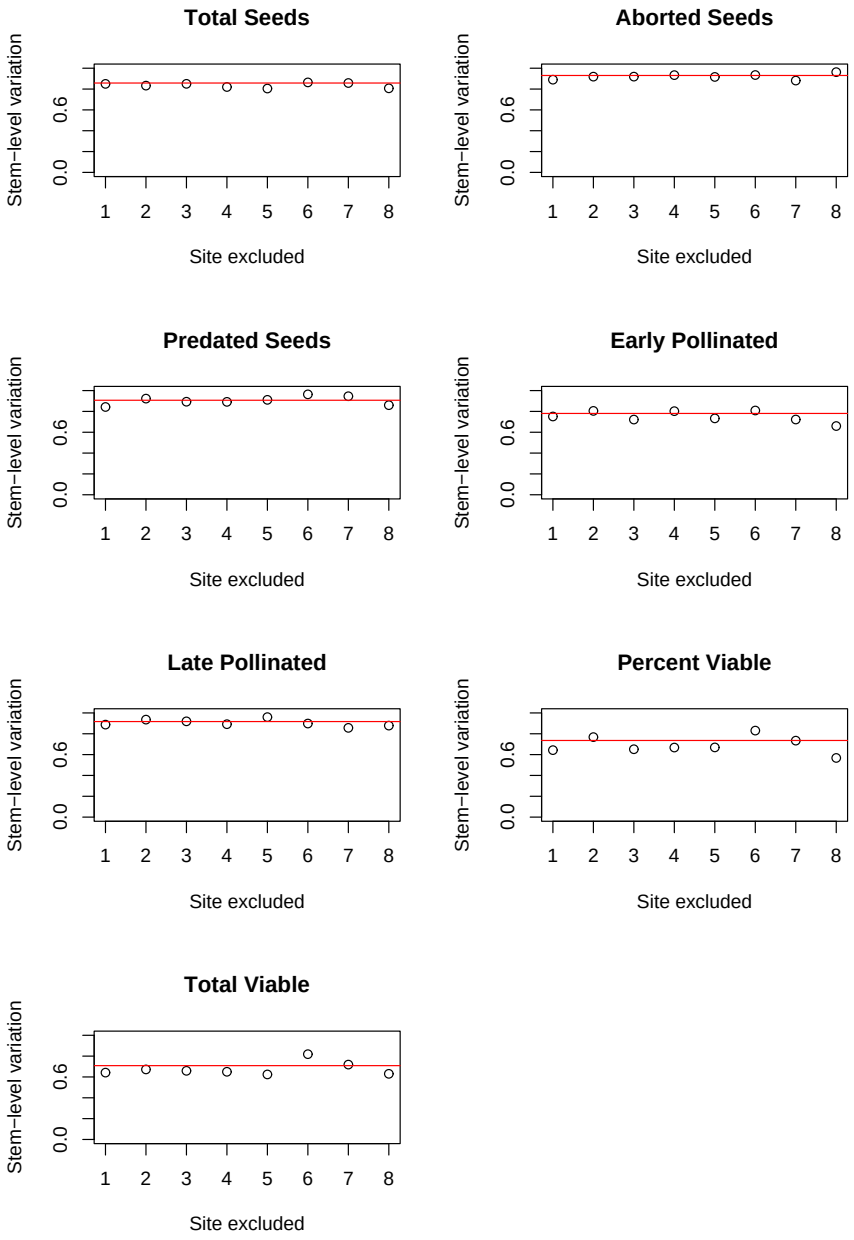


Figure 11: Bootstrap of seed fates holding out patches. Red line shows value of variation with all patches included.

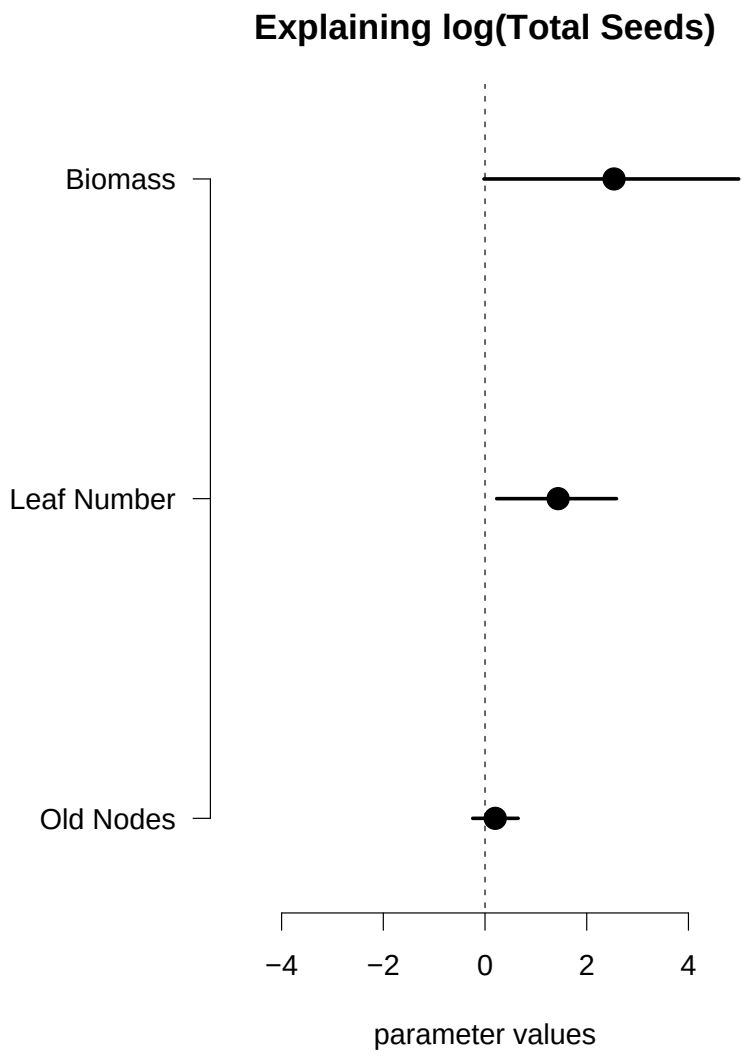


Figure 12: The log of total seeds was regressed against first-level predictor variables.

