BIODIVERSITY AND SCALE: DETERMINANTS OF SPECIES RICHNESS IN GREAT SMOKY MOUNTAINS NATIONAL PARK

R. Todd Jobe

A dissertation submitted to the faculty of the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Curriculum of Ecology.

Chapel Hill

2006

Approved by:
Advisor: Peter S. White
Reader: Robert K. Peet
Reader: Dean L. Urban
Reader: Michael A. Jenkins
Reader: Alan S. Weakley

ABSTRACT

R. Todd Jobe: Biodiversity and scale: determinants of species richness in Great
Smoky Mountains National Park
(Under the direction of Peter S. White)

Species richness is the number of species in a given area or sample and is the most fundamental measure of biodiversity. It results from the aggregation of individual species whose distributions are influenced by processes operating on a wide range of scales. Estimating and understanding species richness at landscape scales (10^3-10^6 ha) is not easily achieved from small sample areas that can be completely inventoried. In particular the spatial structure of environments makes the richness observations across a landscape non-additive. This dissertation develops the vital links between the spatial structure of ecological factors that are hypothesized to control species richness, spatial variation in species composition, and the sampling strategies used to measure species richness.

I present a method for objectively and iteratively assessing patterns of biodiversity. This method builds upon “ecological zipcodes” that classify the landscape by energy flux, temperature, and precipitation. I also present a model of human energetic expenditure during walking that can be applied at landscape scales. I use this model to analyze sampling bias associated with accessibility for vegetation surveys. I used both the "ecological zipcodes" and the model of accessibility to design efficient and representative biodiversity samples based on clustered-stratified sampling. Finally, I assess the reliability of richness estimators that incorporate turnover in species composition.

My results illustrate that efficient and representative richness assessment is possible, even with little *a priori* knowledge about the spatial structure of species richness. They also demonstrate that typical biodiversity assessments show a strong bias in accessibility that is both a product of the spatial structuring of samples as well as environment. This bias is significant even for small biases in sample accessibility. Also, I show that though clustered sampling designs capture multiple scales of aggregation, their representativeness is very sensitive to stratification. Finally, my results show that species richness estimates that incorporate turnover are confounded by the interaction between sample size and environmental heterogeneity. Only when controlling for these effects, can information about the spatial turnover in species composition be effective in estimating species richness.

To my mom who taught me self-sacrifice.

To my dad who taught me perseverance.

To my wife who understands me.

ACKNOWLEDGEMENTS

I gratefully acknowledge all the people and organizations that have helped me in the research and writing of this dissertation. Peter S. White, my advisor, is a cornucopia of deep ecological thoughts. He has always been supportive of my work and believed that I could accomplish great things. He is an intellectual giant, and I humbly and gratefully stand upon his shoulders. I also thank my committee. Robert K. Peet, head of the Curriculum in Ecology at UNC, has helped me with his careful eye for detail, supportive words, and common love for all things technological. I also thank him for providing the excellent computing resources that have been integral in my completion of this dissertation. Michael A. Jenkins, ecologist for Great Smoky Mountains National Park, has provided me with the opportunity to ask interesting questions in one of the worlds greatest reserves of biodiversity. I thank him for providing sources of funding for this work, and being excited about implementing my ideas. Dean L. Urban of Duke University has contributed greatly to the quantitative backdrop of this dissertation. His insights on sampling design and spatial analysis are invaluable resources for me. I also thank Alan S. Weakley, curator of the University of North Carolina Herbarium, for the flexibility and fervor with which he approached my work. I thank him for allowing me to exploit his vast knowledge of taxonomy and plant communities of North Carolina: a resource that has no equal. I must also thank those who at other times served on my committee, but because of changes in topic or scheduling conflicts were replaced: John F. Bruno, and Stephen P. Hubbell. I appreciate their insights and willingness to work with me.

I thank the institutions that provided funding for my research. The University of North Carolina Ecology Curriculum provided a merit fellowship and my teaching assistantship. The University of North Carolina Biology Department provided funding via the Coker Fellowship for my final semester. The Great Smoky Mountains National Park provided financial support throughout my studies in the form of various contracts that eventually became much of the substance of this dissertation.

I want to acknowledge my fellow graduate students, past and present, of the UNC Plant Ecology Laboratory. The free flow of ideas and opinions and the complete absence of competition speak to the character of the people that make up this lab. There are a few members that I must mention by name for their significant contributions to my work. Jason Fridley has been a fierce friend whose clarity of thought has been immensely helpful as I have chased various rabbit trails in my work. Joel Gramling, a friend and cohort, has been a willing sounding board for thoughts and frustrations about my research. David Vandermast, through his broad perspective and open mind, has been instrumental in the development of my ecological schema.

TABLE OF CONTENTS

Chapter

LIST OF TABLES

LIST OF FIGURES

Chapter 1

Introduction: The scale dependence of species richness – Fundamental questions and challenges

Introduction

The most fundamental measure of biodiversity is species richness – the number of species in an area or sample. Understanding what shapes patterns of species richness is one of the central problems of ecological science. Yet, for all the importance of richness, there are many seemingly simple questions surrounding this fundamental ecological property that remain unanswered. In fact, the total amount of richness on Earth is not known within one order of magnitude. What makes quantifying patterns of species richness so difficult? There are three main factors. First, any pattern in richness is really an aggregate pattern of individual species distributions (Gleason 1926). In order to understand the ecological processes that shape patterns of richness, it is necessary first to understand the processes that shape the individual species that make up these patterns. The number of species can be very large with a very complex web of interactions. In this dissertation, my analyses assume that two of the most important structuring processes are niche-environment interactions (disturbance is included here as the main effect is through disturbance effects on environment) and vagility-spatial template interactions. Second, patterns of richness and the processes that shape them happen across many scales simultaneously (Rosenzweig 1995). Ecological processes influencing species richness at a particular scale are not necessarily the same ones operating at larger or smaller scales. Third, richness patterns are ever changing as species disperse, are disturbed, and adapt to their environments. Additionally, modern anthropogenic effects have caused and continue to cause changes in richness patterns (Thomas *et al.* 2004;Vitousek *et al.* 1997;Stohlgren *et al.* 1999). The richness assigned to a spot today is not the richness we will see in future decades nor what was present in the past.

Perhaps these complexities arise because our ecological data are inadequate. If samples were large, continuously resampled, and collected at multiple scales, then the abundance of all species could be completely mapped and all spatial and temporal scales would be present in the data. Patterns of richness would be boiled down to the intersection of measured species ranges whose environmental correlates and response to spatial landscape structure were well-described. However, even if such an effort were practical, it would be a waste of sample effort. The relative abundance distributions of species dictate that most of the data collected would be redundant (Preston 1948;Fisher *et al.* 1943) since there would be unnecessarily large numbers of observations of the most common species. This built-in redundancy highlights the importance of understanding the processes that control the turnover of species across sample units of known richness that are by necessity (or by design) spatially disjunct and restricted to a subset of the study area in terms of sample area. Further, the underlying problem that results from the fact that our richness observations are based on small and disjunct samples is that species turnover complicates the extrapolation of total richness from samples and makes the richness of samples non-additive across landscapes. This dissertation focuses on efficiently quantifying, analyzing, and understanding the factors that structure species richness and turnover, particularly the physical environment, and their influence on species richness across scales.

Much of the turnover associated with species richness patterns is correlated with environmental heterogeneity, especially at a landscape scale (Nekola & White 1999;Whittaker 1956). Understanding how species composition changes along a few environmental gradients, also known as gradient analysis, has a long history in ecology (Whittaker 1967). There are a myriad of processes, however, that control the turnover of species distributions at many different scales, not just environment. Consequently, one of the key issues for understanding species richness across broad scales is learning how to disentangle the effects of environmental heterogeneity from these other processes, and how to do that with necessarily disjunct observations. Further, our assessments of these effects are sensitive to the grain and extent of observations (Palmer & White 1994). Whether it be for the purpose of efficiently capturing as many species as possible across a large area (e.g. the All-Taxa Biodiversity Inventory of Great Smoky Mountains National Park; White *et al.* 2000), estimating total species richness of a landscape, or estimating the effects of disturbance on species richness, it is the establishment of independent correlation between environmental heterogeneity and species turnover, and the relationship of these correlations to grain and extent that are the vital links to understanding species richness at a landscape-scale.

In this dissertation, I develop the linkage between species richness and both environmental gradients and spatial effects related to grain and extent through answering four fundamental questions. 1) Given species richness assessment as a goal and given that we may not know *a priori* how richness patterns associate with environment, how should we go about capturing many species in a wide variety of environments and iteratively improve our understanding of what variables shape species richness patterns? 2) Given that species distributions and samples rarely exhibit random spatial aggregation at a landscape scale, how does the interface between species distributions and sample designs influence our interpretation of ecological processes? 3) How do we design long-term samples that efficiently maximize species capture, while disentangling environmental effects from other spatially autocorrelated processes? And, 4) given that capturing all species is impossible and that most richness estimators are meant to estimate local diversity, how can true richness be extrapolated from the observed richness of a sample for large, heterogeneous landscapes? In this dissertation, I answer these questions in the context of The Great Smoky Mountains National Park (GSMNP). Below, I show how each chapter addresses these questions and describe the fundamental ecological principles that form the base of each chapter.

Chapter Summaries

In Chapter 2, I present a strategy for accurate assessment and monitoring of biodiversity over large, complex landscapes (areas of ~10^3-10^6 ha.). The premise of this chapter is that at the start of any biodiversity assessment little *a priori* knowledge of how environment correlates with species distributions is available. Further, the individual species that make up patterns of biodiversity, respond to a variety environmental factors, and the correlations between environment and species composition are likely to be complex. In spite of these complexities, every biodiversity assessment must begin by answering the question "How can sample effort be most efficiently allocated to maximize total richness?" Once data are collected, managers must also be able to incorporate lessons learned from prior observations to ask more refined questions about particular species or taxonomic groups.

The general strategy I present in chapter 2 addresses the issues outlined above. This strategy is then applied to the specific case of GSMNP. Beginning from the important ecological factors of energy flux, temperature, and precipitation, a landscape classification for stratifying biodiversity assessments is presented. This classification, termed "ecological zipcodes", efficiently represents the combined environmental distribution of these factors. I also introduce a method for clustering samples in a landscape to account for both fine-grained and coarse-grained variation in species distributions. This sample design is elaborated and tested in Chapter 4. Finally, I show how studies can be iteratively designed to incorporate additional variables such as soil chemistry and disturbance, or improve the original ecological zipcodes.

In chapter 3, I move from discussion of the interactions between species and environment to the interface between species distributions and the humans that survey, conserve, or disturb them. Landscape structure controls both the distribution of species and human interactions with species. One major way that landscape structure influences sampling is through accessibility. All else being equal, the more difficult a place is to walk to, the less intensely people visit, disturb, or study it.

In chapter 3, I develop a cost-benefit approach to determining the impact of these human influences on sampling design, patterns of disturbance, the distributions of species, and ecosystem function. I do so through the creation of a model that estimates the energetic cost of walking through a landscape. This model considers the ways that distance, slope, stream crossings, and vegetation influence the energetic cost of walking. I apply this model to GSMNP to assess the correlations between vegetation communities in the Park and survey data. I discuss the myriad ecological questions to which this accessibility model can be applied including the effects of human disturbance as well as conservation planning, management, and restoration.

In chapter 4, I test the effectiveness of using both ecological zipcodes and the accessibility model in survey design for long term biodiversity monitoring. In addition I discuss four issues that influence design effectiveness at landscape scales. First, samples must reflect the environmental variability of the landscape as variation in environment influences variation in species composition. Second, samples should capture a broad array of communities... Third, biodiversity surveys must account for the spatial autocorrelation present in species distributions. Finally, because the cost of sampling increases as accessibility decreases and because no monitoring program will continue if fieldwork is unreasonably difficult, sample sites must be accessible.

I present a sample design that fulfills these qualifications. Following a stratified-clustered design, I generate a set samples for GMSNP. Sample stratification derives from ecological zipcodes developed in Chapter 2. The clustered design allows testing and controlling for spatial autocorrelation in vegetation patterns and in addition improves sampling efficiency. I analyze the effect of weighting sites by the accessibility model developed in Chapter 3 of human accessibility. I show that although stratification by ecological zipcode guarantees environmental variety, such samples do not capture a greater variety in species composition than a random sample. I also show that weighting sights toward more accessible locations causes significant changes in the environmental representativeness.

In Chapter 5, I draw upon lessons learned in chapters 2-4 about the relationship between patterns of species richness, variation in environment, and sampling design to better assess species richness at landscape scales. There is a long history in ecology of estimating total species richness from sample data, and a wide variety of techniques for doing so. In spite of this history, these estimators typically underestimate species richness. One hypothesis for why estimators tend to underestimate total richness is that they do not explicitly account for increases in species richness due to turnover in species composition.

There are estimators, however, that attempt to explicitly incorporate variation in species composition through space. I compare these estimators against classic richness estimators that do not include turnover through analysis of a dataset of native trees in GSMNP, The results of this chapter suggest that separating the biases associated with small sample sizes, while controlling for environmental heterogeneity, can improve the richness estimates based on the turnover of species composition with distance.

In the Chapter 6, I synthesize the results from each chapter and show how this research contributes to closing the gaps in understanding that exist between ecological processes and richness patterns, between human influences and richness patterns, between sampled and true richness patterns, and between local richness patterns and landscape ones. I also discuss the new research questions that emerge from my work.

References

Fisher R.A., Corbet A.S. & Williams C.B. The Relation Between the Number of Species and the Number of Individuals in a Random Sample of an Animal Population. The Journal of Animal Ecology 12[1], 42-58. 1943.

Gleason H.A. The Individualistic Concept of the Plant Association. Bulletin of the Torrey Botanical Club 53[1], 7-26. 1926.

Nekola J.C. & White P.S. (1999) The Distance Decay of Similarity in Biogeography and Ecology. *Journal of Biogeography* 26, 867-878

Palmer M.W. & White P.S. (1994) Scale Dependence and the Species-Area Relationship. *American Naturalist* 144, 717-740

Preston F.W. The Commonness, And Rarity, of Species. Ecology 29[3], 254-283. 1948.

Rosenzweig M.L. (1995) *Species Diversity in Space and Time*. Cambridge University Press, Cambridge, UK.

Stohlgren T.J., Binkley D., Chong G.W., Kalkhan M.A., Schell L.D., Bull K.A., Otsuki Y., Newman G., Bashkin M. & Son Y. (1999) Exotic Plant Species Invade Hot Spots of Native Plant Diversity. *Ecological Monographs* 69, 25-46

Thomas C.D., Cameron A., Green R.E., Bakkenes M., Beaumont L.J., Collingham Y.C., Erasmus B.F.N., De Siqueira M.F., Grainger A., Hannah L., Hughes L., Huntley B., Van Jaarsveld A.S., Midgley G.F., Miles L., Ortega-Huerta M.A., Peterson A.T., Phillips O.L. & Williams S.E. (2004) Extinction Risk From Climate Change. *Nature* 427, 145-148

Vitousek P.M., Mooney H.A., Lubchenco J. & Melillo J.M. (1997) Human Domination of Earth's Ecosystems. *Science* 277, 494-499

White P.S., Morse J., Harris F., Langdon K., Lowe R., Nichols B., Parker C., Pickering J. & Sharkey M. The science plan for the all taxa biodiversity inventory in Great Smoky Mountains National Park, North Carolina and Tennesse. Report to board of directors of Discover Life In america , 15. 2000.

Whittaker R.H. Vegetation of the Great Smoky Mountains. Ecological Monographs 26[1], 1-80. 1956.

Whittaker R.H. (1967) Gradient analysis of vegetation. *Biological Reviews of the Cambridge Philosophical Society* 42, 207-264

Chapter 2

Ecological zipcodes: a strategy of environmental classification for biodiversity assessment

Abstract

There is a great need for simple yet flexible and biologically meaningful ways of classifying landscapes using readily available data for the purposes of inventorying and monitoring biodiversity. In this paper we describe a strategy developed for the All Taxa Biodiversity Inventory (ATBI) in Great Smoky Mountains National Park (NC and TN, USA) that establishes an objective and iterative approach to assessing and understanding patterns of biological diversity. Our strategy begins with the selection of ecologically important and available environmental factors. For the ATBI, we selected variables based on energy flux, temperature, and moisture. We show how to classify and combine these variables into a single classification scheme of "ecological zipcodes" that efficiently represents the combined environmental distribution of these factors. We also present a method for clustering samples of a landscape survey to account for both fine-grained and coarse-grained variation in species distributions. Finally, we show how studies can be iteratively designed to incorporate additional variables such as soil chemistry and disturbance, either as continuous variables or as contrasts for direct hypothesis testing using the ecological zipcodes. By translating a complex landscape into a simple yet biologically relevant distribution of ecological zones, our strategy can contribute a key structural component to current efforts of biodiversity management.

Keywords

landscape classification, All-Taxa Biodiversity Inventory, spatial autocorrelation, Great Smoky Mountains National Park

Introduction

Accurate assessment and monitoring of biodiversity across large, heterogeneous landscapes is a daunting task. At these scales (~10^3-10^6 ha) exhaustive surveys are impractical, and assessments must be based on a limited number of well-placed samples. The question facing land managers is: “How should sample effort be distributed across a large landscape in a way that maximizes the number of species that are observed?” A common approach to this question would be to consider the variety of environments that are present in the landscape, since environmental variation is fundamentally correlated with variation in species composition.

For many taxa we know very little about how species distributions correlate with environment, though we do know the major environmental determinants of vegetation pattern on the landscape. Environmental gradients are complex and overlain by patterns of human and natural disturbance. In addition, spatial factors, including the size and isolation of habitats (Hanski & Gilpin 1991;Ricketts 2001) as well as dispersal history (Skellam 1951;Clark *et al.* 1998) also influence species distributions. Regardless, the important environmental variables controlling species distributions are likely to change between taxa. The possible number of variables that could serve to predict species distributions forms a long list (Table 2.1). Further, the influences of environment as well as patterns of biodiversity vary with scale (Rosenzweig 1995). The result of this complexity is that, even in the most intensely researched areas in the world, biodiversity inventories are vastly inadequate. There is a great need for efficient biodiversity surveys that capture the wide array of environments and spatial scales that species are distributed across. When fieldwork starts, we are unlikely to know *a priori* which variables are the most important environmental predictors of species distributions—and, in any case, the most important variables are likely to vary from one taxonomic group to another. Therefore, we must make some educated decisions at the outset about which environmental variables apply to a wide range of taxa, and what spatial scales that must be captured by the sample. Then, we must be able to iteratively improve our choices as data are collected and correlations with other variables become apparent (Turner 1989;Thompson 1991). Also, by creating a framework for structured observation, we create a scheme for plot sampling that will add a repeatable dimension to biological inventory that will support monitoring. This framework brings together two related scientific fields—ecology and taxonomy.

We advocate a landscape classification approach to biodiversity survey design that achieves these goals and is easily implemented by managers. We describe this approach with particular reference to the All-Taxa Biodiversity Inventory (ATBI) conducted in Great Smoky Mountains National Park. The ATBI Science Plan (White *et al.* 2000) divides the overall approach into two components: traditional taxonomic inventory (simple exploration without quantitative stratification and without explicit quantification of sample grain and extent) and structured observation. The structured observation strategy creates an objective and iterative approach to assessing and understanding patterns of biodiversity. This strategy does not address specifically the design of sampling units. That, of course will vary between studies and taxonomic groups (e.g. quadrates for vegetation; trappings or timed searches for animals). Rather, we describe how a plot-monitoring network should be arranged to capture the important environmental variation in the study area.