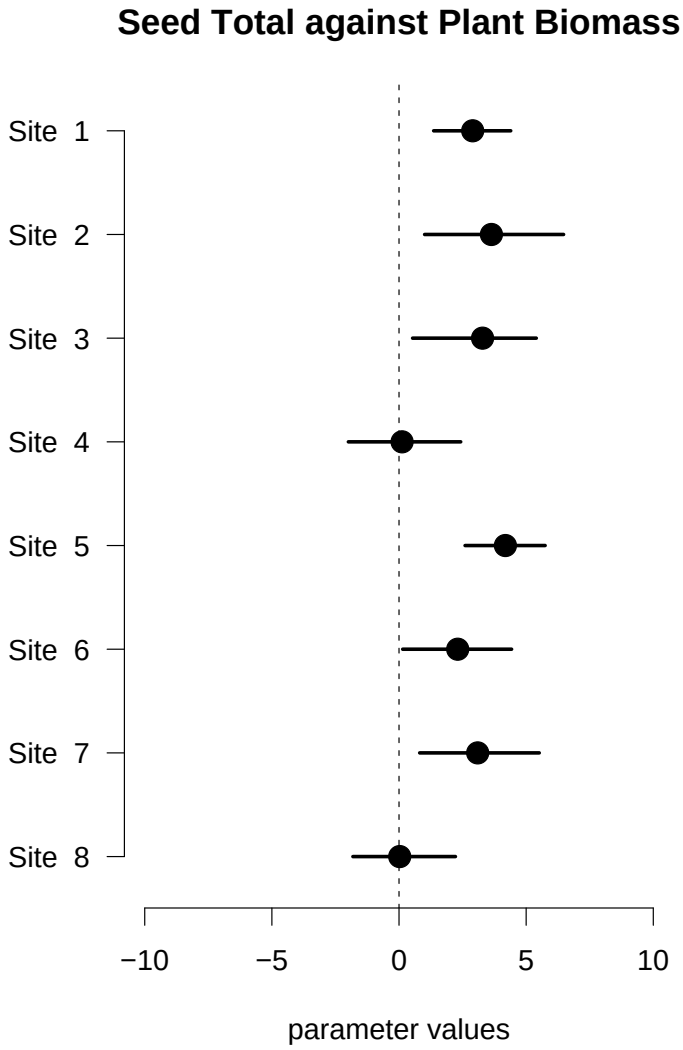


Figure 13: The patch-to-patch differences in the relationship between biomass and total seeds.

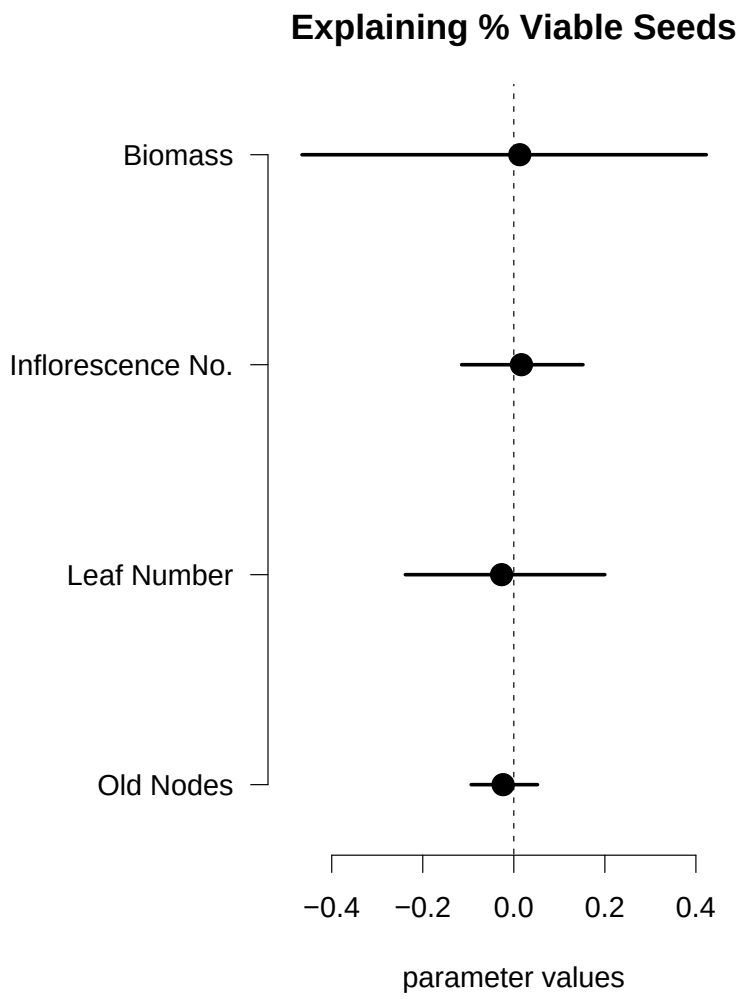


Figure 14: The percent of viable seeds produced by a plant was regressed against first-level predictor variables.

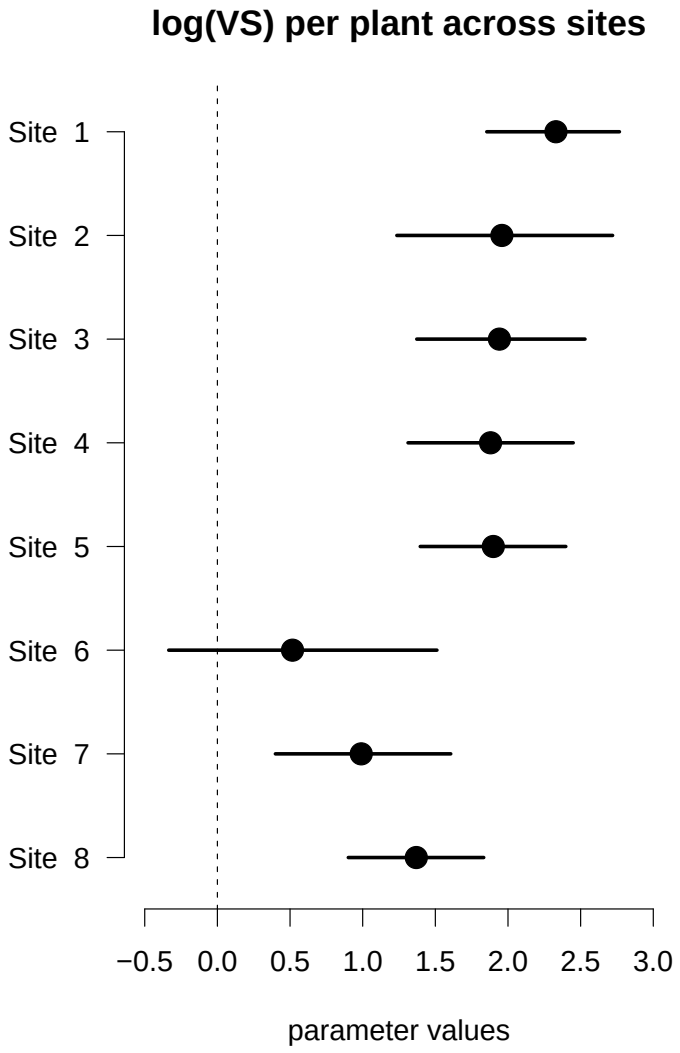


Figure 15: The patch-to-patch differences in the log of viable seeds per plant.

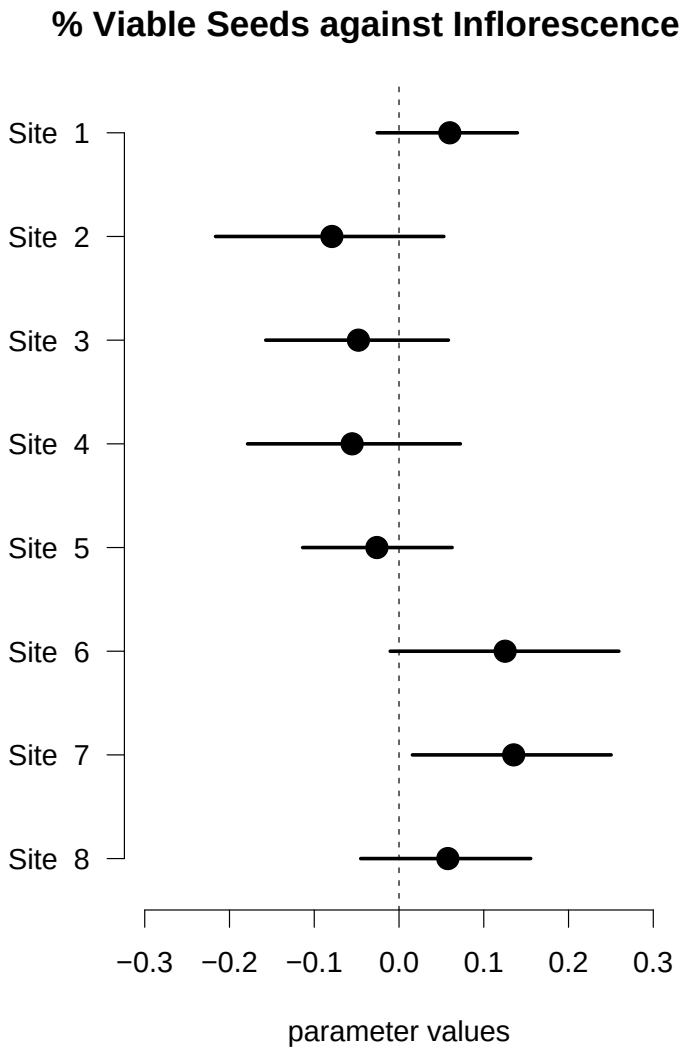


Figure 16: The patch-to-patch differences in the relationship between inflorescence number and absolute number of viable seeds.

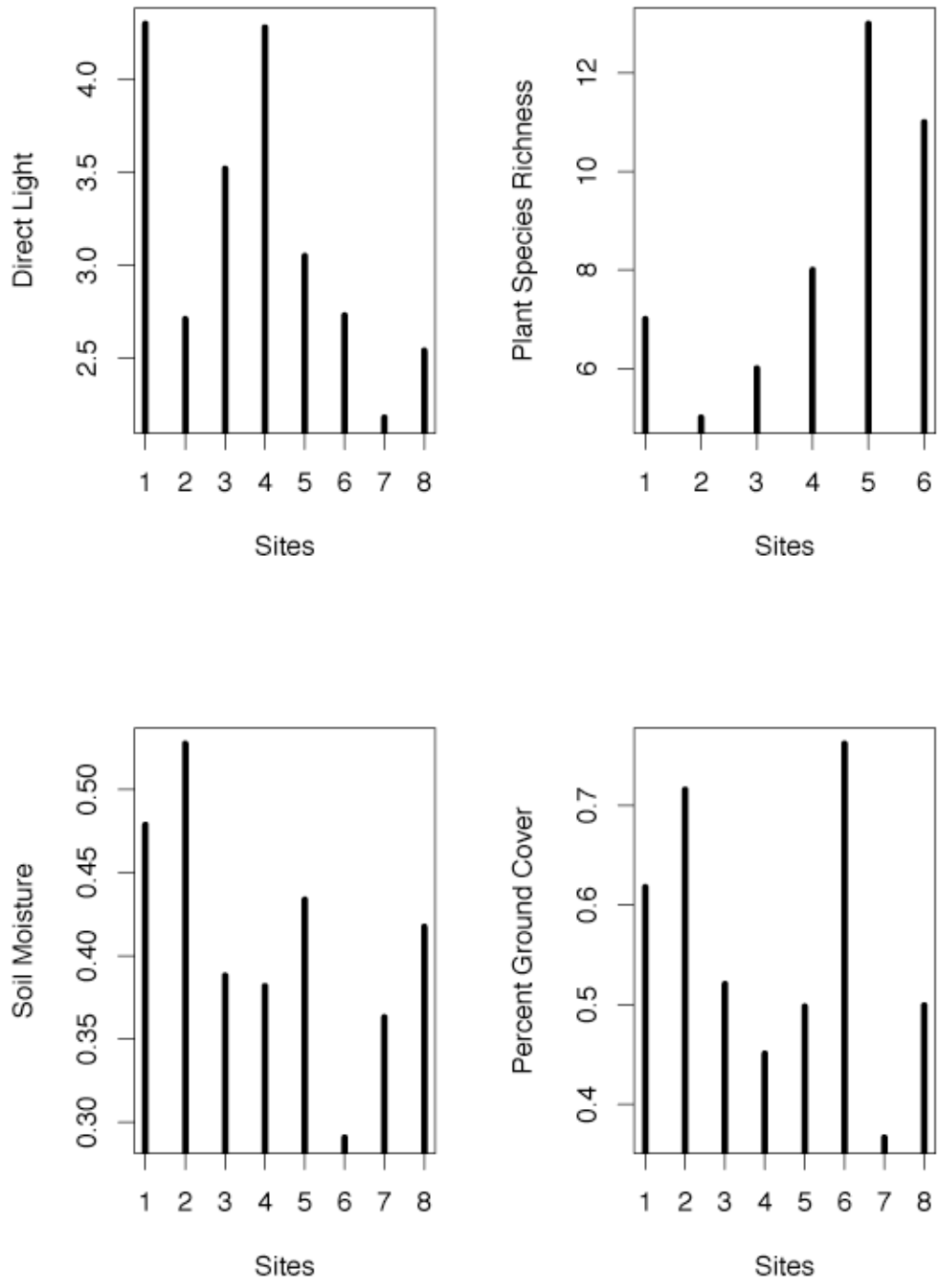


Figure 17: The inter-patch differences for predictor variables.

Part V

Bottom-up effects of a canopy invader

Introduction

The hemlock wooly adelgid (*Adelges tsugae*) is transforming forest ecosystems in the eastern United States by killing up to 99% of the eastern hemlock trees (*Tsuga canadensis*) it infests (Orwig and Foster, 1998). Native to Asia, this insect was most likely introduced to North America and was first observed in British Columbia in 1922 and two years later in Oregon (Annand, 1924). A subsequent introduction occurred in Eastern North America in 1951 (Stoetzel, 2002). By 2005, it had spread to the southern Appalachians where *T. canadensis* is an important component of old growth cove forests, but had not yet defoliated trees in its new range.