This strategy begins with the selection of ecologically important and available variables. We show how to classify and combine these variables into a single classification that efficiently represents their states: an ecological “zipcode”, so named because it assigns an ecological address, in which values of particular digits correspond to states of a particular environmental factor, to all locations in the Park. (Unlike postal zipcodes, however, these ecological zipcodes define a unique environmental address, but one that repeats in space, so it is an address that is not spatially unique) We also present a method for clustering sampling units to account for small-scale variation in species distributions. We describe how to increase sampling efficiency through the inclusion of a measure of site accessibility. Finally, we show how studies can be iteratively designed to incorporate soil variables, disturbance, and spatial pattern, either as continuous variables or as contrasts using the ecological zipcodes.

The first decision: which environmental factors are important?

Researchers usually have reasonable “first guesses” about which ecological variables are responsible for the variation in species composition across a landscape, at least for more macroscopic and well-studied organisms such as vertebrate animals and vascular plants. Taken from ecological first principles, these variables form the base and justification of the ecological zipcodes. Of course, *a priori* knowledge of which variables predict variation in species composition is limited before data are collected. As data are collected and analyzed, new variables may show correlations with species composition. Some or all of the initial set of variables may be uncorrelated with changes in species composition. These initial variables represent a type of informed null hypothesis to be supported or rejected by the data. Iterative improvement in our knowledge of important ecological variables predicting variation in species composition is one of the goals of biodiversity assessment, and of the ecological zipcodes.

In our study, we selected temperature, water availability, and insolation as three important ecological variables for variation in species composition in GSMNP. Each of these variables has been shown to correlate with species composition for a wide variety of taxa and across many scales (Hawkins *et al.* 2003). Temperature controls the metabolic rates of animals (West *et al.* 2001), and cold temperatures limit potential investment in plant growth (Korner 1998). These physiological impacts have consequences for both species abundance and diversity (Allen *et al.* 2002). At a landscape scale, the major variation in temperature of GSMNP comes from adiabatic cooling. Elevations in the Park range from 266m to 2029m. Given an adiabatic lapse rate of ~5°C/1000m, the mean temperature difference for the elevation ranges of GSMNP is about 9°C. The adiabatic cooling associated with increased elevation is one of the most significant driving factors of plant community composition in park (Whittaker 1956). Temperature variations can also be seen as distance from streams increases within the Park. Nighttime katabatic winds from high elevations tend to follow stream drainages and lower the temperature close to streams (Geiger 2003).

Water availability, like temperature, is also correlated with variation in species composition in a variety of systems. Water and carbon dioxide provide the chemical reactants for photosynthesis in plants. Plants vary widely in their water use efficiency. Water is required by animals either for respiration or as an environmental medium, and correlates strongly with composition of animal communities (Andrewartha & Birch 1954;Root 1988). Energy input from the sun interacts with precipitation (Stephenson 1990). Biomass and evapotranspiration of plants as well as soil and bedrock types also influence water availability. GSMNP receives 150-200cm of rainfall per year. Precipitation in the Park increases with elevation because the cooler air at high elevations increases condensation resulting in orographic rainfall. Precipitation runoff and the percolation of water through the soil tend to concentrate water in coves and stream drainages.

Closely allied with and strongly influencing temperature, insolation is the number of photons reaching the earth's surface. Insolation controls the amount of energy available for photosynthesis in plants, which forms the base of the carbon chain in terrestrial ecosystems. Insolation also influences phenology directly through temperature or photoperiod, and other life-history traits may be sensitive to photoperiod or the intensity of solar radiation (Rathcke & Lacey 1985). Local insolation also impacts evaporation and water availability. At local scales, insolation is dependent upon slope, aspect, elevation, and relative exposure. At a global scale, the correlation between aspect and insolation increases toward the poles. South-facing slopes have a higher angle of incidence to solar radiation than do north-facing or slopes and are more strongly insolated. Insolation increases with elevation, because the atmospheric scattering of light is reduced at higher elevations. The most notable variations in insolation for GSMNP are seen in areas of relative shade versus areas receiving direct sunlight such as north-facing coves and south-facing ridges.

These three ecological variables (temperature, water availability, and insolation) served as our first estimation of the most important variables controlling variation in species composition across GSMNP. In the earliest gradient analysis of GSMNP, elevation and relative exposure as surrogates of energy input, temperature and water availability were strongly correlated with vegetation patterns (Whittaker 1956). More generally, some combination of these variables control species composition for most terrestrial ecosystems (Holdridge 1947;Stephenson 1990;Stephenson 1998). We suggest them as the best starting point for classifying landscapes for biodiversity assessment. As with most ecological variables, however, measuring each of these variables at scales appropriate for capturing variation in species composition is difficult, if not impossible. This limitation highlights the next step in selecting the appropriate variables to sample across, when monitoring biodiversity: selecting appropriate measures for the ecological variables of interest.

In general, the ecologically important variables cannot be measured directly, but rather must be viewed through the lens of correlated variables. The number of possible data sources is large (Table 2.1). The selection of all possible variables is constrained because we must be able to represent these variables continuously across large areas and at a spatial resolution small enough to capture meaningful patterns of compositional variation. These chosen measures must be correlated with the ecological variables selected earlier. Further, they must be implemented in a GIS, which provides spatially explicit framework for selection of survey sites and the analysis of data (Scott *et al.* 1987). The set of available data may constrain the list of ecological variables that can be used as correlates with changes in species composition. Soil chemistry, for instance, is a major determinant of vegetation patterns, which forms the carbon base of the ecosystem, and would be a good candidate for inclusion in the landscape classification. Estimates of soil chemistry can be developed from maps of major geologic units. Soil chemistry can vary at relatively small scales (Cain 1931;Bratton 1976), however, and little or no reliable data is available at these small scales with comprehensive coverage for GSMNP. Temperature, water availability, and insolation, on the other hand, can be modeled at a small scale for the entire extent of the Park (Table 2.1). We chose elevation, hillshade and a topographic measure of relative wetness to represent the combined influences of our latent environmental variables (Figure 2.1). All three of these variables may be derived from digital elevation models (DEMs), which are readily available at a small grain (10m^2) for all of the United States (see http://seamless.usgs.gov). Elevation is correlated with temperature (higher elevations have lower temperatures due to adiabatic cooling) and water availability (higher elevations receive greater precipitation from orographic rainfall). Hillshade is a measure of insolation and is calculated as the proportion of direct sunlight striking the landscape for a given solar azimuth and altitude, accounting for topographic shading. We used a solar azimuth and altitude corresponding to a SW exposure to the sun. Calculated in this way, hillshade is also correlated with both temperature (SW-facing slopes have greater insolation than north-facing slopes) and water availability (the greater incoming radiation on exposed SW-facing slopes results in greater evaporative losses). As a measure of relative wetness, we used the topographic convergence index (TCI) (Moore *et al.* 1991;Wolock & Mccabe 1995;Yeakley *et al.* 1998). It is calculated as the logarithm of the ratio between upslope contributing area and the tangent of local slope. In general, low-elevation, cove-like habitats have high relative wetness and high-elevation ridges have low relative wetness.

This process of selecting ecologically important variables and correlated latent variables results in a list of variables to serve as digits in the final ecological zipcode classification. At this point, it is important to minimize the redundancy in the set of variables. For the practical purpose of keeping the classification as simple as possible and to maximize the independence of explanatory variables within the ecological zipcodes (Mendenhall *et al.* 1998), we checked to see that each of the explanatory variables were uncorrelated. Verifying independence among explanatory variables is especially important when the variables are derived from the same data source as in the case of elevation, hillshade, and TCI, which are all derived from a single DEM. Table 2.2 shows the correlation and covariation matrices for elevation, TCI, and hillshade for GSMNP. Since the Pearson’s correlation coefficients (*r*) of paired variables are all relatively close to 0, we have qualitative confidence that they are distributed independently in space, even though they were derived from the same DEM. With the correlation analysis serving as the final filter, we are left with a set of variables that are thought to be ecologically important for species distributions and are relatively independent of one another.

The second decision: how should factors be classified?

The second step in using continuous variables to create ecological zipcodes is transforming them into a ranked list of nominal variables, each with a fixed number of levels. This process begins with a determination of the total number of classes desired. The total number of environmental categories in the final classification will be the product of the number of levels in each variable. In our GSMNP classification, we selected 5 levels of elevation, 3 levels of relative wetness, and 3 levels of hillshade for a total of 5 x 3 x 3 = 45 classes. Choosing the total number of classes is partly a function of available sampling effort, and the desired number of locations that each class should be sampled. At least one sample per zipcode should be included in the sample for a full representation of environments in the biodiversity assessment. The total number of classes can change as data are collected. As data are collected, variables showing a stronger correlation with compositional variation should be more finely divided. Initially, however, a reasonable stating point must be determined.

This process can be made easier by considering the number of variables in the classification, the number of levels in each variable, and the number of sites that would be needed if the diversity of each class were sampled at an equal number of locations. In our GSMNP classification, three variables were used. If three levels were assigned to each variable, then there would be 3^3 = 27 classes. Assuming 3 survey locations per class, a balanced sample would require 81 separate survey sites. If four levels per variable were used, then the total number of classes would be 4^3 = 64. Surveying 3 sites per sample, this classification would require 192 sites. Of course, the total number of sites constituting a reasonable biodiversity assessment will depend of on sampling effort and time required to survey each site. For a given set of explanatory variables, however, increasing the number of levels per variable results in geometric growth of the number of classes in the final ecological zipcode classification. This property should place reasonable limits on the total number of classes for any given study.

Once the approximate number of classes is determined, the number of levels for each variable must be determined. While a balanced design may be appropriate when nothing is known about the relationships between species distributions and the selected environmental variables, it makes sense to employ the same *a priori* ecological hypotheses we used in selecting the environmental variables to select the number of classes for each variables. The environmental variable with the strongest hypothesized correlation with variation in species composition should receive the greatest number of levels in the classification. If species composition varies more strongly along these gradients than others, finer divisions of these variables will result in more obvious differences in species composition among the categorical classification than if the extra levels were evenly distributed among the explanatory variables. Variables hypothesized to have weaker correlations with compositional variation should receive fewer classes. For our implementation of the ecological zipcodes, we hypothesized that elevation had the strongest impact on species distributions. We assigned it five classes. Hillshade and relative wetness both were split into three classes. The total number of classes was 45.

Once the number of levels for each variable has been selected, we must choose the how to partition these gradients into classes. For continuous data we must consider equal-area versus equal-interval methods. The former splits the data into classes that all occupy the same total area in the landscape. This results in intervals of varying ranges in the classified data. Equal interval classification on the other hand splits the dataset along ranges of equal magnitude regardless of the occupancy of those intervals on the landscape. Figure 2.2 illustrates the difference in classification between these two approaches for relative wetness in GSMNP. It is important to note that as the number of categories for a variable increases, equal area classifications approach equal interval classifications.

The choice of which gradient partitioning method to use is largely dependent on the purpose of the sample. If the study requires a relatively balanced population of zipcodes within the study area, then equal area classification is most appropriate. With variables having a skewed distribution of values, however, equal area classifications tend to produce categories that span a very small range of the variable (Figure 2.2a). Equal area classifications also tend to miss rare environmental combinations, which are demonstrably important in characterizing the environmental heterogeneity of (and therefore the biodiversity component of) a landscape, as rare situations often harbor a disproportionate amount of the overall richness of the landscape. Equal interval classifications, on the other hand, produce categories of equal range regardless of their abundance on the landscape. This can result in some classes that cover very little area. This problem is multiplied in the final ecological zipcodes classification where multiple variables, each with very a rare category, are intersected. The abundance of rare zipcodes may be so small that sampling them in the field is impossible (Figure 2.2b).

To balance the objectives of having both realistic variable ranges and a large relative abundance of even the rarest zipcodes, we opted for a compromise between equal area and equal interval classifications. We chose a classification scheme that was equal interval, but the interval choice was based on a data set for each variable in which 2% of the most extreme values were removed. This classification method tended to produce better “high-middle-low” classifications in which the mode of the data was captured mainly in the middle category, and the high and low tails were captured in the other two categories.

The above methods assume that little is known about the functional form of the relationship between the species distributions and the variable used in the zipcodes. As data are collected and the classification is improved, the functional relationships between one or more explanatory variables and species composition may become apparent (Urban *et al.* 2002). In these cases, it makes sense to transform the environmental variables into a compositional turnover surface based on a regression of species composition on the environmental variable. Then, equal interval classification of the new species surface would be most appropriate for sampling regimes, because each interval would represent some linear amount of species turnover.

Creating the ecological zipcodes

The result of variable selection and factorization is a list of nominal variables ordered by the expected strength of their correlation to species distributions. From this set of variables, the ecological zipcodes can be created (Figure 2.3). The basic form of the ecological zipcodes is an *n*-digit number where *n* is the number of explanatory variables. The ecological zipcodes are created through a simple algebraic procedure. For a given location the ecological zipcode (*z*) is calculated as

$$z = \sum_{i=0}^{n-1} 10^i x \qquad \{x | x \in \mathrm{Z}, 0 \leq x < 10\} \tag{1}$$

where *n* is the number of zipcode variables, and *x* is the vector of variable values. Though variable order does not matter for actual creation of the zipcodes, there are qualitative reasons to sort the zipcode variables such the variable with the strongest hypothesized effect is given the highest place value and the weakest variable is given the lowest place value. When the zipcodes are mapped, this variable of greatest value will form the dominate gradient an integral coloring scheme followed by the variable of second greatest place value and so on. The order we chose for our classification was elevation, TCI, and hillshade. For example, zipcode 511 indicates a high elevation (score of 5 for elevation), dry (1 for TCI) site of low insolation (1 for hillshade), while 533 would indicate a high elevation wet site of high insolation.

The ecological zipcodes approach to landscape classification has some distinct advantages for interpretation, analysis, and visualization. First, having a multi-digit classification allows users to easily interpret the classification in multivariate space. For the ecological zipcodes created for GSMNP, a zipcode 511 is immediately identifiable as a high elevation, dry site, with northern exposure. From a practical standpoint, having the variables ordered into a multi-digit zipcode according to the expected strength of their correlation with biodiversity allows more intuitive display of the ecological zipcodes. If the zipcodes are displayed as unique values on an interval scale, the left-most digit dominates the visual pattern of the zipcode raster. The second digit is the second-most visible, and so on. Thus, the zipcodes can be displayed with the relative importance of each variable visible.

Using the ecological zipcodes for biodiversity assessment and monitoring

Once the ecological zipcodes are created, the result is a landscape partitioned into discrete environments that vary across a set of gradients hypothesized to correlate with variation in species composition. The next step in developing a biodiversity assessment is to select a series of sites among these environments to survey. Depending on the focus of the assessment, all of the environments laid out by the ecological zipcodes could be sampled, or a set of environments that are of particular interest in the study. In any case, a scheme must be created for allocating sample effort to locations that correspond to these various environments.

Designing a sample for assessing biodiversity requires balancing two, sometimes conflicting goals. First, the sample must capture a wide variety of environments. That is the primary function of the ecological zipcodes. The second goal is to design a sample that maximizes sampling efficiency. Space is correlated with many factors that control species distributions (Legendre 1993). Regardless of all external environmental factors, samples close together are likely to be more similar than samples far apart (Nekola & White 1999). Dispersal history, in particular, covaries with environmental variables across space, and disentangling the effects of dispersal history from environment is particularly important. With such a profound effect on species distributions, it could be suggested that some distance component be added to the zipcodes. Unfortunately, the effect of distance can only be measured relative to at least two points. It is impossible to assign a single distance value to an ecological zipcode. Geospatial location values could be added to the ecological zipcodes. These measures tend to capture spatial trends as opposed to the structure imposed by spatial autocorrelation (Urban *et al.* 2002). The landscape could be divided into a small number of categories that have similar geographic locations. Watersheds would be the logical choice, because they represent one of the larger repeated units of landscapes. This watershed variable could be included as a digit in the ecological zipcodes. However, the spatial heterogeneity at smaller scales that is independent of the variation in the other variables in the ecological zipcodes is absent. Only very large-scale trends could be seen using watersheds. A solution to this problem might be to increase the number of spatial categories in the ecological zipcodes. Each additional level in this spatially explicit variable greatly increases the total number of zipcodes. Relatively small increases in the number of spatial categories can cause the number of zipcodes to grow quickly beyond practicality or usefulness.

The alternative to incorporating spatial autocorrelation in the ecological zipcodes is to include it in the sample design itself. By designing samples with multiple scales of spatial aggregation, we can disentangle the effects of the ecological zipcodes variables from other, unmeasured factors that are spatially autocorrelated. This is accomplished by using a clustered sampling design which tends to more faithfully capture the underlying spatial autocorrelation of both environment and species (Urban *et al.* 2002;Tobin 2004). Creating clusters of sites can be done randomly, but also could be done to maximize environmental representation by stratifying both within and between clusters by ecological zipcodes. Within-cluster stratification allows the separation of measured effects (the ecological zipcodes) from unmeasured, yet spatially autocorrelated effects.

In addition to designing a survey that represents a variety of environments using the ecological zipcodes in a way that controls for spatial autocorrelation, a good biodiversity assessment should allow robust statistical testing of collected data. Statistical tests will be required to determine if the survey is actually capturing biodiversity at a rate faster than random in the Park. Statistical tests will also be used to determine stopping rules for the assessment; to know when an inordinate amount of sampling effort will be necessary to observe the next new species. Finally, tests will be needed to determine which environmental variables are significantly correlated with changes in species composition as well as determine what new variables should be added into the ecological zipcodes. Nested ANOVA is the typical parametric framework for hypothesis testing in clustered designs. Since the joint inclusion probability of each pair of zipcodes is not fixed within a cluster, parametric methods are not appropriate. Many randomization methods are available, however, since computers can enumerate many if not all of the possible outcomes of any given sample design. This process is made somewhat easier and more robust in the case of the ecological zipcodes since the joint inclusion probabilities of zipcodes within samples can be calculated directly. These inclusion probabilities can be used in hypothesis tests of correlations between the variation in environment and observed variation in species composition.

The final consideration for biodiversity survey design using the ecological zipcodes is sampling efficiency. Though this requirement is often overlooked when designing samples to fit a certain statistical test, it is of primary importance to researchers in the field. No survey design will be successful if plots are placed at random in a large area such as GSMNP. Even without the added benefit of controlling for spatial autocorrelation, clustering of plots is beneficial from a data-collection perspective. Having three plots close together allows more efficient data collection than three plots randomly spaced in the landscape. Maximum cluster size, therefore, should be controlled not only by the spatial autocorrelation structure of the environmental variables of interest, but also by the ability of the researcher to sample all plots in a single day.

The benefit of clustered samples, however, also depends on the accessibility of the clusters. A measure of landscape accessibility to humans, as measured by the energetic-cost associated with round-trip hiking to the survey sites, allows the majority of survey units to be selected in such a way as to minimize the time and money required to collect data. For instance, using a model of accessibility (Chapter 3) samples were selected for GSMNP for the purpose of permanently monitoring vegetation stratified by the ecological zipcodes; all within the 10% most accessible sites (Chapter 4). Obviously, some bias may be introduced by limiting the possible sample locations to only the most accessible sites, so survey designs may incorporate one or more survey units that test for the effect of accessibility on species composition. Accessibility could actually be included in the ecological zipcodes as an explanatory variable in cases where human disturbances that are correlated with accessibility have a profound effect on species composition

In addition to the original purpose of biodiversity monitoring, samples generated from the ecological zipcodes can be implemented for specific studies involving additional variables in which a range of environments need to be sampled within each factor. Having a predefined, multivariate, categorical value assigned to each survey unit broadens the possibilities for meta-analysis and future applications of the original sample. For instance, studies contrasting species distributions in logged versus un-logged areas may use the original zipcodes sample to capture similar environments in logged and un-logged areas. A subset of the original data, with the specific environments of interests would be included. The great advantage of this approach would be that environmental contrasts important for species distributions would be already built into the data. Similarly, direct contrasts employing the zipcodes could be designed for additional categories of interest. For instance, we can sample the same zipcodes: in limestone and sandstone areas for a direct contrast of bedrock influence; in contrasting geographic areas (in GSMNP, for example, this might involve the eastern versus western part of the Park or the generally north-facing Tennessee side and the generally south-facing North Carolina side; and in small and isolated patches versus large and contiguous patches of the same zipcode types.

Evaluation and Extension

After a sample is selected and the initial field work is conducted, the relationship between species distributions and the ecological zipcodes should be evaluated and improved. How well did variation in species composition correlate with the variables of the ecological zipcodes? Do environmental data collected on-site during fieldwork correlate more strongly with species distributions than do the ecological zipcodes? The answer to these questions will vary among species, but it is important to remember that the ecological zipcodes and their corresponding sampling design were meant to capture a wide range of environments, regardless of species identity or taxonomic group. Water, temperature and insolation data could also be collected to assess the relationship between the ecological variables important for species composition and the surrogate measures used to in the ecological zipcodes.

If a particular collected environmental variable shows stronger correlations with richness than the original variables in the ecological zipcodes, then that variable could be added to the classification. The variable selection and categorization procedure can be repeated for this variable, and it can be added as a digit in the zipcodes. If the variable is not measured at the grain and extents necessary for incorporation into the zipcodes, then a surrogate measure may be necessary. Alternatively, the variable may be sampled at its own characteristics scales and modeled across the landscape (e.g. fine-scale temperature models; Lookingbill & Urban 2003). Researchers could plan for the addition of variables in the initial zipcode creation and stratify across fewer categories in the first sample. Additional plots could be added to the sample the following field season stratified by the newly incorporated variable. As an example, Urban *et al.* (2002) outline a multi-stage sampling improvement for modeling vegetation patterns that could be applied to biodiversity patterns in general.

While there are many species correlated with the ecological zipcodes, there may also be a few interesting habitats or species that are not. Further, a species of ecological or conservation significance may be conspicuously underrepresented in the dataset. In these cases, the original sample should serve as a baseline from which to collect species-specific data. New samples can be selected based upon continuous or categorical variables such as soil type or disturbance that are not permanently incorporated into the zipcodes.

In addition to rare or unique species not captured by the sample, the ecological zipcodes could be also used to find the rarest or most extreme environments in the Park, which are likely to be correlated with unique species. Further, extreme and unique environments are also most susceptible to environmental change, because their representation in the landscape is so small. If conditions change to reduce these environments, their demise will occur sooner than environments occupying a larger portion of the landscape. Ecological zipcodes generated via equal interval data splits provide researchers with categorical representations of such environments (Figure 2.2b).

Finally, if zipcodes and other environmental data are used routinely for the labels that accompany specimen collections, we can produce a coarse description of the habitat niche for species observed. The variety of ecological zipcodes for recorded observations would describe the breadth of environmental tolerance for a species and to some extent distinguish species occupying specialized habitats from those occupying a broad range of habitats. This could be extended to species assemblages or communities as well.

Conclusion

The strategy for assessing biodiversity that we have outlined here is applicable to many systems beyond the one for which it was originally developed. Exactly which variables are used in any single implementation of this strategy will change between regions. The procedure, however, of hypothesizing the important ecological variables, categorizing those variables into class that can be sampled, and iteratively evaluating and refining the resulting ecological zipcodes remains the same regardless of region. Given the importance of cataloging biodiversity in the face of global change, we see this strategy as a powerful tool for assessing biodiversity in a great variety of ecosystems.

References

Allen A.P., Brown J.H. & Gillooly J.F. (2002) Global Biodiversity, Biochemical Kinetics, and the Energetic- Equivalence Rule. *Science* 297, 1545-1548

Andrewartha H.G. & Birch L.C. (1954) *The Distribution and Abundance of Animals*. University of Chicago Press, Chicago.

Bratton S.P. Resource Division in an Understory Herb Community: Responses to Temporal and Microtopographic Gradients. American Naturalist 110[974], 679-693. 1976.

Cain S.A. Ecological Studies of the Vegetation of the Great Smoky Mountains of North Carolina and Tennessee. I. Soil Reaction and Plant Distribution. Botanical Gazette 91[1], 22-41. 1931.

Clark J.S., Fastie C., Hurtt G., Jackson S.T., Johnson C., King G.A., Lewis M., Lynch J., Pacala S., Prentice C., Schupp E.W., Webb T. & Wyckoff P. (1998) Reid's Paradox of Rapid Plant Migration - Dispersal Theory and Interpretation of Paleoecological Records. *Bioscience* 48, 13-24

Geiger R. (2003) *The Climate Near the Ground*, 6 edn. Rowman & Littlefield, Lanham, MD.

Hanski I. & Gilpin M. (1991) Metapopulation Dynamics - Brief-History and Conceptual Domain. *Biological Journal of the Linnean Society* 42, 3-16

Holdridge L.R. Determination of World Plant Formations from Simple Climatic Data. Science 105[2727], 367-368. 1947.

Korner C. (1998) A Re-Assessment of High Elevation Treeline Positions and Their Explanation. *Oecologia* 115, 445-459

Legendre P. (1993) Spatial Autocorrelation - Trouble or New Paradigm. *Ecology* 74, 1659-1673

Lookingbill T.R. & Urban D.L. (2003) Spatial Estimation of Air Temperature Differences for Landscape-Scale Studies in Montane Environments. *Agricultural and Forest Meteorology* 114, 141-151

Mendenhall W., Wackerly D. D. & Scheaffer R. L. (1998) *Mathematical statistics with applications*. Duxbury Press, Belmont, CA.

Moore I.D., Grayson R.B. & Ladson A.R. (1991) Digital Terrain Modeling - a Review of Hydrological, Geomorphological, and Biological Applications. *Hydrological Processes* 5, 3-30

Nekola J.C. & White P.S. (1999) The Distance Decay of Similarity in Biogeography and Ecology. *Journal of Biogeography* 26, 867-878

Rathcke B. & Lacey E.P. (1985) Phenological Patterns of Terrestrial Plants. *Annual Review of Ecology and Systematics* 16, 179-214

Ricketts T.H. (2001) The Matrix Matters: Effective Isolation in Fragmented Landscapes. *American Naturalist* 158, 87-99

Root T. (1988) Environmental-Factors Associated With Avian Distributional Boundaries. *Journal of Biogeography* 15, 489-505

Rosenzweig M.L. (1995) *Species Diversity in Space and Time*. Cambridge University Press, Cambridge, UK.

Scott J.M., Csuti B., Jacobi J.D. & Estes J.E. (1987) Species Richness - a Geographic Approach to Protecting Future Biological Diversity. *Bioscience* 37, 782-788

Skellam J.G. Random Dispersal in Theoretical Populations. Biometrika 38[1/2], 196-218. 1951.

Stephenson N.L. (1990) Climatic Control of Vegetation Distribution - the Role of the Water-Balance. *American Naturalist* 135, 649-670

Stephenson N.L. (1998) Actual Evapotranspiration and Deficit: Biologically Meaningful Correlates of Vegetation Distribution Across Spatial Scales. *Journal of Biogeography* 25, 855-870

Thompson S.K. (1991) Adaptive Cluster Sampling - Designs With Primary and Secondary Units. *Biometrics* 47, 1103-1115

Tobin P.C. (2004) Estimation of the Spatial Autocorrelation Function: Consequences of Sampling Dynamic Populations in Space and Time. *Ecography* 27, 767-775

Turner M.G. (1989) Landscape Ecology - the Effect of Pattern on Process. *Annual Review of Ecology and Systematics* 20, 171-197

Urban D., Goslee S., Pierce K. & Lookingbill T. (2002) Extending Community Ecology to Landscapes. *Ecoscience* 9, 200-212

West G.B., Brown J.H. & Enquist B.J. (2001) A General Model for Ontogenetic Growth. *Nature* 413, 628-631

Whittaker R.H. Vegetation of the Great Smoky Mountains. Ecological Monographs 26[1], 1-80. 1956.

Wolock D.M. & Mccabe G.J. (1995) Comparison of Single and Multiple Flow Direction Algorithms for Computing Topographic Parameters in Topmodel.

Water Resources Research 31, 1315-1324

Yeakley J.A., Swank W.T., Swift L.W., Hornberger G.M. & Shugart H.H. (1998) Soil Moisture Gradients and Controls on a Southern Appalachian Hillslope From Drought Through Recharge. *Hydrology and Earth System Sciences* 2, 41-49

Figures

Figure 2.1 The relationships between observed environmental variables used in the ATBI "ecological zipcodes" and their latent ecological counterparts. Environmental variables that have a more direct relationship with species distributions (the right hand side of the figure), are often unavailable at the resolution or extent necessary for stratifying samples. Data that are readily available at the appropriate scale may predict the more ecologically meaningful variables, but the exact form and strength of the relationship may be unknown.

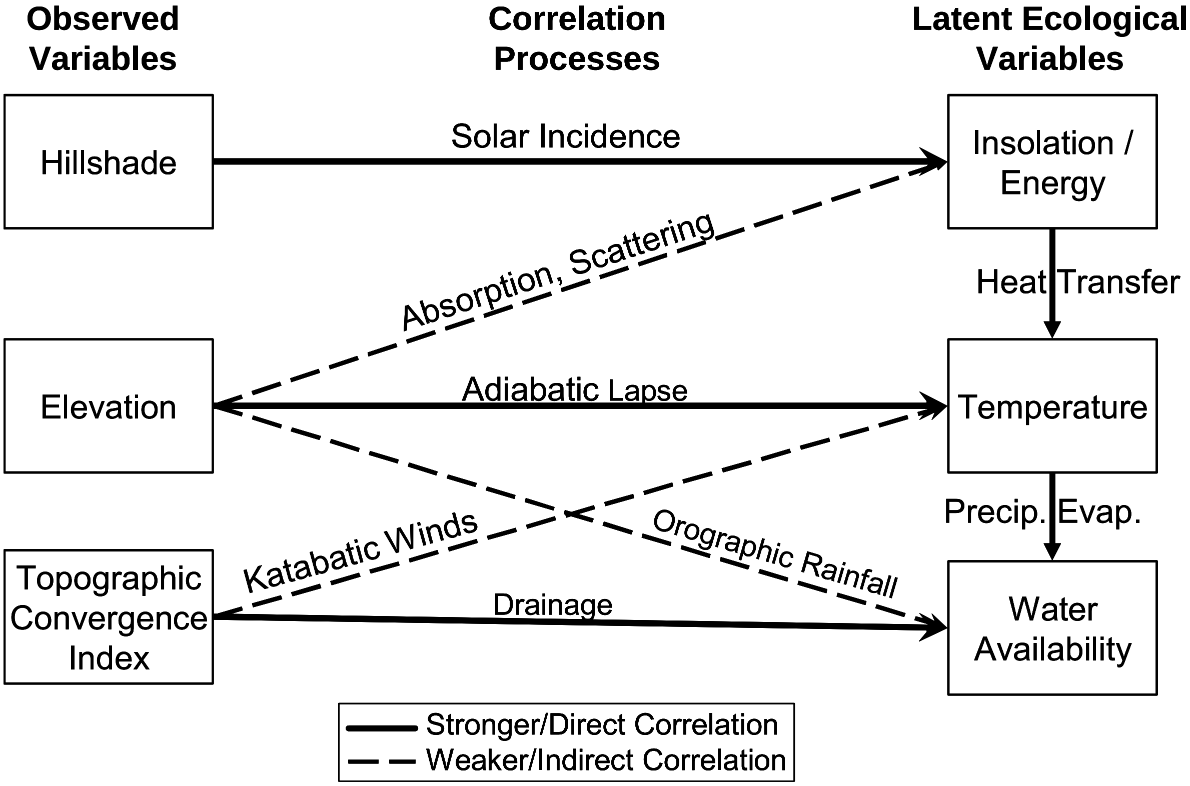


Figure 2.2 Relative abundance of ecological zipcodes using various classification algorithms. a) Histogram of relative wetness in GSMNP (as measured by the topographic convergence index) highlighting the vast difference in the categorical frequency distribution for different break methods. b) The ranked abundance distribution of GSMNP zipcodes created using different break methods. The equal interval method based on data with 2% outliers removed provides a good balance between true categorical differences and rare zipcode abundance.

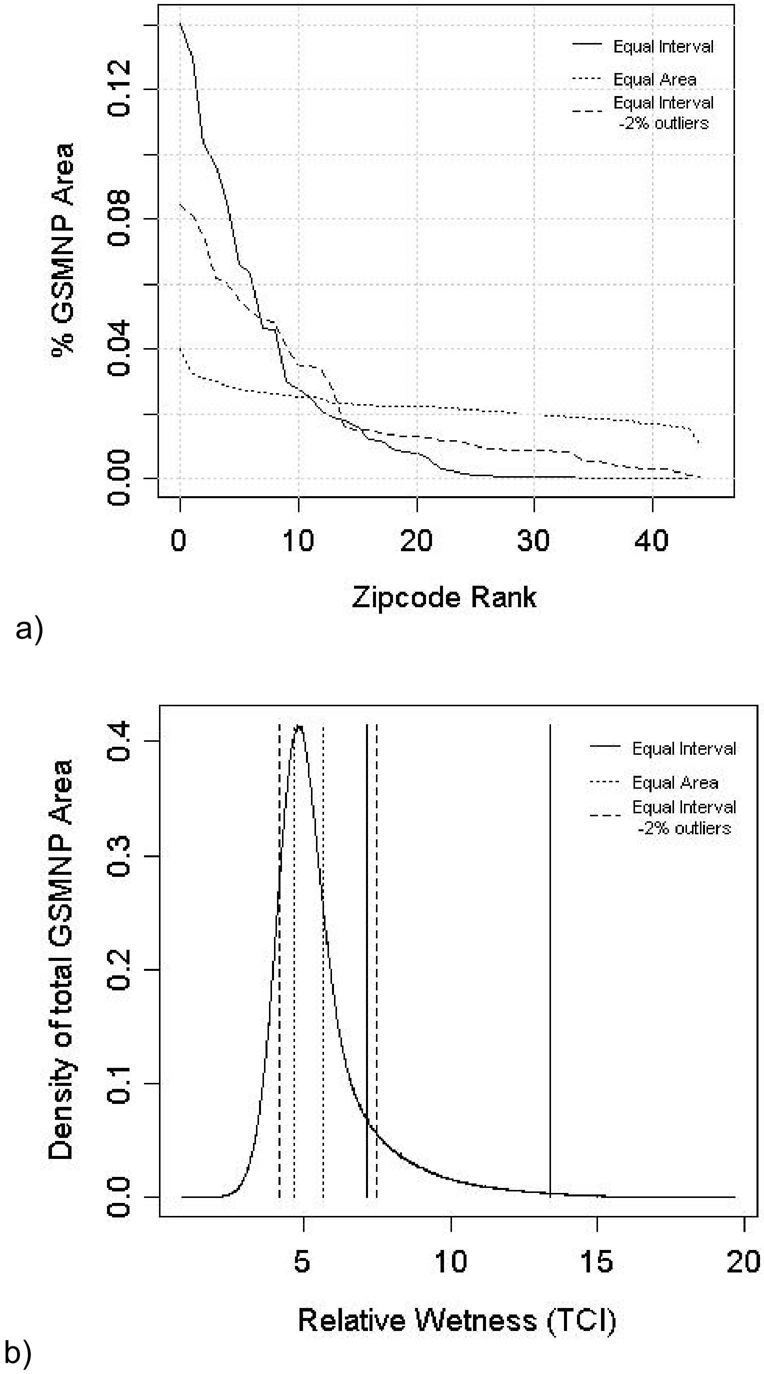
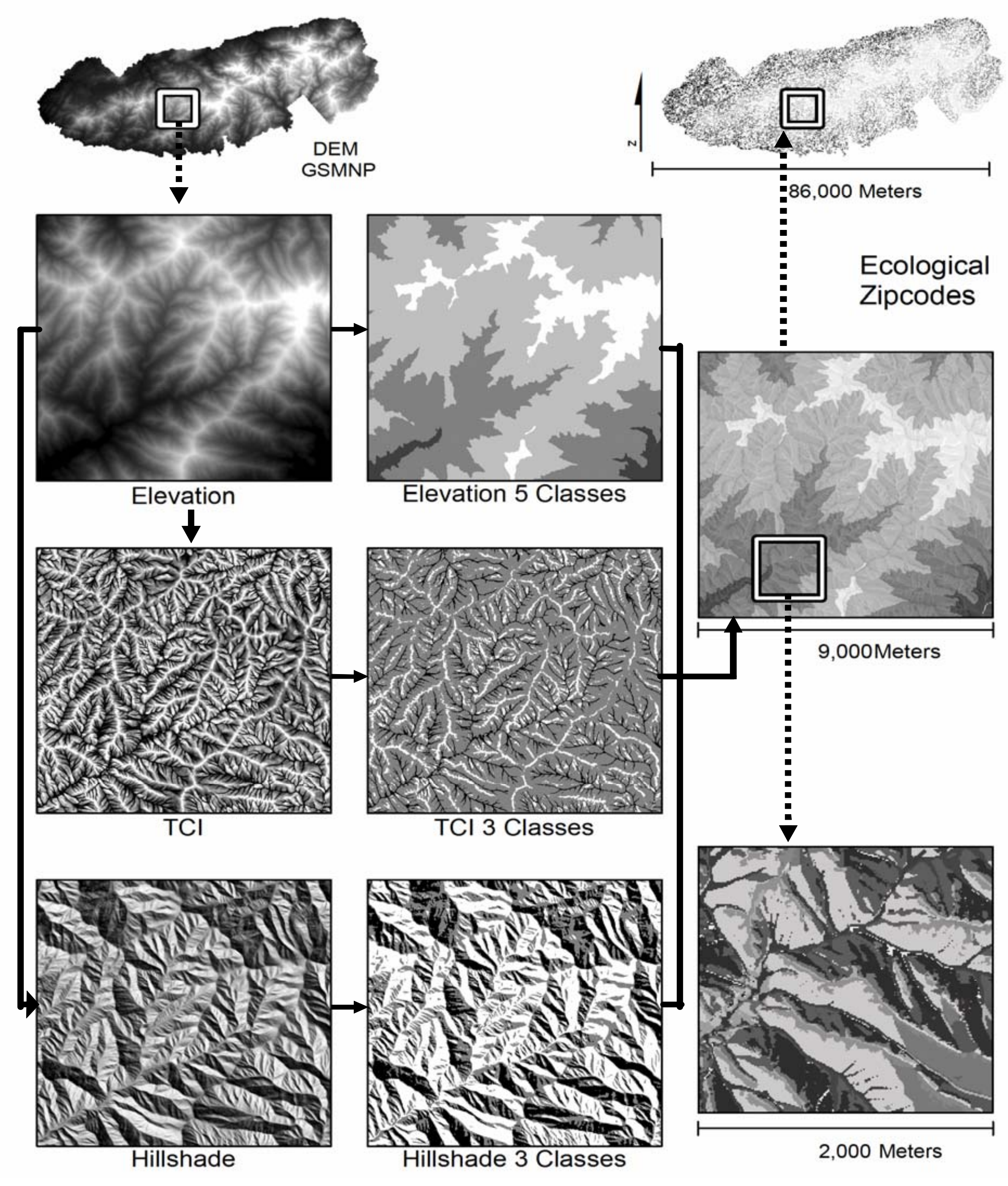


Figure 2.3 Workflow for creating ecological zipcodes. For our study, all zipcode variables were derived from a single DEM. They were classified and then combined into a single raster in which each variable represented 1 digit in the zipcode. The classifications shown here used area equal intervals based upon the original variable with 2% of outliers removed.