The influence of an invading species on the ecological systems it colonizes is rarely limited to the direct effects it exerts on species through interactions such as competition and predation. Cascading or indirect effects exist by default in dynamical systems such as forests, especially when a regime can be shifted by the input of one or more invasive pest species (Parker et al., 1999; Simberloff and Von Holle, 1999; Ghazoul, 2004). The effects of *T. canadensis* mortality on forest understory plants is an indirect effect of the invading *A. tsugae* because defoliation immediately alters the understory light environment and can ultimately change soil pH, nutrient levels, and moisture availability (Kizlinski et al., 2002; Orwig, 2002). Clearly the response of understory plants species to such change is difficult to predict, as are any longer-term changes in species composition and population and community dynamics. Nevertheless, the influence of *T. canadensis* on understory herbs depends largely on how tightly species in the herb community adhere to specific environmental niches in the understory that depend on the canopy trees. This invader is a candidate for such indirect effects. Within the mixed-deciduous forests of the eastern United States, *T. canadensis* is

one of the few evergreen species in moist cove forests in which hardwoods predominate. The soil profile and light regime beneath *T. canadensis* canopies are distinct from deciduous trees that share habitat with *T. canadensis*, and *T. canadensis* lies at the end of the continuum of low nutrient tolerance and light response functions compared to deciduous trees (Bigelow and Canham, 2002). High *T. canadensis* mortality has been shown to drastically change the forest ecosystem (Jenkins et al., 1999).

Even though *T. canadensis* shows a distinct contribution to the forest ecosystem, the impact of its loss on understory plant populations is not simple. The determinants of population growth can be uncertain and high dimensional in natural systems (Clark, 2003). It is not straightforward to map a species to the areas it is found. Soil quality (e.g., texture, cation exchange capacity, organic matter content, pH, and nutrients), soil moisture, climate variables, and light regimes constitute a few of the many potential environmental variables that can dictate where a specific plant species is found. Further, mechanisms such as competition or dispersal limitation can exclude individuals from habitat that would otherwise be considered an appropriate 'niche' (Hutchinson, 1959; Pulliam, 2000). Conversely, when taking into account demographic functions such as reproduction and growth, an observed population may not indicate that the habitat is its appropriate niche, as the population may not be growing or viable in that location (a population 'sink') (Pulliam, 1988). To find influences on the spatial distribution of a plant species, environmental variables, dispersal ability, competition, and population dynamics must all be considered. Further, these mechanisms may not operate consistently or exclusively.

This study models the multiple mechanisms that can govern the spatial pattern of a clonal herb to evaluate whether patchy plant distributions map to particular features of the forest understory. In the study forest, *T. canadensis* is an important species that is distributed among many other tree species, so that although it is clear that

T. canadensis holds an important role in the forest, it is difficult to assign a spatial correlation between the presence of any one tree and the environment experienced by the understory plants (unlike say aspen copses in the western United States). Using observational data and a novel statistical technique that constructs and estimates a network of interactions, however, these ambiguities can begin to be addressed. A population of *Eurybia chlorolepis* (Asteraceae) (Burgess) Nesom (mountain aster) and its habitat features offers a good system to study the role of understory niche in determining causes of population distributions and consequences of forest change.

E. chlorolepis is an understory perennial herb that is found in discrete patches in the old growth cove forests in the Southern Appalachians where *T. canadensis* is a dominant species (Heard, M., unpublished data). Several features of *E. chlorolepis* biology and the forest it inhabits are responsible for its patchy spatial pattern and recommend it for this study. As a clonal herb, *E. chlorolepis* grows in patches of ramets connected by rhizomes. These ramets can have either juvenile or flowering forms. Thus, *E. chlorolepis* can reproduce vegetatively from either form, or sexually from the flowering form. Vegetative growth improves the establishment of a clonal plant, especially in varied or stressful habitats (Stocklin and Winkler, 2004; Takada and Nakajima, 1996). Production of seed through out-crossed pollination allows plants to disperse to new environments, escaping intra-specific competition or resource limitation (Stocklin and Winkler, 2004; Takada and Nakajima, 1996; Volis et al., 2004). *E. chlorolepis* seeds disperse by gravity and water, and as topography of the forest is uneven, they cannot reach all optimal habitat. The percentage of stems within a patch that flower varies greatly across the understory environment, with many patches persisting for years with no sexually reproductive forms at all (unpublished data). This variation in sexual reproduction in different patches creates a clear source-sink dynamic even among long-lived patches. As features of the understory environment lead

to source patches and sink patches, dispersal is limited even in sexually reproductive patches, *E. chlorolepis* is likely to respond significantly to the types of changes in the forest ecosystem predicted by the mortality of *T. canadensis* (Jenkins et al., 1999). The goal of this study is to estimate which parts of this system are most important to *E. chlorolepis* seed production, and therefore to which features population-wide *E. chlorolepis* persistence is most sensitive. To accomplish this, I use Bayesian learning networks (BLN), a flexible method for quantifying a complex network of variables. BLN offers a statistical approach that can structure and quantify direct and indirect associations in a graphical model, can incorporate new information to improve or reassess parameters of that network, and can be used for prediction or inference (McMahon, 2005; Neapolitan, 2001; Pearl, 1988).

Methods

Site

The population of *E. chlorolepis* that was used in this study spanned three sites in a watershed in Cosby Ranger District in Great Smoky Mountains National Park (GSMNP). Twenty patches of plants were located across these three sites. The three sites contained 10, 6, and 4 patches and showed slight differences in both plant density and soil profile. However, all patches demonstrated similar correlational relationships across the patches as within them, so the 20 sites were pooled for analysis. Each patch was at least 20m from any other and had at least a 10m space between its colony and any other colony. Within each of the twenty patches, a single 1m×1m quadrat was established. These quadrats held at least fifteen stems, ensuring that all quadrats were well-established, and were not newly colonized.

Species

Mountain aster, *Eurybia chlorolepis* (Burgess) Nesom, is an understory perennial herb found in the middle elevations of the Southern Appalachians. *E. chlorolepis* reproduces vegetatively through the propagation of new ramets from the base of old ramets and from ramets growing from a rhizome. The rhizome that connects ramets of *E. chlorolepis* is found just below the soil surface. Roots from the rhizome and ramets are generally shallow. Flowering stems produce composite inflorescences that are hermaphroditic and generalist pollinated. The life-cycle of *E. chlorolepis* has a typical perennial stage structure (Figure 18) (Horvitz and Schemske, 1995) (all tables and figures referred to in this part are in an appendix at the end of the part). Ramets sprout new growth in the winter in the form of small ($1 - 3$cm leaf length) two- to four-leaved stems. These remain small until the spring, when they grow as either juvenile plants in rosette form (Figure 19a) or as sexually reproductive individuals with inter-nodal stems (Figure 19b). Flowering stems can reach $0.5m$ in height.

Variable collection

In order to measure niche associations, a number of abiotic measurements were made. Soil moisture was measured as gravimetric water content (GWC), the percent of water in the soil found from dividing the dry weight of the mix of five samples from immediately around a site subtracted from the wet weight, divided by the wet weight. Soil variables were collected from soil samples taken in 2004 in the same manner as the gravimetric water content protocol. Samples were sent to A & L Laboratories in Memphis, Tennessee for soil nutrient analysis, soil pH, organic matter content and NO_3 content. Light was measured using hemispherical photos taken from 1m above the ground at the center of every site. These photos were then filtered into

canopy cover and openness. Using Gap Light Analyzer software (Frazer et al., 1999), estimates of direct, diffuse, and total light transmittance through the canopy at each site were calculated. Direct light transmittance was used in analysis as it is most likely to be important in the dark forest understory.