Tables

Table 2.1 A list of digitally available variables that influence species distributions in GSMNP along with the data form in which they are available. The strength of correlation between these variables and biodiversity is probably unknown prior to sampling. Further, data for many of these variables is unavailable at a small grain over the entire park. In order to select variables for inclusion in the ecological zipcodes, they must have a hypothesized correlation with species distributions, and must be available at the appropriate resolution over the entire study area. The DAYMET data are modelled climatic variables available at a 1km^2 resolution for the entire US (http://www.daymet.org).

		Available Data		
Variable Type	**Variable**	**Resolution**	**Extent**	**Data Source**
Direct				
Energy Input	Temperature	1km^2	park	modelled DAYMET
	Insolation	10m^2	park	DEM derived
Water Availability	PET	N/A	N/A	N/A
	Precipitation	1km^2	park	modelled DAYMET
	Soil Moisture	points	sparse	direct sample
Nutrient Availability	Soil Texture	points	sparse	direct sample
	Soil Depth	points	sparse	direct sample
	Cations	points	sparse	direct sample
	N-P	points	sparse	direct sample
	pH	points	sparse	direct sample
Vegetation	Occurrence	1m^2-100m^2	sparse	direct sample
	Abundance	1m^2-100m^2	sparse	direct sample
Indirect				
Landscape Pattern	Human Accessibility	10m^2	park	modelled (Jobe in prep.)
	Dominant Vegetation	polygons	park	modelled from area photos
Topographic Position	Elevation	10m^2	park	digital elevation model (DEM)
	Aspect	10m^2	park	DEM derived
	Slope	10m^2	park	DEM derived
	Distance from Stream	10m^2	park	DEM derived
Disturbance History	Fire	polygons	park	historic maps, written records
	Logging	polygons	park	historic maps, ownership records
	Gaps	points	sparse	direct sample
Climate Change	Glaciation	N/A	N/A	N/A
	Global Warming	1km^2	entire park	modelled

Table 2.2 The matrix of correlation coefficients for each pair of zipcode variables. The low correlation coefficients suggest a degree of independence among the layers, though they are all derived from the same data source (a DEM).

Correlation Matrix

Layer	Elevation	TCI	Hillshade
Elevation	1.000	-0.058	-0.051
TCI	-0.058	1.000	0.083
Hillshade	-0.051	0.083	1.000

Chapter 3

Calculating human accessibility of conservation landscapes

Abstract

I present a model that estimates the accessibility of a landscape to people walking through that landscape. This model considers the ways that distance, slope, stream crossings, and vegetation influence the energetic cost of walking. I outline the development of this accessibility model, and then apply it to the question of sampling bias in vegetation surveys of Great Smoky Mountains National Park. I also assess the correlations between vegetation communities in the Park, and how collected data may be affected by those interactions. Over 1100 vegetation surveys were consolidated spanning a 40-year period of research in the Park. These data show a strong positive correlation with accessibility. Important vegetation communities in the Park are under-sampled relative to their abundance because they are less accessible. These less accessible communities may have a distinctive species composition because of their distance from sampled sites as well as decreased human disturbance. I discuss the myriad ecological questions this accessibility model can be applied to, including the effects of human disturbance as well as conservation planning, management, and restoration.

Keywords

Human accessibility, sampling bias, disturbance ecology, conservation planning, Great Smoky Mountains National Park

Introduction

Roads, trails, and landscape roughness influence how humans interact with landscapes. At broad scales, this is intuitive. Many of the remaining undisturbed areas of the world are in its most inaccessible locations (Scott *et al.* 2001). As more roads are built, these areas become more accessible. The frequency and intensity of human disturbance then increases. This process is evident in regions such as the Amazon (Nepstad *et al.* 2001;Maki *et al.* 2001;Laurance *et al.* 2004). The impacts of decreasing friction to human access are evident not only at these regional scales but also at landscape and local scales. In general, human disturbance decreases as areas become remote from points of access, although narrow gauge railroads in my own study area (Great Smoky Mountains National Park), allowed logging on steep, high elevation, and remote sites. In most national parks, the density of visitors drops dramatically at minimal distances from the main roads. Even biologists rarely stray far from roads to conduct surveys and tend to avoid areas that are difficult to get to, since survey resources are limited and time used for access directly decreases the time available for actual survey.

Despite the obvious importance of the interaction between landscape structure and human accessibility, it has been measured only crudely in science. Distance from road is the typical measure used as a surrogate for human accessibility of a landscape. The influences of topography, vegetation, and water features are ignored in this measure. Yet, even with this crude measure, studies have shown that increased distance from road results in increased frequency of native grasses (Gelbard & Harrison 2003) and changes in movement behavior for a wide variety of animals (e.g. Brody & Pelton 1989;Mclellan & Shackleton 1988;Grover & Thompson 1986;Whittington *et al.* 2005;Rost & Bailey 1979). Trombulak & Frissell (2000) provide a thorough review of the impacts that road building has on natural ecosystems.

The impacts on natural systems are only one of the areas in which the interaction between landscape structure and human accessibility are important. Since most biological surveys are conducted close to roads, there may be a significant sampling bias in much of the data we now use for inquiry. Also, the patterns of human accessibility on a landscape may play an important role in deciding how to fight fires or control the spread of exotic species. Since they are less prone to human disturbance, areas that are more inaccessible are also likely candidates for the reintroduction of extirpated species. In short, landscape accessibility plays a vital role in our understanding of a wide variety of biological, ecological, and conservation topics. Some of these correlations are strong enough to be obvious with even relatively coarse measures, such as distance from road. Other correlations may be subtle, but heretofore we have not had measures of accessibility sensitive enough to capture them. As a consequence, there is a need for measures of landscape accessibility that are more realistic and accurate than coarse distance-from-road measure that has been used in the past.

Here, I develop such a model of accessibility that estimates the energetic-cost associated with hiking along least-cost path from the nearest road to any point on a landscape. This model incorporates not only surface distance, but also the isotropic friction associated with landscape features such as trails, vegetation, and streams.

Further, I include an anisotropic factor that estimates the energetic cost associated with hiking along slopes of different gradients. This model provides a far more realistic and accurate picture of landscape accessibility than mere Euclidean distance. I apply this model to a particular issue that receives little attention: the sampling bias of biological surveys for accessibility. I apply the energetic-cost model to the Great Smoky Mountains National Park (GSMNP TN and NC, US), and analyze the correlations between accessibility and vegetation surveys that have been conducted in the Park over the last 40 years. I show how particular plant communities have been sampled in the Park and how both the relative abundance of the communities and the distribution of accessibility among communities have interacted to generate a sampling bias in vegetation samples that was previously undocumented. I conclude with an observation that there are many studies where a positive correlation between accessibility and sample location is present, and I discuss how future studies should use this accessibility model to develop less costly and unbiased surveys. I also discuss the myriad applications of the energetic-cost of hiking model.

Methods

Study area

The accessibility model I developed for this study encompasses the area of GSMNP. GSMNP is part of the Southern Appalachian Mountains and comprises ~2×10^5 ha. The Park has a main ridge running ENE to WSW that is followed closely by the Appalachian Trail. Elevation varies from 184m to 2029m within the Park.

The average slope in the Park is 25°. There are 845km of roads within and surrounding the Park with a main road (Newfound Gap Road) running N-S through the center of the Park (Figure 3.1). There are 1295km of hiking trails in Park.

GSMNP is rich in endemic plants and amphibians. There are at least 129 trees native to GSMNP. The Park contains 56 different vegetation formations following the U.S. National Vegetation Classification (NatureServe 2006). The most abundant vegetation communities are part of the broad categories of *Quercus rubra-Carya glabra* forests, cove forests dominated by *Acer saccharum* and *Aesculus flava*, *Tsuga canadensis* forest, Northern Hardwood forests dominated by *Betula alleghaniensis* and *Fagus grandifolia*, and *Picea rubens-Abies fraseri* forest above 1400m.

One of the most significant physiognomic features of GSMNP, even to a casual observer, is the distribution of ericaceous shrubs of the genera *Rhododendron* and *Kalmia*. Though their abundance varies, approximately 24% of the Park can be considered to have moderate or heavy dominance of these taxa. These thickly tangled, evergreen stands sometimes referred to as “hells” are extremely difficult to travel through, and represent a significant barrier to the accessibility of interior regions of the Park.

Data

The data required for the development of the accessibility model were a series of GIS layers available for the Park. These include a digital elevation model (DEM, resolution $10m^2$) and vector data for roads, trails and streams. I have also used the overstory vegetation polygon layer developed by the University of Georgia Center for

Mapping and Remote Sensing (Center for Remote Sensing and Mapping Science (CRMS) 2004). This map classifies GSMNP into 172 vegetation types based on the interpretation of aerial photographs. The classification used by CRMS can be mapped directly onto the more familiar U.S. National Vegetation Classification whose identifiers are known as “CEGL”-codes. It is this classification that I have used in this analysis. For mapping the distributions of *Rhododendron* spp. (*R. maximum, R. catawbiense, R. minus,* and *R. carolinianum*) and *Kalmia latifolia*, I have used an understory map developed by CRMS that records the dominant evergreen species present in the understory. This classification gave a low, medium, and high density to the distribution of these two genera based on cover. Low densities had <50% cover of these species, medium had 50-80% and heavy had more than 80% cover. For use in the model, I considered low densities to be equal in difficulty to the surrounding vegetation, while medium and high densities were assigned greater energetic costs.

The vegetation survey data used in this study were obtained from a recently compiled database of the vast majority of vegetation surveys conducted in GSMNP beginning in 1972 and continuing until 2004. The data are composed of 1104 individual survey units. The geolocations of each survey unit in these data are known within ~100m. The data are compiled across many studies, whose target communities varied widely from rich cove hardwood communities, to grassy balds, to spruce-fir forest, to rocky outcrops.

Making the energetic-cost model

The energetic-cost model is based upon an accumulative least-cost path algorithm. The model is computed on a lattice of equal area cells. Resolution for this model was $10m^2$, and the extent was the entire extent of GSMNP. This least-cost path algorithm computes the total cost of travel from a focal area (the Park) to a source area: in this case, roads. The algorithm follows the least cost path by starting from the focal cell and successively finding the adjacent cell (in an 8-cell neighborhood) that is least costly. This process continues until a source cell is reached. The energetic-cost assigned to a focal cell is the sum of the individual cell costs along the least cost path.

For a given cell, the energetic cost of traversal is calculated by:

$$Cost = S \times I \times A \tag{1}$$

where *S* is the surface distance, *I* is the product of the isotropic costs, and *A* is the anisotropic cost. The energetic cost is calculated in (J/kg). Surface distance (*S*) is merely the linear distance of travel across the surface of the earth, and is dependant on the slope, resolution of the raster and the direction of travel across the grid. Isotropic costs (*I*) are those energetic costs that do not change with direction of travel. These include the friction associated with the trail surface, the vegetation, and any water features that are present. Anisotropic costs (*A*) are those costs that are dependant upon the direction of travel. For the energetic-cost model, the only anisotropic cost was that of slope. The effect of slope on energetic cost is anisotropic because the perceived slope of a surface changes depending on the direction from which it is approached. For instance, the slope of a steep hill is 0 when traveling perpendicular to the hill face, is positive when traveling up the hill, and negative when traveling down it. I now explain in detail the calculation of the both the isotropic and anisotropic costs.

Isotropic costs

Thee landscape classes isotropically influence the energetic cost of hiking: trails, vegetation, and streams. There is some history of calculating the increase in energetic costs associated with different terrains, especially in military applications (e.g. Soule and Goldman 1972). Imhof (1950) suggested that the effect of walking on- versus off-trail reduces speed by 60%. In a more detailed study, Soule and Goldman (Soule & Goldman 1972) report coefficients for the energetic effects of walking on roads and trails, through light and heavy brush, and through swamps. I have used these coefficients to train the isotropic factors in this energetic-cost model (Table 3.1). Unfortunately, Soule and Goldman do not explicitly define what is meant by light brush and heavy brush. It is reasonable to assume, though, that any off-trail walking in GSMNP could be considered light brush in the context of Soule and Goldman (1972). I considered the effect of a medium to heavy density of *Rhododendron* and *Kalmia* to be equivalent to heavy brush in the context of Soule and Goldman (1972).

The impact of streams on accessibility is not as straight-forward to calculate as those for vegetation and trail vs. off trail walking. High elevation, 1st order streams are rarely difficult to cross, if they are flowing at all, while lower-elevation streams can be virtually impossible. Taking this into account, I considered the energetic expenditure associated with crossing a stream to be a function of stream discharge. The more water flowing in a stream, the more difficult it is to cross. Stream discharge is controlled two factors: the stream area and the stream velocity. Since neither of these factors can be calculated directly for all streams in the Park, I used surrogate measures for each. I assumed stream area to be proportional to the logarithm of accumulated flow, which is the merely the upslope area that contributes to the stream. I assumed stream velocity to vary with the tangent of slope. Steeper slopes have faster streams. Using these surrogates, I calculated the proportional increase in energetic cost associated with stream crossing (*C*$_{st}$) to be:

$$C_{St} = a\ln(F)\tan(S) + b \tag{2}$$

where *F* is the accumulated flow of the stream, *S* is the slope, *b* is the baseline friction associated with crossing a stream of slope 0 as specified for swamps in Soule and Goldman (1972), and *a* is fitted parameter used to standardize the minimum stream crossing friction to be equal to the friction of walking on a trail. The final isotropic cost grid was the intersection of the weights for trails, the weights for vegetation and the calculated stream crossing cost.

Anisotropic cost: Slope

The only anisotropic energetic cost in the model was that of walking on a slope. Within limits it is less energetically costly to walk down slopes than to walk up slopes. The energetic cost associated with traversing a slope is also dependant upon the direction from which the slope is approached. So, a location does not have a single energetic cost. Tobler (1993) was the first to use an anisotropic function for calculating the cost of slope walking in a GIS framework. He employed the “hiking function” of Imhof (1950). A few anthropological studies have used similar functions to calculate the probability of interactions between tribes (e.g. Van Leusen 2002;

Hare 2004). The Imhof “hiking function” is calculated in terms of velocity (km/hr). There are a few problems associated with the Imhof “hiking-function”. It is symmetric so that walking speeds are identical for gradients of equal deviation from the -1° of maximal speed. It is also unknown how additional factors, such as vegetation affect walking speed.

Instead, of using the Imhof “hiking function” to calculate the cost of walking on slopes, I have synthesized recent work in biomechanics to generate a new hiking function based on energetic cost rather than velocity. There have been a variety of studies calculating the impact of slope on the energetic expenditure of walking and running (see Rose Jessica *et al.* 1994 for a comprehensive treatment). Only a few, however, focus on the energetic cost of walking on relatively steep terrain (e.g. Minetti *et al.* 1993;Minetti *et al.* 1994;Minetti 1995;Minetti *et al.* 2001;Minetti *et al.* 2002). Minetti and other (2002) determined the energetic cost associated with walking up and down slopes ranging from -24° to +24°. They found a slope of -6° to have the least energetic cost of walking. At slopes greater than this, costs rise sharply. Costs rise slowly at slopes less than -6°. To estimate energetic cost from these data, I fit a 2nd order polynomial to data of Minetti *et. al.* (2002) with slopes greater than -6° and a 2nd order polynomial to data with slopes less than -6°:

$$W = \begin{cases} 18.827S^2 + 3.766S + 1.240 & -60 < S < -6 \\ 46.869S^2 + 9.374S + 2.382 & -6 < S < 60 \end{cases} \qquad (3)$$

where *W* is the energetic cost (J/kgm) and *S* is the local slope in degrees. Some extrapolation from the maximum was necessary since slopes in the Park are often greater than 24° and people can traverse steeper slopes than these. Energetic costs were estimated for a maximum slope of 60°. Slopes steeper than 60° were considered impossible to cross.

The functions that I derived from Minetti *et al.* (2002), consider the cost of traversing a slope in one direction only. It is not merely the one-way energetic costs that must be considered when modeling accessibility, however. The energetic cost for leaving a trail and walking down the side of a mountain might be low, but the return trip back up the mountain would have a great energetic cost. I wanted to estimate the round-trip energetic costs to assess the accessibility of the GSMNP landscape. To incorporate the energetic cost of a round-trip path to a particular location, I considered the least cost path to be the one in which energetic cost of both the uphill and downhill legs of the trip were minimal. In other words, the least cost path considered not only the energetic cost of going to a location, but also the cost associated with going back to the starting point from that location along the same path.

Analysis of Sampling Bias in Vegetation Surveys

The path distances that are derived from the model are the estimated round-trip energetic expenditure in J/kg. To test for a correlation between accessibility and sample location in the vegetation data, accessibility estimates at the geographic location for each vegetation sample unit were derived from the model. Since the geographic locations of vegetation samples were known to within only 50m, the energetic cost estimate I used was the mean of all energetic costs within 50m of the sample.

I divided the range of round-trip energetic costs into 50 equal interval bins, each with a range of 3090 J/kg. This allowed me to compare the observed distribution of both plots and vegetation communities to their expected distributions if they were uncorrelated with accessibility. I calculated the expected and observed abundance of plots across accessibility to highlight any correlations between the location selected for plots and accessibility. Second, I calculated the observed and expected number of plots in each vegetation community. This highlighted those communities that are chronically under-sampled or over-sampled relative to their abundance in the Park. Third, I calculated the expected and observed area of each vegetation community with respect to accessibility to determine those communities that were typically very close to the road or far from it. Finally, I calculated expected and observed number of plots for each community individually by accessibility. This analysis highlighted those communities that have been sampled only in their most accessible locations, while larger, more inaccessible areas of the same community type are under-sampled.

Results

Estimated round-trip energy costs for the Park ranged from 0 J/kg to 1.5×10^5J/kg (Figure 3.1). The distribution of energetic cost in the Park was skewed toward more accessible sites with a mean of 36755 J/kg and a standard deviation of 31886 J/kg. The estimated energetic costs can be measured in kilocalorie (kcal) expenditures if the weight of the person is known. Based on the model, an average male weighing 75 kg would expend 650 kcal to travel to and from a random site within the Park and would expend 2768 kcal to travel to and from the most inaccessible site in the Park along the least cost path. As a means of comparison, a 75 kg male expends about 950 kcal playing a full-court game of basketball.

Due to the skewed distribution of energetic costs, a very low proportion of the Park could be classified as extremely inaccessible. This suggests that access points are over-dispersed and abundant in the Park. The area north of Fontana Lake in the SW part of the Park is the largest contiguous area that is quite inaccessible, followed closely by an area in the NE part of the Park. The area in the NE section of the Park is inaccessible because of the steepness of the landscape and the abundance of *Rhododendron* and *Kalmia* in that area (Figure 3.1). The area north of Fontana Lake is inaccessible because the closest access points are at the east and west ends of the lake. It is possible to access some of the trails that occur north of the lake via boat, but such access was not accommodated in this analysis.

The estimated round-trip energetic expenditure to vegetation plots ranged from 6 J/kg to 1.2×10^5 J/kg, with a mean energetic expenditure of 2.7×10^4 J/kg. As expected, the distribution of sample units in the Park is not a random sample of accessibilities in the Park (Figure 3.2). 14% of all vegetation plots fell within the very first interval of accessibility, which is twice the expected number of plots for that interval. The vast majority of vegetation samples fell within the top 1/3 of the range of accessibility values. 27% of the Park area or > 6.3×10^5ha is under-sampled because of its inaccessibility. All but 16 of the 1104 vegetation plots occurred in the first half of the range of accessibility values. The top 1.2% or ~3,000ha of most inaccessible places in the Park have no vegetation samples at all in them. The few plots that were in relatively inaccessible locations tended to occur in relatively localized areas of inaccessibility, not the larger, more contiguous areas of inaccessibility.

Most vegetation communities that occur in GSMNP are under-sampled relative to their abundance in the Park. Table 3.2 highlights the top 10 over-collected and under-collected communities in GSMNP. The most over-collected community in the Park is the rich Northern Hardwood Forest dominated by *Aesculus flava*, *Betula alleghaniensis*, and *Acer saccharum*. It occupies only 3% of the Park area, yet 7% of the vegetation samples come from this community. The most under-sampled community in the Park is the Appalachian Montane Oak-Hickory Forest dominated by Chestnut oak (*Quercus prinus*). It occupies 11% of the Park area, yet only 5% of the vegetation samples occur in this community type. More species-rich communities, tended to be over-sampled while more species-poor communities were under-sampled. For example, Southern Appalachian Cove Forests are very over-sampled in the species rich areas dominated by *Acer saccharum*, but the typic cove forests dominated by *Liriodendron tulipifera* are under-sampled relative to their abundance in the Park. Some communities are under-sampled in the Park, but are well sampled in the Southern Appalachians as a whole (Newall & Peet 1998).

Most communities that were over-sampled in the Park were positively correlated with accessibility. This is especially true for mesic Chestnut Oak Forest, Xeric Pine Woodlands, and Appalachian Montane Alluvial Forests (Figure 3.3a). Appalachian Montane Alluvial Forest dominated by *Platanus occidentalis* and *Liriodendron tulipifera* occurs at lower elevations along rivers. The major rivers in the Park are followed closely by roads at these lower elevations. The Red Spruce-

Fraser Fir Forest was the only over-sampled community that was negatively correlated with accessibility. More area of Red Spruce-Fraser Fir forest occurred in inaccessible locations. Surprisingly, most under-sampled communities were either randomly distributed with respect to accessibility, or were positively correlated with accessibility (Figure 3.3b). Notable exceptions were Southern Appalachian Mixed Hardwood Forest dominated by *Acer rubrum, Nyssa sylvatica, Magnolia fraseri,* and *Oxydendrum arboreum*, which showed a negative correlation with accessibility.

The arrangement of plots within community types with respect to accessibility was varied. Communities that were over-sampled tended to have surveys aggregated in accessible areas, even though a range of accessibilities may have been sampled. Plots located in Red Spruce-Fraser Fir forest were strongly correlated with accessibility. There are large, inaccessible tracts of Red Spruce-Fraser Fir Forest that have no vegetation surveys from our complied data. The same is true of species-rich Northern Hardwood Forests. Also, pure Fraser Fir forest, which is quite rare in the Park now because the Balsam Woolly Adelgid (*Adelges piceae*) has killed most Fraser Fir trees, has an abundance of inaccessible tracts. Though the Fraser Fir forest is over-sampled relative to its abundance in the Park, there are no surveys in these more inaccessible tracts, which seem to have the highest abundance of Fraser Fir. For under-sampled communities, the plots that did exist were located in accessible areas. Plots located in Oak-Hickory Forest dominated by Chestnut oaks showed the strongest correlation with accessible areas. The Southern Appalachian Mixed Hardwood Forest had the greatest area in inaccessible locations without a plot. The plots in the most under-sampled community, Oak-Hickory Forest dominated by Red Oak, were actually among the most well dispersed samples with regard to accessibility.

Discussion

Assessment of the model

The model of human accessibility that I have developed estimates the energetic expenditure necessary to hike round-trip along the least cost path to any location from the nearest road. It incorporates not only the effect of linear distance, but also of walking on slopes and walking across different terrain types including: trails, vegetation, and streams. It also estimates the friction associated with crossing the stands of *Rhododendron* and *Kalmia* that occur throughout the Park. This round-trip energetic expenditure is the best available estimator of human accessibility to date. The accessibility model, however, likely underestimates the true energetic cost associated with reaching interior locations. There are many factors in addition to energetic cost that influence the path that is chosen to reach an interior destination. One of the most important is safety. Footing is more precarious off-trail than on-trail. Scratches and bruises from vegetation are also more likely when walking off-trail. Finally, since the least-costs paths used in this model can have much larger distances than the straight-line distance from the nearest road, orienteering is more difficult. The greater the off-trail distance, the more likely a hiker is to become lost.

The current model does not account for the psychological impact of walking through a *Rhododendron-Kalmia* thicket, nor the fact that some stands are, for all practical purposes, impenetrable. The psychological impact of vegetation, in general, is quite high for most people. Typically, when a person wants to make their way to a point off-trail, the easiest way to do so is to walk along the trail to a point where off-trail distance will be minimized. The impact of walking through vegetation versus walking on a trail, then, is more than just energetic cost. Precarious footing as well as uncertainty about the path impacts the choice of path once you leave the trail. Part of the uncertainty is that the easiest path may not be apparent, especially if the destination is not within sight of the trail. The least-cost path calculated by the model is very useful in such situations. For any interior point in the Park, the least-cost path can be re-constructed from the model. From there, the path may be downloaded into a handheld GPS unit and then used to navigate to the location in the field.

There are weaknesses associated with the using the least-cost path, however. When calculating the least-cost path, the algorithm looks only in the immediate neighborhood (radius 10m) for the next step in the path. The neighbor with the least cost is the one chose for the next step in the least cost path. The algorithm is not aware of the larger scale structure of the landscape as a person would be, at least in some circumstances (Figure 3.5). A person might employ a series of switchbacks to get up a steep slope rather than merely find the shallowest gradient to follow along the slope. A person might also know that going over a small, steep knoll would be most efficient, while a least-cost algorithm would always go around it (Figure 3.5). Nevertheless, the least-cost path typically represents an improvement over the typical human algorithm of minimizing the off-trail distance

Accessibility in Conservation Areas

GSMNP is well-divided by roads and trails. Accessible sites are far more abundant than inaccessible sites. From the standpoint of conservation area planning, the balance of accessible and inaccessible sites is important. Making conservation areas accessible is vital for maintaining public interest in conservation and increasing awareness of conservation issues. When people can physically interact with the landscape, those personal experiences hopefully result in an improved land-ethic and deeper concern for conserving natural areas. Most people are either unwilling or unable to hike to remote locations, and increasing accessibility of the conservation landscape increases human interaction with the landscape. Conversely, increased accessibility results in increased landscape fragmentation, increased disturbance, increased arson fires, and increased spread of exotic invasive species. These particular problems loom large in the minds of conservationists and biologists. Inaccessible areas are less prone to these problems.

In general the area of inaccessible locations required by conservation is much larger than the area required to increase awareness. The goals of increasing conservation awareness and deepening appreciation of conservation areas can be fulfilled with a few well-placed roads. Conserving ecosystem processes in habitats where human disturbance is minimized requires very large areas, especially where large mammals play in important role. Increasing accessibility degrades these ecosystem processes (Nielsen *et al.* 2004;Mclellan & Shackleton 1988;Woodroffe 2000;Whittington *et al.* 2005). There are only two, relatively inaccessible tracts in

GSMNP. Relative to other National Parks of similar size, GSMNP, is very accessible. Increasing fragmentation through road building in park is ill-advised, given that the balance of human use and conservation already leans heavily toward human use.

Sampling Bias and Accessibility

Vegetation surveys in GSMNP occur more frequently in accessible areas than inaccessible areas (Figure 3.1). The results shown here are likely typical of most vegetation surveys. Survey plots tend to be congregated in the most accessible areas of the Park. This effect is most pronounced in the most accessible locations, typically within 100m of a road. There are almost twice as many plots in GSMNP right next to roads than would be expected given a random sample. Samples of intermediate accessibility, on the other hand, have abundances that are typical of a random survey. At a certain threshold, however, this relationship breaks down. For GSMNP, the most inaccessible quarter of the Park has virtually no plots in it. This threshold probably represents an upper limit on time rather than energy expenditure. There is a limit to how far one can walk in a single day, survey the vegetation, and then return. Most vegetation surveys (at least in eastern North America) are not conducted while staying overnight in the interior of the Park, but rather are conducted as day trips.

What are the implications of an accessibility sampling bias in accessibility for data analysis? Some have suggested that samples correlated with roads have little impact on the prediction of species distributions (Reese *et al.* 2005). This is strongly dependent upon the spatial arrangement of species distributions and environment relative to accessibility. If all environments are randomly arranged relative to accessibility, then a sample weighted toward more accessible areas is unlikely to yield different results than an unweighted sample. I have shown this assumption to be false, however, at least in the case of GSMNP, and in general it is clear that in most landscapes environments are not randomly arranged relative to accessibility. Some vegetation communities show their greatest abundance in the more inaccessible areas of the Park. Some are random, and some are most abundant in accessible areas. Even within broad vegetation groups, such as all spruce-fir or all cove forests, the association with accessibility was varied. So, there is no intuitive reason to expect that a sample weighted for accessibility will capture a representative sample of all vegetation communities in the Park. That said, there were no communities in this analysis that were so strongly correlated with inaccessible areas that accessible tracts of the community were completely absent. Every community has some area in the more accessible parts of the Park. Directed samples, then, should be able to capture at least one representative from each community even when deliberately weighted for accessibility.

The community classification I have used here is a very broad-brush approach to the patterns of species composition within the Park. The communities are identified by a few of their dominant trees. As a consequence, the total species composition within these communities varies widely. In some cases, the variation in species composition within a community is likely larger than that between communities. It is unknown what correlations exist between accessibility and species composition within a community. Correlations are likely, however, since accessibility exhibits strong spatial autocorrelation as does species composition. The question remains, then, “Do samples correlated with accessibility capture the full range of variation in species composition within communities?” The answer to this question is dependent upon turnover of species in space and the abundance of the community in inaccessible locations. Larger patches in inaccessible locations are likely to have greater richness than smaller communities. Communities with greater spatial turnover are more likely to differ between accessible and inaccessible areas.

Consider, for instance, the distribution of Spruce-Fir forest in the Park. These forests are dominant across the Park above 1500m. The fir trees (*Abies fraseri*) in these forests were decimated throughout the Park by the Balsam Woolly Adelgid beginning in the 1963. The distribution of these forests is bimodal with respect to accessibility within the Park. There are a few patches that are very accessible because they occur near Clingman’s Dome, the highest point in the Park that has a road running nearly to the top. There are also significant areas of Spruce-Fir Forest in the more inaccessible northeastern section of the Park, and north of Fontana Lake in the southwestern section of the Park. Virtually all the Spruce-Fir Forest samples in the dataset come from the most accessible areas. If, the impact of the balsam woolly adelgid was spatially heterogeneous at scales of greater than ~10km, then these spatially disjunct populations are likely to exhibit differences in species composition not present in the current data.

It is the differences between disjunct communities, and communities that are under-sampled where our understanding of biodiversity patterns is lacking. We have a great understanding of how to evaluate communities at a local scale. Measuring species richness and understanding community dynamics is easier at local scales because they can be fully censused or nearly so. More understanding is needed about the processes and patterns that shape disjunct communities at scales greater than the local scale. Currently, many of our surveys are missing particular scales of community organization because their correlation with accessibility. This may be especially important considering that more inaccessible communities are likely to have less human disturbance. Surveys of any group of organisms should incorporate some sites that are in less accessible locations.

Applications of the Accessibility Model

The potential applications of the accessibility model I have developed here reach far beyond assessing sampling bias in biological surveys. There is the potential for correlations with accessibility between any process that includes interactions between humans and landscapes. Here I discuss the important ecological and conservation questions that can be addressed using the accessibility model.

First, accessibility is a vital correlate in understanding the spread of exotic invasive species. The spread of exotic invasive species tends to track corridors of access such as trails or roads (Gelbard & Belnap 2003). The species composition of inaccessible areas includes fewer exotic-invasive species than more accessible areas. Most measures of this, however, are concentrated immediately along roads where cars are the main dispersal vector. The typical vector for human dispersal of exotic invasive species within the interior of parks, however, is on the clothes and horses. Most parks keep record of the intensity of use for trails and campsites. Combining this information with the model of accessibility allows the creation of a measure of the intensity of human use for the entire park. This composite model will give an estimate of the actual number of humans that walk past a spot during a given period of time. This measure will likely be strongly correlated with the distribution of exotic invasive species as they move from the edges of a park toward more interior locations. It will also predict the most probable locations for new introductions of invasive species.

The accessibility model can improve understanding not only of species that are introduced, but also of species that are being removed. The harvesting of economically important plants is a multi-million dollar industry. Ginseng (*Panax quinquefolius*) and ramps (*Allium tricoccum*) are among the two most important economically harvested species in the eastern United States, but other species such as Black Cohosh (*Actaea racemosa*) and Bloodroot (*Sanguinaria canadensis*) are also harvested. The probability of harvesting an individual from any of the species is likely a function of accessibility. In areas where collection is legal, intensity of harvest is likely to be negatively correlated with accessibility. In areas where collection is illegal, such as in GSMNP, the relationship between harvesting and accessibility is less intuitive. It may be that harvesters not wanting to be observed will travel to more inaccessible locations leaving accessible locations untouched. In either case, the accessibility model can provide a metric for locating areas that are likely to be harvested and show those areas where populations are likely to be undisturbed. Also, as a species is depleted more inaccessible areas are searched.

Since virtually any human disturbance is likely to be correlated with accessibility, the accessibility model provides a unique opportunity to locate communities that are relatively undisturbed. Researchers are always on the lookout for the most undisturbed and unique patches in a landscape. These sites could harbor rare species, and at least provide a baseline for comparison against more disturbed communities. The accessibility model can be used in conjunction with measures of environmental uniqueness to find these patches on the landscape. These inaccessible patches are also good candidates for reintroduction of extirpated species since they are less likely to be disturbed. For these reasons, inaccessible patches of communities represent areas of great conservation import.

The accessibility model outlined here is also a valuable tool for land-use planning in conservation areas. As discussed above, the accessibility model can be used to establish a balance between the abundance of areas that are prioritized for human use and those prioritized for conservation. Understanding the accessibility of a landscape is vital for deciding where new trails should be built. A trail that follows the least-cost path to a destination is likely to be traveled much more frequently than a trail that does not follow that path. Finally, knowing the least cost path to any location in the Park is a valuable tool for fighting forest fires in conservation areas where fires are controlled. Knowing the best point of access and the best path to follow to a fire can save valuable time and lives.

As a final application of the accessibility model, I note that virtually any animal obeys movement rules similar to humans. Animals often choose the least cost path to traverse a landscape, having preference for walking on level slopes than uphill or downhill and avoiding large stream crossings while maximizing protective cover. The actual energetic costs associated with movement scale with body size (West *et al.* 2003), and the biomechanics of the particular animal, but the rules of movement used in the accessibility model remain the constant. A similar model to the human accessibility model I have built could be built for elk, bear, or any of myriads of animals. These models could have important implications for our understanding of metapopulation dynamics as well as our understanding of the scales at which animals interact with landscapes.

Conclusion

Understanding the accessibility of conservation landscapes to humans is vital for understanding a variety of important topics for ecology and conservation. The model of human accessibility I have outlined here currently represents the only developed measure of accessibility that incorporates the effects of slope walking, as well as vegetation and streams. Further, this model estimates accessibility in terms of the round-trip energetic cost of walking (in J/kg), which is a more accurate measure of accessibility than any pure distance metric. Using this model, I have shown that the vast majority of vegetation surveys collected in GSMNP (over 1100 plots) are correlated with accessible areas. Further, I have shown that this sampling bias has caused some important vegetation communities to be under-sampled. These results highlight the need for studies to include at least a few survey plots in more inaccessible areas. Doing so can increase the numbers of communities that are sampled, as well and the variation in species composition that is observed within communities.

The estimates from this model can be incorporated into any existing data that has geolocations information. It can then be used as a surrogate for the probability of human disturbance. There are myriad of species and ecosystem distributions for which this measure will be an important correlate. Further, the distribution of accessibility on a landscape is an important tool for making conservation and management decisions.

References

Brody A.J. & Pelton M.R. (1989) Effects of Roads on Black Bear Movements in Western North-Carolina. *Wildlife Society Bulletin* 17, 5-10

Center for Remote Sensing and Mapping Science (CRMS). Digital Vegetation Maps for the Great Smoky Mountains National Park. Madden, Marguerite, Welch, Roy, Jordan, Thomas, and Jackson, Phyllis. 2004. Center for Remote Sensing and Mapping Science, Department of Geography, The University of Georgia. March 1, 2006.

Gelbard J.L. & Belnap J. (2003) Roads as Conduits for Exotic Plant Invasions in a Semiarid Landscape. *Conservation Biology* 17, 420-432

Gelbard J.L. & Harrison S. (2003) Roadless Habitats as Refuges for Native Grasslands: Interactions With Soil, Aspect, and Grazing. *Ecological Applications* 13, 404-415

Grover K.E. & Thompson M.J. (1986) Factors Influencing Spring Feeding Site Selection by Elk in the Elkhorn Mountains, Montana. *Journal of Wildlife Management* 50, 466-470

Hare T.S. (2004) Using Measures of Cost Distance in the Estimation of Polity Boundaries in the Postclassic Yautepec Valley, Mexico. *Journal of Archaeological Science* 31, 799-814

Imhof E. (1950) *Gelaende und Karte*. Rentsch, Zurich.