Plant species richness was recorded at each site in mid-July. Percent ground cover is the total plant cover calculated from digital photos taken 1.6m above patches. These photos were filtered using Adobe Photoshop to a threshold of white and black reflecting vegetative cover and bare ground respectively. Ground cover, as a variable, is the percentage of pixels in the photograph of a site that indicate cover (white pixels). The variable 'Old nodes' (ON) refers to nodes on the caudal root of plant stems resulting from the die-back of previous flowering stems. The number of nodes on a root gives a rough estimate of stem and site age and past flowering patterns.

Statistical Methods

To construct the graphical model of the sites, pairwise correlations were calculated. Sixteen soil variables were recorded from the soil analysis and entered into a principal component analysis to reduce the data set to fewer, uncorrelated variables. The components returned by that analysis were then set in a correlation analysis with other collected varaibles. One component was shown to correlate with population variable of *E. chlorolepis*. that component was then regressed on the soil variables and significant ones were re-entered into the correlation analysis. By returning those original soil variables into the network, a subset of the many soil variables could be used for model building. Correlated variables are confusing when there are many, but a network approach to ecology requires correlation insomuch as they can be interpreted. From the final correlation analysis, a set of variables were selected that

showed strong associations ($P < 0.10$) (Figure 20). A graphical model was then constructed according to these associations and plant ecology and biology.

To analyze the graphical model, I used Bayesian Learning Networks (BLNs). In this approach, the graphical model constructed from the correlations is translated into a directed acyclic graph (DAG). The arrows connecting those variables reflect conditional dependencies of the variables lower on the graph ('children') on those above ('parents'). A variable with no parents is called a 'root' node. Two assumptions must hold for the graphical representation of the variables to be a Bayesian learning network. The graph must be a DAG; that is, it must contain no cycles and have arrows which delimit direction. The graph must also obey the Markov condition, whereby each node in the graph is independent of all other nodes given the value of that node's parents.

When combined with the data, a DAG can be represented as a multivariate normal distribution with a covariance matrix that incorporates the correlations implied by the structural model. An important step in translating a DAG into that multivariate normal distribution is building the conditional precision matrix. Shachter and Kenley (1989) offer an algorithm to transform the variance vector, $\vec{v}$, and the dependence parameters $(b_{ij}|i < j)$ into the precision matrix T. From (Shachter and Kenley, 1989), we define $T(i)$ as the $i \times i$ upper left submatrix of T, $\vec{b_i}$ as the column vector $(b_{1,i}, \cdots, b_{i-1,i})$, and $\vec{b_i'}$ as its transpose. $T(1)$ is simply $1/v_1$, or the precision of the first variable. From there we iteratively build the precision matrix as

$$T_{i+1} = \begin{pmatrix} T(i) + \frac{\vec{b_{i+1}}\vec{b'_{i+1}}}{v_{i+1}} & -\frac{\vec{b_{i+1}}}{v_{i+1}} \\ -\frac{\vec{b_{i+1}}}{v_{i+1}} & -\frac{1}{v_{i+1}} \end{pmatrix}. \tag{19}$$

Parameter estimation for BLN advances directly from this six by six cell precision matrix, T, built with the algorithm in (19). The DAG developed from pairwise correlation analysis was used for parameter estimation. This DAG was entered into the algorithm in (19). Instead of using prior estimates to build the variance-covariance matrix quantitatively, the symbolic Maple kernel in Matlab (The Mathworks, 2003) was used to build the variance-covariance matrix symbolically. Each cell in the six by six covariance matrix contains the symbolic representation of the conditional variance and covariance of the variables. For example, cell (2,2) of the solution gives a symbolic representation as:

$$\beta_{13} * \sigma_1^2 * \beta'_{12} + \beta_{23} * \sigma_2^2 + \beta_{23} * \beta_{12} * \beta'_{12} * \sigma_1^2 = -0.1456. \tag{20}$$

With the symbolic representation of each cell in the updated variance-covariance matrix and the cell values (ten equations in all, from the diagonal and the upper triangle of the matrix), the Maple kernel can be used to solve for each posterior parameter value.

Results

Sixteen soil variables were recorded from the soil analysis and entered into a principal component analysis to reduce the data set to four principal components. These components explained over %93 of the variation in the 16 variables. The third principal component of the soil variables showed significant associations with plant variables in the pairwise correlations. All variables were standardized to make regression and variance components in the network more easily interpretable. The sixteen soil variables were then entered into a multiple regression with the third principal component as a

response variable. Percent potassium cation exchange capacity, total cation exchange capacity, nitrate, and buffer pH were highly significant predictors of this principal component ($R^2 = 0.97$). These were reentered into the pairwise correlation analysis. In this way, from sixteen variables, a subset were determined to be orthogonal and capture variation in soil values. Figure 20 shows the results of the pairwise correlations. These included the total *E. chlorolepis* density (plant number per quadrat), percent ground cover, the listed soil variables, plant species richness in the patches, the average number of old nodes per plant per site, and the response variable, the proportion of all stems in the site that were sexually reproductive.

A network was constructed from the variables with significant correlations. After running the covariance algorithm, a network was estimated. Figure 21 shows the completed graphical model, the posterior estimates of the regression coefficients (besides the arrows), and the conditional variances of the variables (in the legend). Conditional variances reflect the variation in a variable taking into account its dependency on parent variables. Because these variables were standardized, their unconditional variances are all 1.0. In this network 57% of the variation in percent flowering reproductive stems was explained by this network.

Discussion

The network resolved from the correlation analysis combined two soil variables (buffer pH and cation exchange capacity (CEC)) with four biotic variables (plant species richness in a patch, the average number of old nodes per stem at a patch, the number of stems per patch, and the response variable for this graphical model, the percentage of all stems that were at the reproductive stage in a patch). The resolved model (Figure

21) shows a combination of direct and indirect associations influencing the percentage of reproductive stems in a patch. The graph also shows the resolution of two independent collections of variables ('cliques' in the terminology of graphical models) relating to the two soil variables: one shows buffer pH directly and positively associated with the percentage of flowering stems, and the other collection incorporates cation exchange capacity (CEC) into several indirect pathways to the percentage of flowering reproductive stems. CEC is associated negatively with old nodes and plants per patch. Those variables are positively associated with the percentage of reproductive stems. Richness positively associates with the average number of old nodes. The two soil components are closely related in their ecosystem function, yet given the rest of the network, they are not associated.

The soils in these sites have extremely high acidity. Average active soil pH (not buffer pH) was 3.98, with a range from 3.6 to 4.6 (upper and lower %95 intervals were 4.12 and 3.84) (4.0 is the pH of acid rain). In soils this acidic, changes in pH (which are on the log scale) can influence nutrient availability and uptake Xiong et al. (2003); Zak et al. (1994). The active pH, or pH to which plant roots are exposed at any given time can fluctuate significantly. Buffer pH indicates how resistant a soil is to changes in pH. Low buffer pH indicates soil with high stored acidity. Low buffer pH can indicate high organic matter content in the soil or organic content that is not easily broken down (low nitrogen exchange). In the case of these sites, that organic matter can vary from leaf litter from hemlocks, rhododendrons, or deciduous trees(Finzi et al., 1998). Hemlock have been shown to have particularly strong effects on soil acidification and the concomitant effects on nutrient availability in the soil (Finzi et al., 1998), showing high lignin content and low nitrogen turnover **?**Jenkins et al. (1999) .

CEC measures the potential for ion exchange in soils, specifically positively charged ions (cations). Although CEC is commonly associated with the ability to resist acidification, because high CEC can indicate a large fraction of base cations(Ca, Mg, K and Na), high CEC also can indicate a resistance to changing acidification when acidity is high. This combination indicates that although low active pH may be tolerated by plants, the consistency of that low pH may keep *E. chlorolepis* from garnering enough nutrients to recruit juveniles to reproductive forms. The pattern of hemlocks, the most common tree in this watershed (Heard, M., unpublished data) may determine the soil buffer pH and CEC levels, and therefore structure the source-sink (juvenile-reproductive) patterns of *E. chlorolepis*.

In a study comparing hemlock forests with varying ranges of mortality, Jenkins et al. (1999) found that forests with high hemlock mortality had higher net N mineralization, nitrification, and N turnover. The negative association between features of a hemlock dominated soil profile and population sinks of *E. chlorolepis* predicts two potential, opposite responses in this population. *E. chlorolepis* would be predicted to respond positively to the higher N turnover in the soil initially. However Jenkins et al. (1999) warn that this release of N could result in significant leaching, whereby in a high-precipitation environment like these cove forests, the released nitrogen could be lost from the system.

E. chlorolepis demonstrates many of the difficulties that arise in interpreting plant population response to the environment. Because it is perennial, polycarpic (flowering in multiple years), and clonal, *E. chlorolepis* patches are well equipped to persist through difficult years. This life-history strategy can lead to difficulties in generalizing, or predicting population dynamics (Crawley, 1990). In the forest understory, there are many candidate explanations for perennial plant recruitment to sexually reproductive forms. Examples of processes that both positively and negatively influence

recruitment include light Brokaw (1985); Silvertown (2004), soil quality (Xiong et al., 2003; Wardle et al., 2003), herbivores Carson and Root (1999); Maron et al. (2002), competition Denslow (1980); Gustafsson and Ehrlen (2003), and climate Bell et al. (1995); Dzwonko and Gawronski (2002). This study suggests that soils are indeed important to *E. chlorolepis* population dynamics by determining the ability for ramets to transition from juvenile to adult status. This study sheds light on a possible shift in these soil influences if hemlock mortality is substantial. Because fungi, bacteria, and the litter of other plants are all a part of this forest system, further research on the more complex cycles and spatial patterns of soils would aid in predicting the indirect effects of the adelgid invasion on the forest understory community.

Bibliography

Annand, P. N. 1924. A new species of *Adelges* (hemiptera, phylloxeridae. Pan-Pacific Entomologist **1**:79–82.

Bell, D. T., D. P. Rokich, C. J. McChesney, and J. A. Plummer. 1995. Effects of temperature, light and gibberellic acid on the germination of seeds of 43 species native to western australia. Journal of Vegetation Science **6**:797–806.

Bigelow, S. W. and C. D. Canham. 2002. Community organization of tree species along soil gradients in a north-eastern USA forest. Journal of Ecology **90**:188–200.

Brokaw, N. V. L. 1985. Gap-phase regeneration in a tropical forest. Ecology-Washington-D-C. **66**:682–687.

Carson, W. P. and R. B. Root. 1999. Top-down effects of insect herbivores during early succession: Influence on biomass and plant dominance. Oecologia Berlin **121**:260–272.

Clark, J. S. 2003. Uncertainty in ecological inference and forecasting. Ecology **84**:1349–1350.

Crawley, M. J. 1990. The population dynamics of plants. Philosophical Transactions of the Royal Society Series B - Biological Sciences **330**:125–140.

Denslow, J. S. a. 1980. Gap partitioning among tropical rainforest trees. Biotropica **12 (supplement)**:47–55.

Dzwonko, Z. and S. Gawronski. 2002. Influence of litter and weather on seedling recruitment in a mixed oak-pine woodland. Annals of Botany **90**:245.

Finzi, A. C., N. Van Breemen, and C. D. Canham. 1998. Canopy tree–soil interactions within temperate forests: Species effects on soil carbon and nitrogen. Ecological Applications **8**:440–446.

Frazer, G., C. Canham, and K. Lertzman, 1999. Gap light analyzer (GLA).

Ghazoul, J. 2004. Alien abduction: Disruption of native plant-pollinator interactions by invasive species. Biotropica **36**:156–164.

Gustafsson, C. and J. Ehrlen. 2003. Effects of intraspecific and interspecific density on the demography of a perennial herb, sanicula europaea. Oikos **100**:317.

Horvitz, C. C. and D. W. Schemske. 1995. Spatiotemporal variation in demographic transitions of a tropical understory herb - projection matrix analysis. Ecological Monographs **65**:155–192.

Hutchinson, G. E. 1959. Homage to santa rosalia or why are there so many kinds of animals? American Naturalist **93**:145–159.

Jenkins, J. C., J. D. Aber, and C. D. Canham. 1999. Hemlock woolly adelgid impacts on community structure and n cycling rates in eastern hemlock forests. Canadian Journal of Forest Research-Revue Canadienne De Recherche Forestiere **29**:630–645.

Kizlinski, M. L., D. A. Orwig, R. C. Cobb, and D. R. Foster. 2002. Direct and indirect ecosystem consequences of an invasive pest on forests dominated by eastern hemlock. Journal of Biogeography **29**:1489–1503.

Maron, J. L., J. K. Combs, and S. M. Louda. 2002. Convergent demographic effects of insect attack on related thistles in coastal vs. continental dunes. Ecology **83**:3382.

McMahon, S. M. 2005. Quantifying the community: using Bayesian learning networks to find structure and conduct inference in invasions biology. Biological Invasions **7**:833–844.

Neapolitan, R. E. 2001. Learning Bayesian Networks. Pearson, Prentice Hall, Upper Saddle River, NJ.

Orwig, D. A. 2002. Ecosystem to regional impacts of introduced pests and pathogens: historical context, questions and issues. Journal of Biogeography **29**:1471–1474.

Orwig, D. A. and D. R. Foster. 1998. Forest response to the introduced hemlock woolly adelgid in southern new england, usa. Journal of the Torrey Botanical Society **125**:60–73.

Parker, I. M., D. Simberloff, W. M. Lonsdale, K. Goodell, M. Wonham, P. M. Kareiva, M. H. Williamson, B. Von Holle, P. Moyle, J. E. Byers, and L. Goldwasser. 1999. Impact: toward a framework for understanding the ecological effects of invaders. Biological Invasions **1**.

Pearl, J. 1988. Probabilistic reasoning in intelligent systems: networks of plausible inference. Morgan Kaufmann, San Mateo, California.