Laurance S.G.W., Stouffer P.C. & Laurance W.E. (2004) Effects of Road Clearings on Movement Patterns of Understory Rainforest Birds in Central Amazonia. *Conservation Biology* 18, 1099-1109

Maki S., Kalliola R. & Vuorinen K. (2001) Road Construction in the Peruvian Amazon: Process, Causes and Consequences. *Environmental Conservation* 28, 199-214

McClellan B.N. & Shackleton D.M. (1988) Grizzly Bears and Resource-Extraction Industries - Effects of Roads on Behavior, Habitat Use and Demography. *Journal of Applied Ecology* 25, 451-460

Minetti A.E. (1995) Optimum Gradient of Mountain Paths. *Journal of Applied Physiology* 79, 1698-1703

Minetti A.E., Ardigo L.P., Capodaglio E.M. & Saibene F. (2001) Energetics and Mechanics of Human Walking at Oscillating Speeds. *American Zoologist* 41, 205-210

Minetti A.E., Ardigo L.P. & Saibene F. (1993) Mechanical Determinants of Gradient

Walking Energetics in Man. *Journal of Physiology-London* 472, 725-735

Minetti A.E., Ardigo L.P. & Saibene F. (1994) The Transition Between Walking and Running in Humans - Metabolic and Mechanical Aspects at Different Gradients. *Acta Physiologica Scandinavica* 150, 315-323

Minetti A.E., Moia C., Roi G.S., Susta D. & Ferretti G. (2002) Energy Cost of Walking and Running at Extreme Uphill and Downhill Slopes. *Journal of Applied Physiology* 93, 1039-1046

NatureServe. Nature Serve Explorer: An online encyclopedia of life . 2006. NatureServe. March 1, 2006.

Nepstad D., Carvalho G., Barros A.C., Alencar A., Capobianco J.P., Bishop J., Moutinho P., Lefebvre P., Silva U.L. & Prins E. (2001) Road Paving, Fire Regime Feedbacks, and the Future of Amazon Forests. *Forest Ecology and Management* 154, 395-407

Newall C.L. & Peet R.K. (1998) Vegetation of Linville Gorge Wilderness, North Carolina. *Castanea* 63, 275-322

Nielsen S.E., Herrero S., Boyce M.S., Mace R.D., Benn B., Gibeau M.L. & Jevons S. (2004) Modelling the Spatial Distribution of Human-Ccaused Grizzly Bear Mortalities in the Central Rockies Ecosystem of Canada. *Biological Conservation* 120, 101-113

Reese G.C., Wilson K.R., Hoeting J.A. & Flather C.H. (2005) Factors Affecting Species Distribution Predictions: a Simulation Modeling Experiment. *Ecological Applications* 15, 554-564

Rose Jessica, Ralston H.J. & Gamble J.G. (1994) Energetics of Walking. In: *Human Walking* (eds Rose J. & Gamble James G.), 2nd Edition edn, pp. 45-72. Williams & Wilkins, Baltimore, Maryland.

Rost G.R. & Bailey J.A. (1979) Distribution of Mule Deer and Elk in Relation to Roads. *Journal of Wildlife Management* 43, 634-641

Scott J.M., Davis F.W., Mcghie R.G., Wright R.G., Groves C. & Estes J. (2001) Nature Reserves: Do They Capture the Full Range of America's Biological Diversity? *Ecological Applications* 11, 999-1007

Soule R.G. & Goldman R.F. (1972) Terrain Coefficients for Energy Cost Prediction. *Journal of Applied Physiology* 32, 706-&

Tobler, W. Three Presentations on Geographical Analysis and Modeling: Non-isotrophic modelling, speculations on the geometry of geography, global spatial analysis. 1993. National Center for Geographic Information and Analysis.

Trombulak S.C. & Frissell C.A. (2000) Review of Ecological Effects of Roads on Terrestrial and Aquatic Communities. *Conservation Biology* 14, 18-30

Van Leusen P.M. Pattern to process: methodological investigations into the formation and interpretation of spatial patterns in archaeological landscapes. 2002. Groningen, Netherlands, University of Groningen.

West G.B., Savage V.M., Gillooly J., Enquist B.J., Woodruff W.H. & Brown J.H. (2003) Why Does Metabolic Rate Scale With Body Size? *Nature* 421, 713

Whittington J., St Clair C.C. & Mercer G. (2005) Spatial Responses of Wolves to Roads and Trails in Mountain Valleys. *Ecological Applications* 15, 543-553

Woodroffe R. (2000) Predators and People: Using Human Densities to Interpret Declines of Large Carnivores. *Animal Conservation* 3, 165-173

Figures

Figure 3.1 Energetic cost estimates for GSMNP based on a least-cost path model of round trip energetic expenditures. There are two major areas of inaccessibility: one located north of Fontana Lake and the other in the high-elevation, eastern portion of the Park. The collection of 1104 vegetation surveys that were used to assess sample correlation with accessibility are also shown.

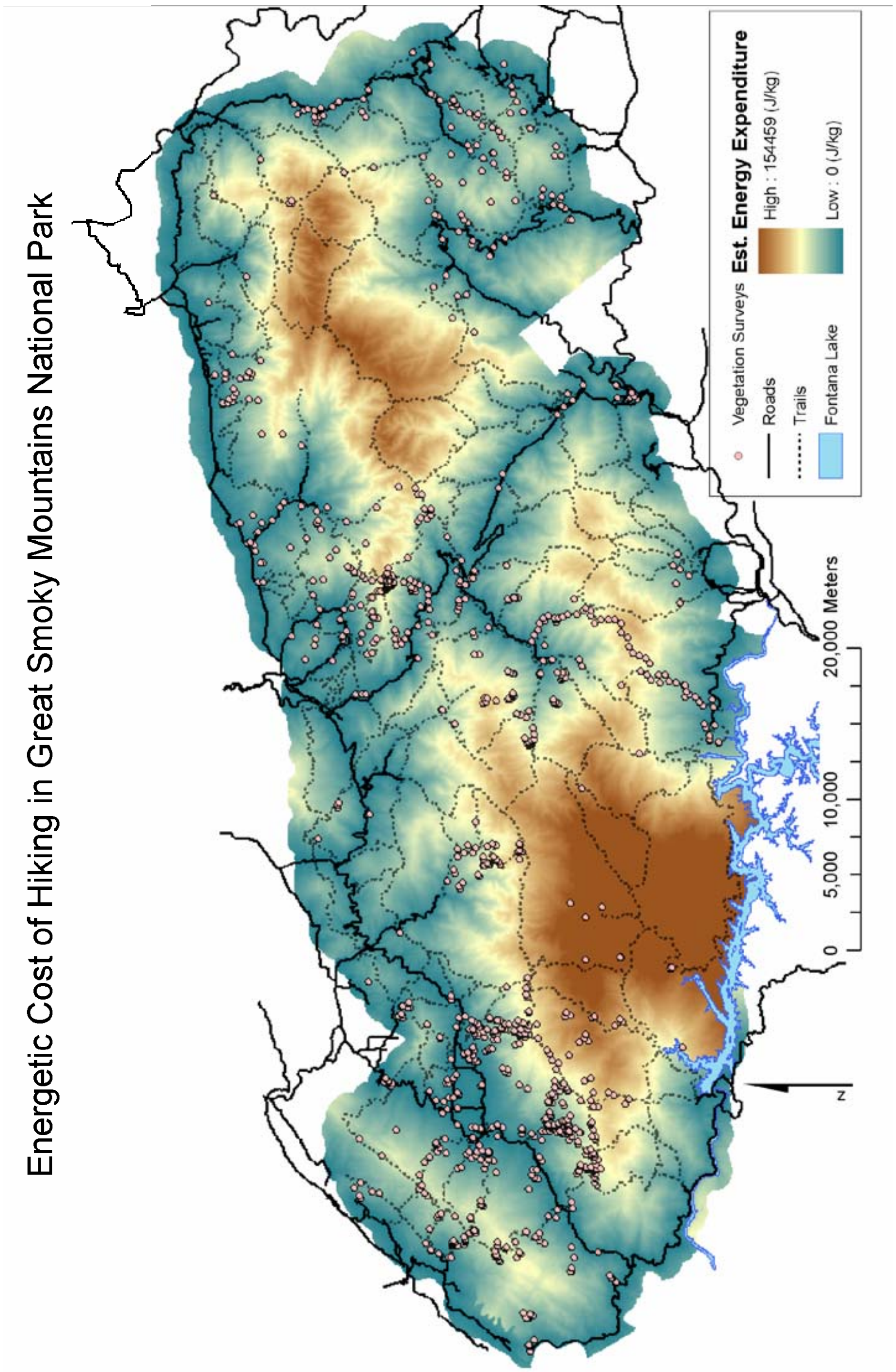


Figure 3.2 Expected and observed frequencies of vegetation surveys by accessibility. Expected frequencies are the number of plots that would be observed if vegetation surveys were uncorrelated with accessibility. Plot frequencies are binned into 50 equal interval quantiles, each spanning 3090 J/kg of energetic cost. The data show a strong correlation with accessible sites, especially in the most accessible sites.

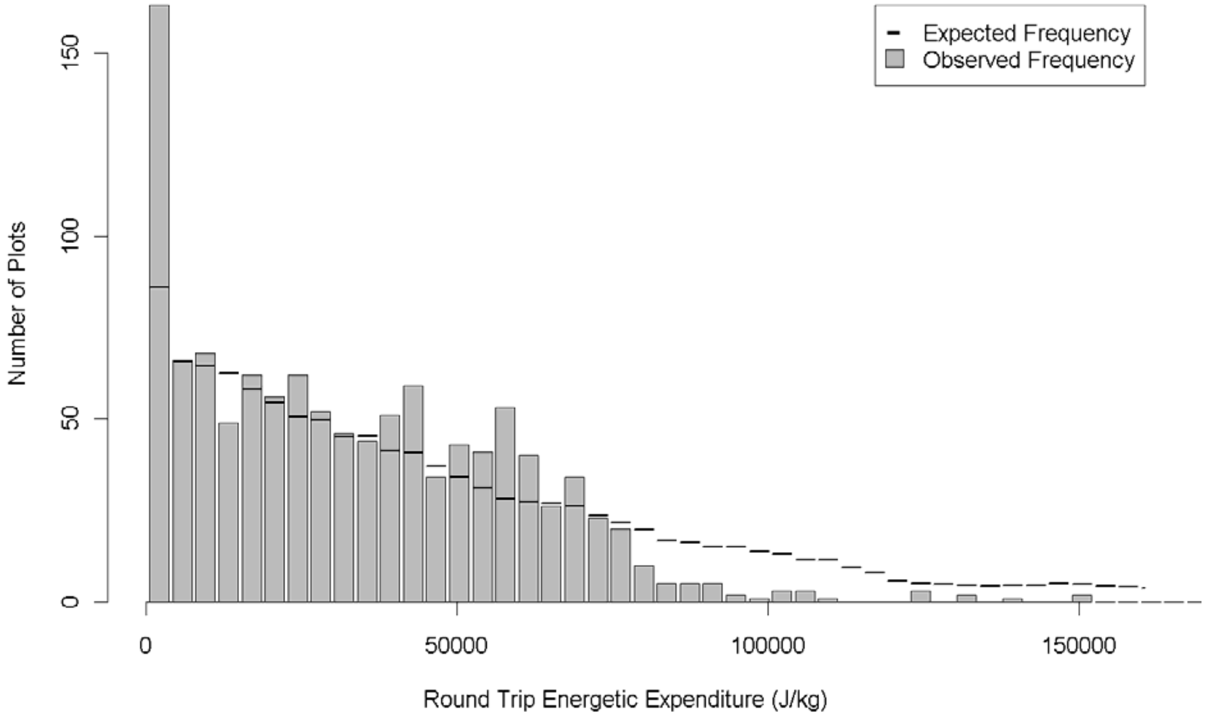


Figure 3.3 The areal distribution of vegetation communities with respect to accessibility. The selections of communities shown are a) over-sampled and b) under-sampled relative to their abundance in the Park. These communities illustrate the range of affinities for accessible locations between over-sampled vegetation communities.

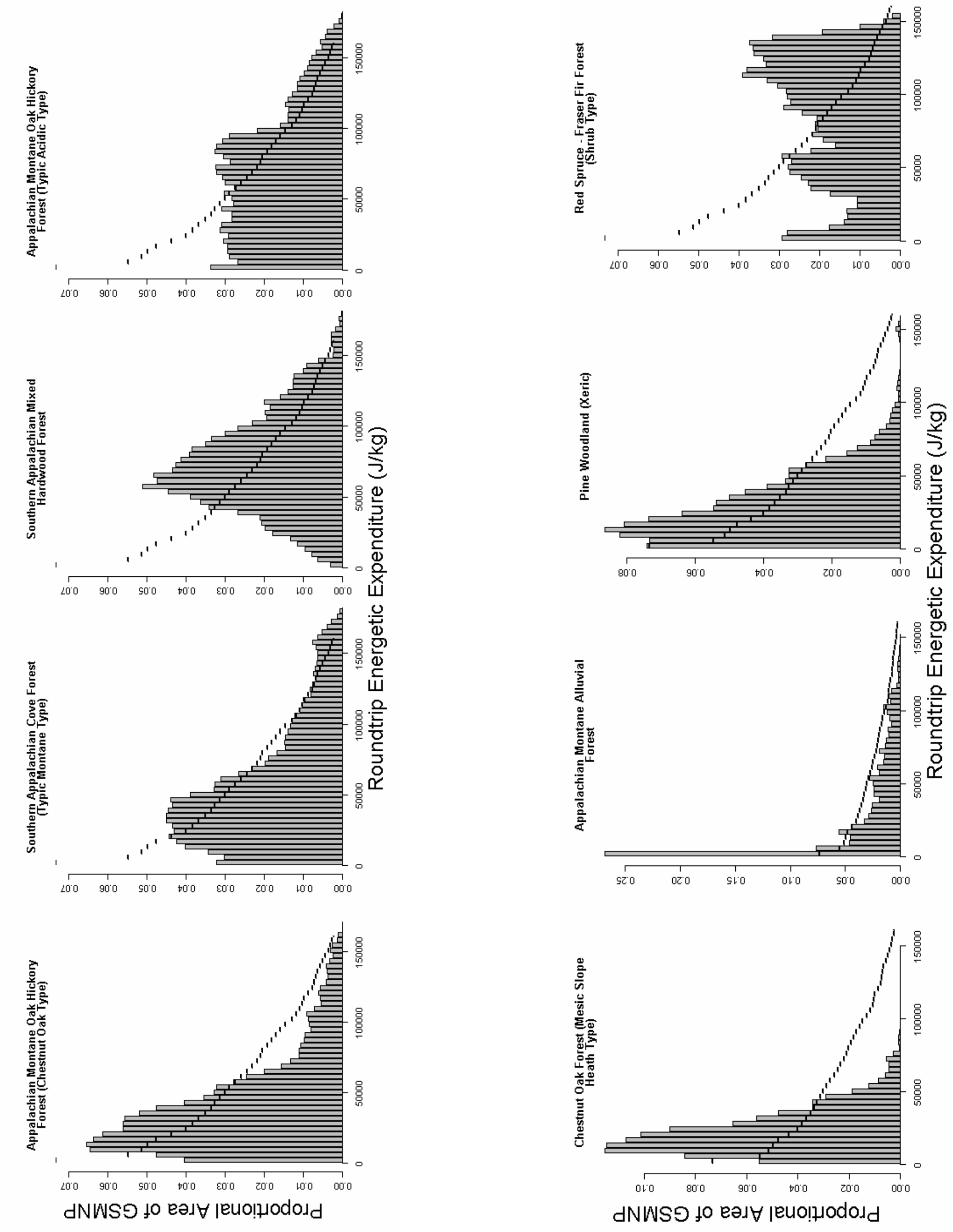


Figure 3.4 The distribution of vegetation surveys with respect to accessibility for selected communities in a) over-sampled communities and b) under-sampled communities. More inaccessible areas of communities tend to be under-sampled, regardless of the community type. This may be of great importance for communities that have a large proportion of their total area in inaccessible parts of the Park: spruce-fir forest, for instance.