Pulliam, H. 1988. Sources, sinks, and population regulation. American Naturalist **132**:652–661.

Pulliam, H. R. 2000. On the relationship between niche and distribution. Ecology Letters **3**:349–361.

Shachter, R. D. and C. R. Kenley. 1989. Gaussian influence diagrams. Management Science **35**:527–550.

Silvertown, J. 2004. Plant coexistence and the niche. Trends in Ecology and Evolution **19**:605–611.

Simberloff, D. and B. Von Holle. 1999. Positive interactions between nonindigenous species: invasional meltdown? Biological Invasions **1**:21–32.

Stocklin, J. and E. Winkler. 2004. Optimum reproduction and dispersal strategies of a clonal plant in a metapopulation: a simulation study with *Hieracium pilosella*. Evolutionary Ecology **18**:563–584.

Stoetzel, M. B., 2002. The Smithsonian Institute National Collection of Insects. Chapter history of the introduction of *Adelges tsugae* based on voucher speciments *in* B. Onken, R. Reardon, and J. Lashomb, editors. Hemlock woolly adelgid in the Eastern United States symposium, volume 12. New Jersey Agricultural Experiment Station and Rutgers University New Brunswick, NJ.

Takada, T. and H. Nakajima. 1996. The optimal allocation for seed reproduction and vegetative reproduction in perennial plants: An application to the density-dependent transition matrix model. Journal of Theoretical Biology **182**:179–191.

The Mathworks, I., 2003. Matlab.

Volis, S., K. Verhoeven, S. Mendlinger, and D. Ward. 2004. Phenotypic selection and regulation of reproduction in different environments in wild barley. Journal of Evolutionary Biology **17**:1121–1131.

Wardle, D. A., G. W. Yeats, W. Williamson, and K. I. Bonner. 2003. The response of a three trophic level soil food web to the identity and diversity of plant species and functional groups. OIKOS **102**:454–56.

Xiong, S. J., M. E. Johansson, F. M. R. Hughes, A. Hayes, K. S. Richards, and C. Nilsson. 2003. Interactive effects of soil moisture, vegetation canopy, plant litter and seed addition on plant diversity in a wetland community. Journal of Ecology **91**:976–986.

Zak, D. R., D. Tilman, R. R. Parmenter, C. W. Rice, F. M. Fisher, J. Vose, D. Milchunas, and C. W. Martin. 1994. Plant production and soil microorganisms in late-successional ecosystems: A continental-scale study. Ecology Tempe **75**:2333–2347.

Appendix: Tables and Figures

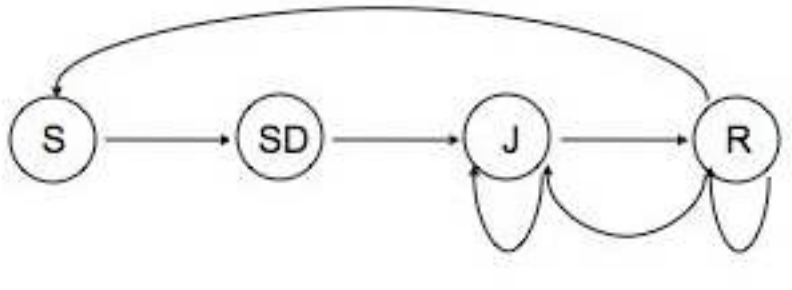


Figure 18: Stage diagram of *E. chlorolepis*. Stages are in circles (S = seed, SD = seedling, J = juvenile, and R = reproductive), and transitions, reproduction, and survival are shown as arrows.

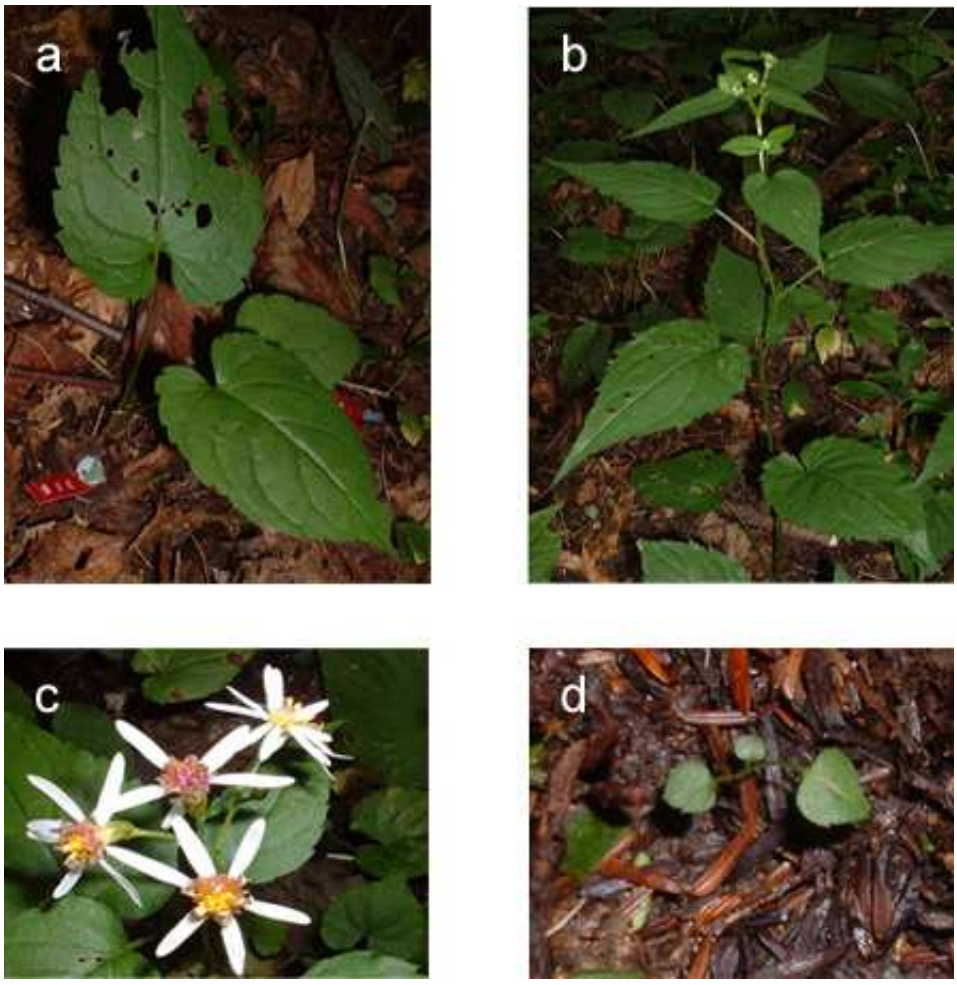
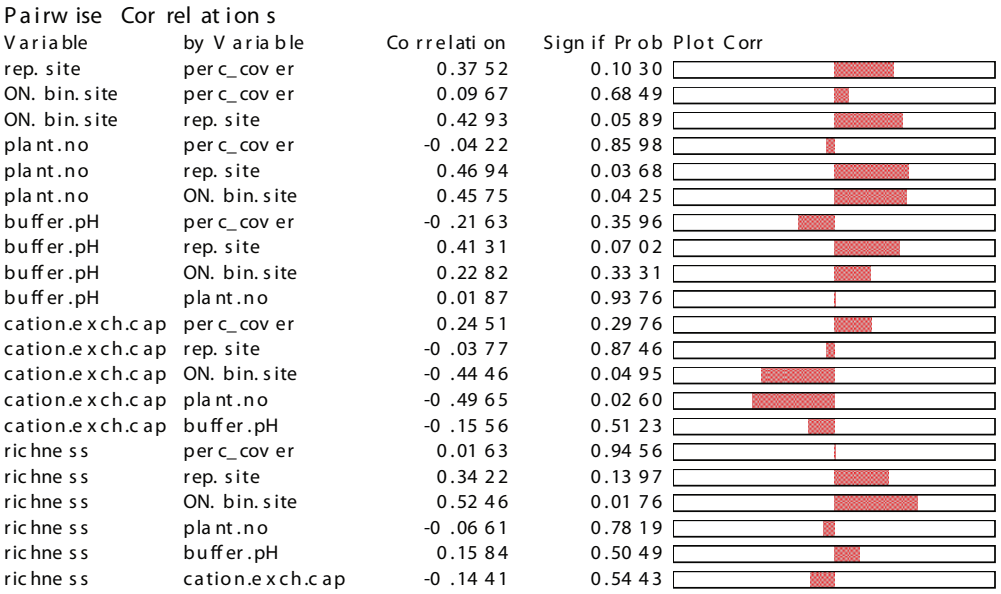


Figure 19: Photographs of *E. chlorolepis*. The juvenile stage shows leaves in a rosette form (a), while the reproductive form has inter-nodal stems (b). Inflorescences show disc and ray flowers (c). Disc flowers show various stages (colors) of reproductive receptivity. A seedling is shown with one of two cotyledons (d).



Pairwise Correlations

Variable	by Variable	Correlation	Signif Prob	Plot Corr
rep. site	perc_cover	0.3752	0.1030	
ON. bin. site	perc_cover	0.0967	0.6849	
ON. bin. site	rep. site	0.4293	0.0589	
plant.no	perc_cover	-0.0422	0.8598	
plant.no	rep. site	0.4694	0.0368	
plant.no	ON. bin. site	0.4575	0.0425	
buffer.pH	perc_cover	-0.2163	0.3596	
buffer.pH	rep. site	0.4131	0.0702	
buffer.pH	ON. bin. site	0.2282	0.3331	
buffer.pH	plant.no	0.0187	0.9376	
cation.exch.cap	perc_cover	0.2451	0.2976	
cation.exch.cap	rep. site	-0.0377	0.8746	
cation.exch.cap	ON. bin. site	-0.4446	0.0495	
cation.exch.cap	plant.no	-0.4965	0.0260	
cation.exch.cap	buffer.pH	-0.1556	0.5123	
richness	perc_cover	0.0163	0.9456	
richness	rep. site	0.3422	0.1397	
richness	ON. bin. site	0.5246	0.0176	
richness	plant.no	-0.0661	0.7819	
richness	buffer.pH	0.1584	0.5049	
richness	cation.exch.cap	-0.1441	0.5443	

Figure 20: Pairwise correlations show the direction and strength of relationships between variables.

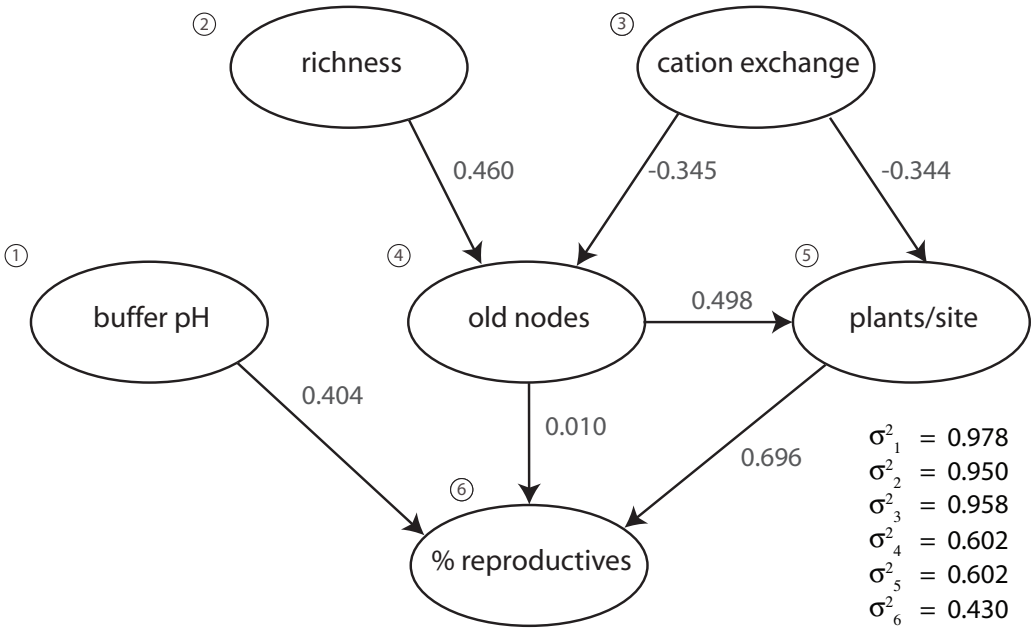


Figure 21: The network illustrates the relationship between soil variables and components of *E. chlorolepis* site life-history characteristics. Conditional variances are in a legend at the bottom right.

Vita

Sean McMahon was born in Manhattan, KS in October 1968 to Sarah Lynne and Adrian Michal McMahon. The family moved to Philadelphia, PA in the summer of 1972, just months before adding Jeff McMahon to the already lively family dynamic. In 1992, Sean received a B.A. in Honors in Liberal Arts from the University of Texas at Austin with special honors in English Literature. His senior thesis was on the psychoanalytic theory of Jacques Lacan and Joseph Conrad's *Heart of Darkness*. Sean then received a M.A. from University College Dublin in Anglo-Irish literature in 1993. His masters thesis was entitled *The violence of style*. It sought to unite political and psychoanalytic theory using James Joyce's *Ulysses* and Samuel Beckett's *The trilogy: Malloy, Malone Dies, and The Unnamable*. Returning to the United States, Sean decided to shift topics. He re-enrolled at the University of Texas and began taking biology courses. He worked in Stuart Gilbert's lab with Greg Sword, and then began a three-year stint in Eric Pianka's lab investigating lizard diets in a fire regime. After being accepted to the University of Tennessee, Knoxville, Sean moved to Knoxville and began his Ph.D. studies under Professor Jim Drake. At Tennessee, Sean combined field work in the Smokies on an understory herb with intense study of statistics. He received his M.S. in Statistics in the Summer of 2006. He now has a post-doctoral research position at the Nicholas School of the Environment at Duke University studying forest dynamics under the supervision of Professor Jim Clark.