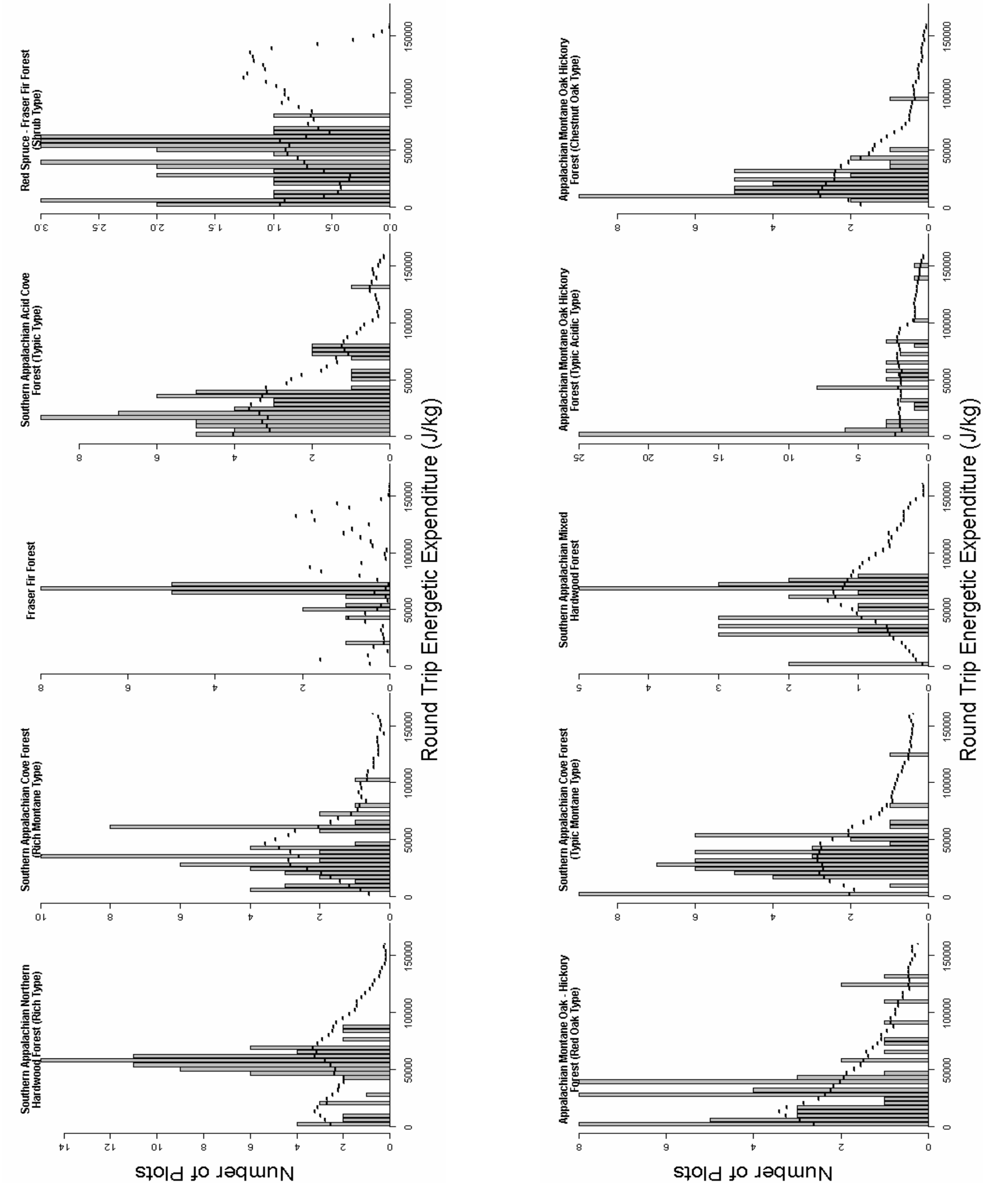
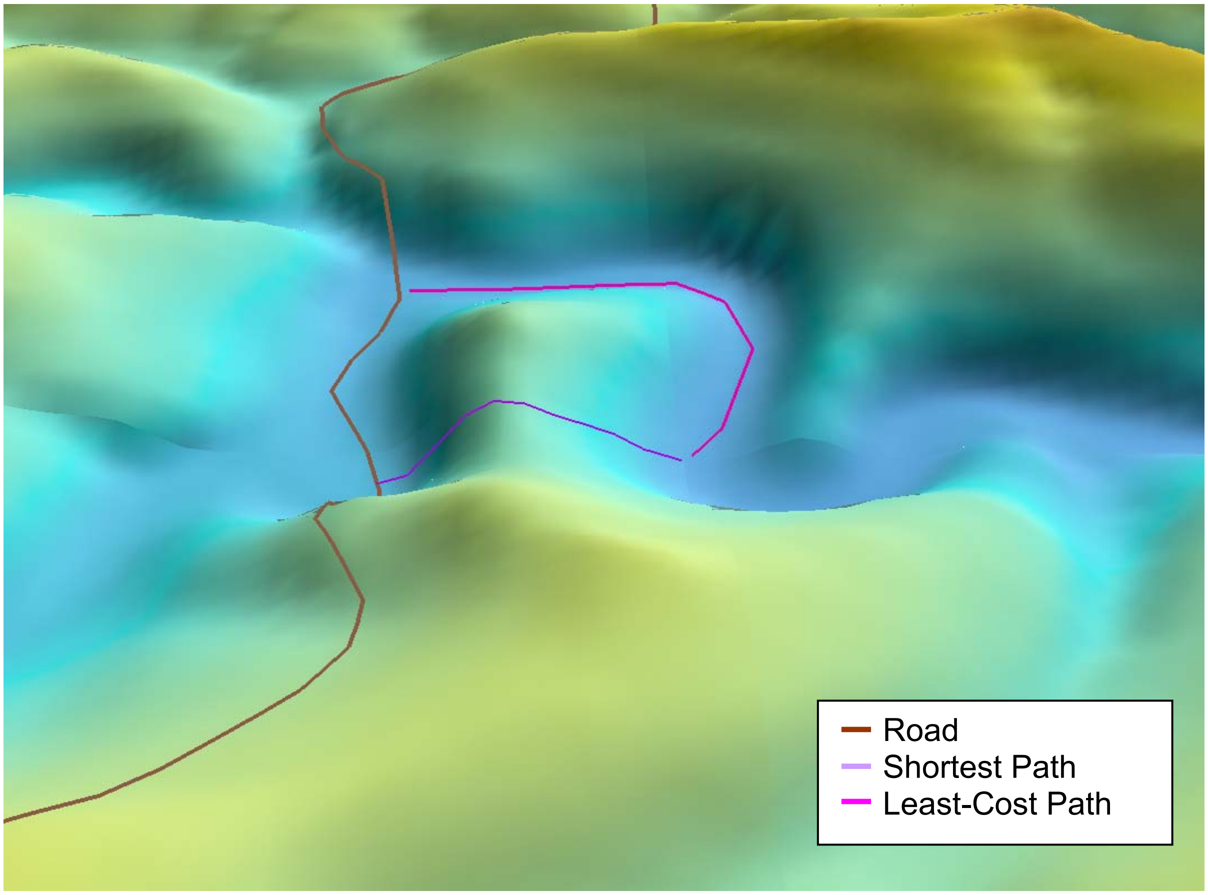


Figure 3.5 3D rendering of a trail, shortest surface path and the least-cost path calculated by the accessibility model to an interior point. Lower elevations are shown in cool colors, while higher elevations are hot colors. This figure highlights a weakness of the least-cost path algorithm used to create the accessibility model. The algorithm with calculates the least-cost path looks only in its immediate neighborhood for the easiest path. In this case, the shortest path is probably the most energy efficient. The least-cost path algorithm could not discern that, however, because the cells close to the trail have a steep knoll separating the trail and the destination.



Tables

Table 3.1 Energetic cost coefficients for different terrains in the least-cost path model of human accessibility. The coefficients used are based on the study of Soule and Goldman (Soule & Goldman 1972). These coefficients determine the relative difficulty of different terrains that are encountered while hiking in the GSMNP. The swamp coefficient of Soule and Goldman

Isotropic Cost Coefficients of GSMNP Energetic-cost Model		
Terrain	**Soule and Goldman 1972**	**Coefficient**
Trail	Dirt	1.2
Off-Trail	Light Brush	1.31
Rhododendron sp.-Kalmia sp.	Heavy Brush	1.59
Stream (Slope → 0)	Swamp	1.87

Table 3.2 The 10 most over-sampled and under-sampled communities in GSMNP relative to their proportional area in the Park. The observed (Obs.) number of plots is the number of vegetation samples falling within the community. The expected (Exp.) number of plots is the number of plots that would be expected if communities were sampled in proportion to their total area in the Park. The difference between the observed and expected number of plots for each community represents the number of plots that the community has been over-sampled or under-sampled.

Top 10 under-collected and over-collected communities in GSMNP					
Code	**Community Description**	**% Area of GSMNP**	**Number of Vegetation Plots**		
			Obs.	**Exp.**	**Obs.-Exp.**
	Over Collected				
4973	Southern Appalachian Northern Hardwood Forest (Rich Type)	3.28	82	37.6	44.4
7695	Southern Appalachian Cove Forest (Rich Montane Type)	2.46	57	28.3	28.7
6000	Fraser Fir Forest	0.20	24	2.3	21.7
7543	Southern Appalachian Acid Cove Forest (Typic Type)	4.18	68	48.0	20.0
6286	Chestnut Oak Forest (Mesic Slope Heath Type)	0.99	29	11.4	17.6
4691	Appalachian Montane Alluvial Forest	1.19	27	13.7	13.3
7285	Southern Appalachian Northern Hardwood Forest (Typic Type)	0.77	22	8.9	13.1
1001	Pine Woodland (Xeric)	6.42	86	73.7	12.3
7299	High-Elevation Red Oak Forest (Evergreen Shrub Type)	2.13	35	24.4	10.6
9000	Red Spruce - Fraser Fir Forest (Shrub Type)	NA	32	21.6	10.4
	Under Collected				
6192	Appalachian Montane Oak - Hickory Forest (Red Oak Type)	11.16	58	128.2	-70.2
7710	Southern Appalachian Cove Forest (Typic Montane Type)	8.76	63	100.7	-37.7
8558	Southern Appalachian Mixed Hardwood Forest	0.06	28	51.9	-23.9
7230	Appalachian Montane Oak Hickory Forest (Typic Acidic Type)	7.98	69	91.6	-22.6
7267	Appalachian Montane Oak Hickory Forest (Chestnut Oak Type)	5.37	43	61.6	-18.6
7300	High-Elevation Red Oak Forest (Deciduous Shrub Type)	1.39	1	15.9	-14.9
7692	Appalachian Montane Oak - Hickory Forest (Rich Type)	1.17	2	13.5	-11.5
7219	Early Successional Appalachian Hardwood Forest	3.59	34	41.3	-7.3
7130	Red Spruce - Fraser Fir Forest (Evergreen Shrub Type)	0.43	0	5.0	-5.0
7097	Blue Ridge Table Mountain Pine - Pitch Pine Woodland (Typic Type)	0.43	0	5.0	-5.0

Chapter 4

Sampling biodiversity at landscape scales: four major obstacles and solutions

Abstract

Four issues of sampling design influence the effectiveness of biodiversity surveys at landscape scales. First, samples must reflect the environmental variability of the landscape since variation in environment influences variation in species composition. Second, samples should capture dominant species compositional patterns. Third, biodiversity surveys must account the spatial autocorrelation present in species distributions. Finally, the costs of sampling increase as the cost of logistics increases and no monitoring program will continue if fieldwork is unreasonably difficult. Sample sites must be accessible though they must also consider the possible biases introduced by accessibility. I describe a sampling protocol designed for the Great Smoky Mountains National Park vegetation-monitoring program that satisfies these design issues. The protocol follows a stratified-clustered design. My results suggest that the environmental variation the spatial scaling of stratification variables must mesh with those of the sampling design to avoid spurious environmental distinctions. Weighting samples toward more accessible locations significantly altered the join-inclusion probabilities and frequencies of environments in samples. The spatial clustering of sample sites tended to reduce the variety of vegetation communities captured by the samples from the random expectation. My stratification scheme tended to capture environmental variety within broad vegetation communities, rather maximizing the total number of communities sampled.

Keywords

sample design, clustered sampling, stratification, spatial autocorrelation, *Tsuga canadensis*, Great Smoky Mountains National Park

Introduction

Biodiversity monitoring programs are vital tools for conservation and ecological research (Noss 1990;Janzen & Hallwachs W. 1994;Savage 1995). As spatial patterns of biodiversity change through time due to climate change (Thomas *et al.* 2004), landscape fragmentation (Vitousek *et al.* 1997), and exotic species invasion (Stohlgren *et al.* 1999), the cataloging and monitoring of species becomes ever more important. This is not an easy task. There are many, sometimes conflicting goals that must be considered when selecting a sampling design for a biodiversity survey. First, pattern and process within a landscape occur across a wide range of scales. Thus, effective sampling must also address pattern and process at multiple scales (Pettitt & Mcbratney 1993;Tobin 2004;Lookingbill & Urban 2005). Further, species distributions are influenced by both endogenous (e.g. reproduction, dispersal) and exogenous factors (e.g. physical environment, disturbance). Separating endogenous effects from exogenous effects predicates the explanatory power of any sample design (Wagner & Fortin 2005).

In this chapter, I present the four main challenges in the design of biodiversity surveys. I then present a sampling protocol that addresses these design issues. This protocol utilizes ecological zipcodes (Chapter 2) to capture a broad spectrum of environments. It aggregates sites at multiple scales, which allows distinguishing between environmental and dispersal processes. The protocol also weights sites by accessibility in a way that maximizes data collection efficiency while sampling a broad spectrum of environments. Using Great Smoky Mountains National Park as an example, I discuss how to implement and analyze data collected using the protocol. I analyze the effect of incorporating accessibility bias into samples changes their representation of the landscape. Since one of the important functions of a biodiversity survey is to provide a baseline for additional studies, I analyze the effectiveness of this protocol at capturing the variety of environments in which eastern hemlock (*Tsuga canadensis* (L.) Carr.) occurs. Finally, I discuss a balance of sampling efficiency and representation.

Background

The role of sample design in ecological inference

The ability of ecologists to understand pattern and process in nature is tightly bound to issues of sampling design. The conclusions of any ecological study rest on how faithfully data reflect the pattern or process of interest. When data do not conform to natural patterns, research conclusions range from tenuous at best to entirely false.

At scales where complete or nearly complete surveys are possible, debate about the accuracy of samples becomes moot. Instead, distinguishing between possible explanatory variables (Condit *et al.* 1996) and interpreting patterns in the midst of stochasticity (Levin 1992) become more important. At scales greater than this, researchers must rely upon incomplete samples to understand pattern and process. Landscapes on which biodiversity monitoring takes place (typically 10^3-10^6ha) fit the latter scale.

Landscape sampling designs

There are myriad sampling designs for landscape-scale ecological studies. The spatial arrangement of sampling units (e.g. quadrats, plots, traps) in these designs vary along a gradient from clustered to random to regular. Most surveys lie somewhere between clustered and random. The benefit of random sampling is independence among sampling units. As a result, there is no bias added to the survey due to sampling design, though bias due to spatial autocorrelation in species and environments is present in any sample (Legendre 1993). Conversely, regular and clustered sampling designs show strong dependence among sampling units and strong spatial bias. In regular designs, each site is spatially dependant upon the location of every other site.

Given a large area, both random and regular sampling designs have very low efficiency, which is the energetic or monetary cost associated with completing the survey. In these designs, the probability of many sample units occurring in inaccessible locations is high. Further, the distance between sequentially surveyed sample units is great. Random and regular sample designs also tend to miss ecologically important scales in a landscape (Fortin *et al.* 1989). Sample intensity is not high enough to capture both small- and large-scale processes (Urban *et al.* 2002;Tobin 2004). These weaknesses highlight the strengths of clustered sampling: efficiency and broadly ranging scales of spatial aggregation. Clustered designs allow the surveying of many sampling units in a few areas. Further, clustered designs allow short distances between sampling units (10s to 100s of meters) to be well represented in addition to larger, landscape-scale distances (Urban *et al.* 2002;Urban 2000;Fortin *et al.* 1989;Legendre & Fortin 1989).

The four challenges of sampling design for biodiversity monitoring

Regardless of the design, whether clustered, regular or random, there are four main challenges facing landscape-scale biodiversity surveys. First, any survey must faithfully represent landscape biodiversity patterns and must capture the range of variation present in the environment (temperature, moisture, disturbance history). Increasing the environmental variety of samples increases the complementarity in species composition between sites (Faith *et al.* 2004;Faith & Walker 1996;1995). This yields greater species richness for a given sample effort: an important component of biodiversity. Environmental variety in a sample is also important for monitoring changes in species distributions through time. Habitats that initially seem unimportant in terms of diversity may become more important as climate changes or as disturbance increases.

Second, biodiversity surveys must be sensitive to the relative abundance distribution of species. The distribution of species abundances on a landscape controls the observation rate of species in a survey (Fisher *et al.* 1943;Preston 1948). Surveys tend to record many individuals of common species, for any single individual of a rare species. It is these rare species, however, that are often of greatest conservation and monitoring import. Biodiversity monitoring programs must capture a better than random set of these rare species.

Third, biodiversity surveys must account for the spatial aggregation and covariation of environments and species (Legendre 1993). Rarely, do species exhibit random distributions in space. Most of the time species exhibit some degree of spatial contagion, and in rare cases over-dispersion. These patterns may be due to dispersal history, disturbance, or spatial covariation of the environment. Disentangling these effects from each other is not a trivial matter (Koenig 1999). Without accounting for these processes in sample design, it is difficult to separate endogenous effects from exogenous ones.

Finally, biodiversity surveys should maximize efficiency. Even a well-designed biodiversity survey from the standpoint of ecological inference will fail if it is cost-prohibitive. Thus, making data easy to collect is vital. This goal can conflict, however, with other goals such as environmental representation. This chapter discusses how best to balance efficiency and representation.

Methods

Study Area

I generated a series of biodiversity samples for Great Smoky Mountains National Park (GSMNP). The Park area is ~200,000 ha. It is part of the Southern Appalachian Mountains, which are significantly enriched in endemic plants. To date, 129 native tree species have been documented within the Park (taxonomic reference: Weakley 2006, species list for the Park updated from: White 1982 and White & Wafford 1984). The major environmental gradients that influence species composition in the Park are elevation, exposure and relative wetness (Whittaker 1956). Geology and soil chemistry are also important factors (Cain 1931). The Park is over 95% closed canopy forest. At low elevations, cove hardwood forests dominated by *Acer saccharum*, *Liriodendron tulipifera* and *Aesculus flava* occur in moist, protected areas. These coves are the most species rich areas in the Park in terms of vascular plant diversity. On drier sites at lower elevations, *Quercus coccinea* and *Carya* spp. dominate with *Pinus* spp. forests dominating on the most xeric sites. Northern Hardwood forest dominated by *Quercus rubra*, *Betula alleghaniensis*, and *Fagus grandifolia* occur at mid to high elevations. At elevations greater than 1500m, *Picea rubens-Abies fraserii* forest dominates, along with Northern Hardwood Forests in sheltered areas.

Data

Generating samples following the protocol required two GIS data layers. The first was a classification of the Park's important environmental gradients. I used this classification to stratify samples. In doing so sampling units were located in a great variety of the Park's environments. The classification that I used, known as "ecological zipcodes" (Chapter 2), assigns a single number to each location in the Park. This number succinctly represents the states of temperature, insolation, and water availability: gradients that are strongly correlated with species distributions (Hawkins *et al.* 2003;Whittaker 1956). The ecological zipcodes use surrogate variables instead of measuring temperature, insolation, and water availability directly. All of these variables were derived from a single digital elevation model (DEM): 1) Elevation, 2) relative wetness in the form of the topographic convergence index (TCI;Moore *et al.* 1991;Wolock & Mccabe 1995;Yeakley *et al.* 1998), and 3) hillshade (a metric of transformed aspect: azimuth 135 altitude 45). The Park contains 45 different ecological zipcodes comprising five elevation levels, three levels of relative wetness, and three levels of hillshade.

The second GIS data layer needed was an estimate of the energetic expenditure necessary to reach any location in the Park. While a random sample of locations might be easiest to analyze, statistically, the energetic expenditure necessary for a random sample is prohibitive. In a 200,000 ha park with few roads and great topographic complexity a random sample is impractical. A measure of park accessibility, however, allows selection of sample units in such a way as to maximize the efficiency with which data are collected. I have developed such a model for GSMNP (Chapter 3). It estimates round-trip energetic cast associated with surface distance, terrain type, and slope. It also includes frictions associated with the most important features in the Park from an accessibility perspective: the distributions of *Rhododendron* sp*p*. and *Kalmia latifolia*. Their abundance within the Park and their very dense growth habit play an important role in overall accessibility. Limiting the population of possible sampling units to those of low energetic-cost maximizes the efficiency of data collection. To make analysis easier, I transformed the accessibility model from pure energetic costs in J/kg to relative energetic cost. The transformed measured assigns a percentile of accessibility for each location in the Park, based on the number of locations that are more accessible and less accessible. For a given location, a value of 45% indicates that 45% of the Park is more accessible than that site.

The ecological zipcodes and human accessibility data layers provide the necessary information to ensure that samples are both representative and easy to collect. To assess the representativeness of samples, I used an additional data source: an overstory vegetation map of the Park developed by the University of

Georgia Center for Remote Sensing and Mapping (Center for Remote Sensing and Mapping Science (CRMS) 2004). This data layer divides the Park into 171 different vegetation communities. They assigned community classification based on the interpretation of aerial photos. I used this classification to assess how well the samples stratified by environment (i.e. the ecological zipcodes) captured vegetation patterns. I also used the classification to determine the distribution of forests dominated by eastern hemlock.

Sample Design

The sample design I selected is a two-stage stratified-clustered design (Figure 4.1). Stratification occurs in both the inter- and intra-cluster selection of sites based upon the ecological zipcodes. Each zipcode represents a distinct environment varying along elevation, wetness, and insolation gradients. Capturing a wide range of environments should result in a greater than random representation of communities in the sample. Stratification forces some sampling units to be located in rare environments. A random sample of the Park would not capture such environments.

In the first stage of sampling, I selected random points within each ecological zipcode (Figure 4.2). Each point serves as the center of a cluster, and is the location of the first sampling unit within a cluster. Optionally, I added weightings for accessibility at this stage. I used the simplest possible weighting model. The selection probability of a point is proportional to the accessibility percentile raised to an exponent:

$$w = a^z \tag{1}$$

where *w* is the relative weight, *a* is the accessibility percentile, and *z* is the weighting parameter. I generated samples with weighting exponents (*z*) varying from zero (unweighted) to 16.

At the second stage of sampling, I selected points clustered around each of the stage-one points. Before this step, I determined a maximum radius for including points in a cluster. Stratification by ecological zipcodes occurred at this stage, as well (Figure 4.2). The resulting sample contains central sites stratified by ecological zipcode between clusters, and sites within each cluster stratified by ecological zipcode.

Sample Generation and Analysis

I generated a series of samples using the above design for GSMNP. I analyzed the spatial structure of the ecological zipcodes using mantel correlograms (Legendre & Fortin 1989; Legendre and Legendre1998), and measured the environmental variety present at different scales within the Park. I used these data to select variables to include in the stratification of the sample, as well as the maximum cluster radius. I assessed representativeness of samples using the joint probability distribution of ecological zipcodes for the selected maximum cluster radius.

To analyze the effect that sample bias has on accessibility, I generated 1000 samples for each of seven different weightings ranging from unweighted to a weighting exponent (*z* in Eq. 1) of 16. Each sample consisted of 45 clusters (1 cluster per zipcode) with three sites in each cluster. I tested the null hypothesis that weighted samples derived from populations are identical to that of an unweighted sample using a χ^2 goodness-of-fit (GOF) test. To assess how well the sampling protocol captured vegetation communities, I intersected these samples with the vegetation community classification. I performed a χ^2 GOF test on the variety and frequency of vegetation communities captured by the samples against the null hypothesis that samples reflected the actual variety and frequency of communities in the Park. Finally, I tested variety and abundance of hemlock communities in the sample against the distribution of those same communities within the Park to assess the representativeness of the sample for not only the Park as a whole but also individual taxa.

Results

Spatial structure of the environment

The spatial structure of each variable in the ecological zipcodes has a profound impact on the spatial structure of the sample. I found that if the variables used for stratification are spatially autocorrelated at scales greater than the size of a cluster, then stratification within a cluster will results in "false" environmental distinctions. If the average patch size of a particular digit of the zipcode is greater than that of a cluster, then stratifying by ecological zipcode will tend to place sites at the border of patches. The differences suggested by the ecological zipcode, then, really illustrates the weakness of categorical variables rather an actual environmental distinction. These environmental variables would be gradients with respect to a cluster as opposed to spatial structures (*sensu* Legendre 1993).

To address this issue, I sampled the values of each variable making up the ecological zipcodes at 2000 locations within the Park and generated spatial autocorrelation functions based on these samples (Figure 4.3). The distances at which hillshade and TCI are not spatially autocorrelated were quite small relative to the spacing of the 2000 random points throughout the Park. The spatial autocorrelation function did not provide a fine resolution at these scales. So, I randomly sampled 2000 points from a randomly selected circle of radius 2000m. Both hillshade and TCI become uncorrelated at scales between 200m-400m (Figure 4.3b, c). This evidence suggested stratifying samples by TCI and hillshade at the scale of a cluster (<1km) was appropriate. Elevation, on the other hand, showed significant positive autocorrelation at distances out to about 30km and was negatively autocorrelated at distances greater than this. This suggests that stratification within clusters by the elevation variable in the ecological zipcodes is inappropriate. Within-cluster stratification by elevation would merely place sample unit on the border between two larger patches of different elevation. Therefore, when selecting the population of possible cluster locations, I considered only locations showed variety in the hillshade and TCI digits of the ecological zipcodes. Stratification by elevation still occurred in the first stage selection of sites.

The maximum cluster radius (*r*) of the sample is the maximum distance between the central point of a cluster and the additional cluster points. This distance sets both the maximum distance between points in cluster and the minimum distance between clusters at 2*r*. The range of values that *r* can take is bounded on both ends. The spatial scaling of the stratification variable and the size of sampling units set the minimum value of *r*. Accessibility and efficiency determine the maximum value of *r*. If *r* is large, the efficiency gained by clustering samples is lost. Given these boundaries, the ideal cluster size can be determined by analyzing the variation in environment with maximum cluster size (Table 4.1).

Since stratification occurs in both intra-cluster plot selection and infra-cluster selection, there must be reasonable assurance that a randomly selected site has at least as many environments as the number of intra-cluster samples. I calculated the probability of any single site having at least a given number of ecological zipcodes (only considering the hillshade and TCI digits) within circular windows of increasing radii (Table 4.2). For this study, the number of plots within a cluster was 3. The radius at which 95% of sites in park contained at least 3 different ecological zipcodes was 400m. This distance is similar to the lag distances at which TCI and hillshade show little or not spatial autocorrelation (Figure 4.3) and supports the idea that stratification by these variables within clusters results in "true" environmental differences. These results suggested that I select 400m as the maximum cluster radius (*r*).

Implementation of the Sampling Design

Once I determined the maximum cluster radius for the sample design, I could generate the population of sites from which to select the cluster centers. Sites selected from this population are stratified by the full ecological zipcodes. This is the stage 1 population. The stage 1 population is not a random subset of all locations in the Park. Instead, it is limited to those areas that have at least three ecological zipcodes (differences in TCI and hillshade only) in a circle with radius equal to the maximum cluster radius (400m). Though it is not a random subset, it is a representative subset, containing 95% of the total park area (Table 4.2). I randomly selected one point from this population for each ecological zipcode. This set of points served as the cluster centers. I then selected two additional points randomly around each cluster center within the maximum cluster radius. These points were stratified by ecological zipcode so that all three sites within a cluster (the center and the 2 outlying sites) were of different environments (Figure 4.2).

Accessibility Bias

I generated such samples 1000 times for each of six different accessibility weightings (Figure 4.4). Mean accessibility percentile in the samples ranged from 50% for unweighted to 8% for a weighting factor of 16 (*z* in Eq. 1; Figure 4.5). This means that when *z*=16, samples were, on average, located within the 8% most accessible locations in the Park. Weighting samples toward more accessible locations resulted in a strong bias in the joint-inclusion probabilities of ecological zipcodes as well as the frequency distribution of ecological zipcodes (Table 4.2). For joint probabilities of ecological zipcodes, only samples of minimal weighting (*z*=3) did not reject the null hypothesis that the weighted sample population is indistinguishable from the unweighted sample population. Zipcode frequency distributions between weighted and unweighted samples were significantly different for all weightings. This suggests that the topographic environment of GSMNP from which the zipcodes derive correlates strongly with accessibility. This is not unexpected given that one parameter of the accessibility model is the same DEM as that for the ecological zipcodes. Roads and trail tend within the Park tend to follow drainages and to be located landscape positions that are less steep.

Vegetation community representation

The spatial clustering of sample sites tended to reduce the variety of vegetation communities captured by the samples from the random expectation (Figure 4.6). This seems intuitive given the spatial autocorrelation of vegetation communities (Koenig 1999;Legendre 1993). Sites in close proximity should have a greater probability of being from the same vegetation community. Though clustering tended to reduce representativeness of vegetation communities, stratification by ecological zipcode increased the average number of communities captured in a sample (Figure 4.6). This increase in the variety of vegetation communities through stratification was not strong enough, however, to overcome the reduction in variety due to sample clustering.

Forests dominated by Eastern Hemlock

Forests dominated by eastern hemlock occupy about 10% (~20000 ha) of the Park area. Samples generated using the protocol reflected this proportion. Approximately 10% of the samples sites fell in hemlock forest. The proportion of hemlock in a sample did not increase significantly with increasing sample weighting for accessibility. Though the proportion of hemlock forest captured by the sample was typical of a random sample of the Park as a whole, the diversity of ecological zipcodes for sites containing hemlock was greater than the random expectation (Figure 4.7). Hemlock abundance in the sample matched that of the Park, but in a greater variety of environments.

Discussion

I now assess these results by turning back to the four design challenges facing landscape-scale biodiversity surveys. I comment on how this protocol addresses those challenges, and what conceptual basis these results provide for overcoming them.

Environmental variety

Environmental variety in samples is a desirable trait of biodiversity surveys. Samples that maximize environmental diversity tend to capture more species than random samples (Faith *et al.* 2004). Further, environmental diversity is at least as important as species diversity for detecting change in species distributions through time, since species tend to respond to processes like climate change in the context of an environmental template. Sample stratification ensures a wide variety of environments regardless of their abundance. The ecological zipcodes are one approach to environmental stratification and have some distinct advantages. First, they link directly to important ecological processes. Second, they derive from readily available remotely sensed data. Finally, they efficiently classify a multivariate, continuous environment in a univariate, nominal measure.

In order for a sample to represent faithfully environmental variation, however, the spatial scaling of stratification variables must mesh with those of the sampling design (Urban *et al.* 2002;Tobin 2004). If the distance between sites is less than the mean distance between unique patches of the environmental variable, then any stratification across the gradient will result in a false environmental distinction. For this study, the distances between unique patches of elevation were much larger than the distance between sites within a cluster. Stratification within a cluster could not include this variable. Relative wetness (TCI) and insolation (Hillshade) on the other hand, become spatially uncorrelated at distances of hundreds of meters. These variables were good candidates for within-cluster stratification of sample sites. The important message from these results is this. When sampling for environmental diversity, ecological importance may not be the only criteria for variables. Special care must be given to the variety of spatial autocorrelation structures in the set of environmental variables. Given two variables of equal explanatory power, the variable that maximizes the variety of spatial scaling should be chosen.

Relative Abundance Distribution of Species

Though clustered sampling tends to improve efficiency, it also tends to miss rare environments (Figure 4.6). To capture rare species, clustered sampling designs need stratification along a gradient important for species composition. Even with stratification by the ecological zipcodes, stratified-clustered samples tend not to capture rare vegetation communities more frequently than their abundance on the landscape suggests. This result implies that the ecological zipcode classification is coarse relative to the subtle environmental variations that may predict rare vegetation communities. This is not surprising, given the fact that the ecological zipcodes do no include an edaphic component: an important predictor of vegetation communities in the GSMNP (Cain 1931;Bratton 1976).

Hemlock representation in the samples followed much the same pattern as that for all vegetation communities. In general, hemlock occurred in the samples in proportion to its abundance in the Park. The variety of environments, however, in which hemlock was present was greater than would be expected by random. Rare environments such as high-elevation hemlock stands (>1400m) were present in samples. The variety of environments captured in the sample important for threatened species such as hemlock. These extreme environments represent possible escapes or storehouses of genetic variability, which can facilitate species persistence.

Overall, stratified-clustered sampling may not be the most appropriate for capturing the rarest species at a landscape scale. These results suggest that capturing rare environments does not guarantee the presence of rare communities or species. Instead, the protocol I have outlined here captures a wider variety of environments within the broader context of vegetation patterns. If the goal of a survey is just to capture rare species, an adaptive-clustered design may be more appropriate. The basic protocol is the same, except that encounters with rare species require more attention. Rare species necessitate additional samples. Statistical estimators, such as the Horvitz-Thompson estimator, address changes in second-order inclusion probabilities that result from such adaptive sampling (Philippi 2005;Thompson 1991).

Spatial covariation of environment and species distributions

From exogenous factors, such as topography, soil chemistry, and competition to endogenous factors such as reproduction and dispersal, the variables controlling species distributions exhibit spatial structuring. Prioritizing these factors for by their importance to species is fundamental to understanding why species distributions exhibit their own spatial structure. Unfortunately, it is impossible to stratify samples by every variable that influences species distributions. The result is that, even after removing variation from measured variables, geographically distant sites show greater dissimilarity in both environment and species composition than neighboring sites (Nekola & White 1999). At a landscape scale, then, how does one distinguish between the influences of measured variables from unmeasured variables? The answer lies in the spatial structuring of the sample. The protocol has two properties that help to distinguish between measured environmental variables and other unmeasured, yet spatially autocorrelated variables: clustering of sample sites and two-stage stratification.

Clustering sites in a sample imparts two explicitly defined spatial scales: within-cluster and between-cluster. This has some advantage over both random and regular sampling schemes. While random samples possess multiple scales of spatial aggregation, at a landscape scale there will be few sites that are close together (on the order of tens of meters). The same is true of regular sampling. Distances among sample units are not small, unless the landscape is limited in extent and sample size is high. Yet, it is at these distances that many important ecological processes occur. I have shown that for GSMNP both wetness and insolation vary on the order of tens to hundreds of meters. Dispersal processes (Clark *et al.* 1998) and soil gradients (Palmer & Dixon 1990;Bratton 1976) vary at these scales. Clustering of sample units allows for many replicates at small scales to be present in the sample, while retaining the larger scale inter-plot distances.

Stratification of the samples by environment both between clusters and within clusters allows the separation of measured environmental effects from unmeasured effects. There is no stratification within clusters in typical stratified-clustered or randomized-block design. Instead, these designs assume homogeneity with clusters. The problem with this approach is that it is difficult to distinguish between effects due purely to the stratification variables and those due to unmeasured effects that happen to be spatially autocorrelated at the same scale. Having environmental variables stratified at two explicit scales, separates the effects of the stratification variables from those that contribute to larger scale distance decay of similarity. Consider, for example, a sample of relatively wet versus relatively dry sites. If contrasting sites are only present between clusters separated by thousands of meters on average, it is difficult to know whether differences in species composition are due to wetness or to some other unmeasured environmental variable. If, however, contrasting sites are present both between clusters and within clusters, and differences in species composition remain, these differences are likely due to wetness as opposed to some unmeasured variable that just happens to be uncorrelated at large spatial scales. This also highlights the importance of knowing the scales at which stratified variables are spatially autocorrelated. Matching the size of the cluster to the scales at which important variables change is vital to separating the pure effects of these variables from other unmeasured effects.

Sampling Efficiency vs. Environmental Representativeness

Two design components affect the efficiency of a sample: the absolute accessibility of sites, and the relative distance between consecutively surveyed sites. Cluster of samples sites controls the latter. Weighting sites by accessibility controls the former. Increasing sample efficiency comes at the cost decreases the degree to which the sample reflects the landscape. My results suggest that bias in representation is significant for even small weightings in accessibility. This result is profound and disturbing considering that the vast majority of landscape-scale ecological datasets exhibit some bias for accessibility (Chapter 3). The χ^2-GOF test used in this analysis, however, is conservative. In some studies, bias away from a random population may acceptable. In those cases, χ^2 statistics estimate the relative bias that introduced by weighting for accessibility.

An alternative for increased efficiency is increasing the number of sites per cluster. As number of sites per cluster increases, the maximum cluster radius must also increase in order to capture a representative set of environments in the landscape. The relative increase in the maximum cluster radius is dependent upon the spatial heterogeneity of the environment. For GSMNP, doubling the number of sites per cluster increased minimum cluster radius from 200m to 500m in order for 95% of the Park to be included in the population (Table 4.1). Given this radius, plots could potentially be 1km apart. At greater radii, decreases in efficiency outweigh the other benefits of clustered sampling.

Conclusion

I have shown that two-stage stratified clustered sampling using ecological zipcodes is an effective method for capturing a wide variety of environments within the broad patterns of vegetation. The two-stage stratification allows effective assessment of spatial covariation of environment and species distributions as well as separation of environmental effects from other spatially structured processes such as dispersal. I found that while clustered designs improve efficiency and incorporate fine and broad scale pattern, they tend to reduce the number of different communities captured by a sample. This highlighted the need for choosing appropriate stratification variables in any clustered design. The choice of stratification variables must also maximize variety in spatial scaling. Finally, I found that by weighting the sample using a model of accessibility sampling efficiency is increased, though, sampling more frequently in a more broadly dispersed area may offer the best balance between sampling efficiency and environmental representativeness.

References

Bratton S.P. Resource Division in an Understory Herb Community: Responses to Temporal and Microtopographic Gradients. American Naturalist 110[974], 679-693. 1976.

Cain S.A. Ecological Studies of the Vegetation of the Great Smoky Mountains of North Carolina and Tennessee. I. Soil Reaction and Plant Distribution. Botanical Gazette 91[1], 22-41. 1931.

Center for Remote Sensing and Mapping Science (CRMS). Digital Vegetation Maps for the Great Smoky Mountains National Park. Madden, Marguerite, Welch, Roy, Jordan, Thomas, and Jackson, Phyllis. 2004. Center for Remote Sensing and Mapping Science, Department of Geography, The University of Georgia. March 1, 2006.

Clark J.S., Fastie C., Hurtt G., Jackson S.T., Johnson C., King G.A., Lewis M., Lynch J., Pacala S., Prentice C., Schupp E.W., Webb T. & Wyckoff P. (1998) Reid's Paradox of Rapid Plant Migration - Dispersal Theory and Interpretation of Paleoecological Records. *Bioscience* 48, 13-24

Condit R., Hubbell S.P., Lafrankie J.V., Sukumar R., Manokaran N., Foster R.B. & Ashton P.S. (1996) Species-Area and Species-Individual Relationships for Tropical Trees: a Comparison of Three 50-Ha Plots. *Journal of Ecology* 84, 549-562

Faith D.P., Ferrier S. & Walker P.A. (2004) The Ed Strategy: How Species-Level Surrogates Indicate General Biodiversity Patterns Through an 'environmental Diversity' Perspective. *Journal of Biogeography* 31, 1207-1217

Faith D.P. & Walker P.A. (1996) Environmental Diversity: on the Best-Possible Use of Surrogate Data for Assessing the Relative Biodiversity of Sets of Areas. *Biodiversity and Conservation* 5, 399-415

Fisher R.A., Corbet A.S. & Williams C.B. The Relation Between the Number of Species and the Number of Individuals in a Random Sample of an Animal Population. The Journal of Animal Ecology 12[1], 42-58. 1943.

Fortin M.J., Drapeau P. & Legendre P. (1989) Spatial Auto-Correlation and Sampling Design in Plant Ecology. *Vegetatio* 83, 209-222

Hawkins B.A., Field R., Cornell H.V., Currie D.J., Guegan J.F., Kaufman D.M., Kerr J.T., Mittelbach G.G., Oberdorff T., O'brien E.M., Porter E.E. & Turner J.R.G. (2003) Energy, Water, and Broad-Scale Geographic Patterns of Species Richness. *Ecology* 84, 3105-3117

Janzen D.H. & Hallwachs W. All Taxa Biodiversity Inventory (ATBI) of Terrestrial Systems. A generic protocol for preparing wildland biodiversity for nondamaging use. Report of an NSF Workshop, 16-18 April 1993. 1994. Philadelphia, PA.

Koenig W.D. (1999) Spatial Autocorrelation of Ecological Phenomena. *Trends in Ecology & Evolution* 14, 22-26

Legendre P. (1993) Spatial Autocorrelation - Trouble or New Paradigm. *Ecology* 74, 1659-1673

Legendre P. & Fortin M.J. (1989) Spatial Pattern and Ecological Analysis. *Vegetatio* 80, 107-138

Legendre P. & Legendre L. (1998) *Numerical ecology. 2nd English edition*. Elsevier, Amsterdam.

Levin S.A. (1992) The Problem of Pattern and Scale in Ecology. *Ecology* 73, 1943-1967

Lookingbill T.R. & Urban D.L. (2005) Gradient Analysis, the Next Generation: Towards More Plant-Relevant Explanatory Variables. *Canadian Journal of Forest Research-Revue Canadienne De Recherche Forestiere* 35, 1744-1753

Moore I.D., Grayson R.B. & Ladson A.R. (1991) Digital Terrain Modeling - a Review of Hydrological, Geomorphological, and Biological Applications. *Hydrological Processes* 5, 3-30

Nekola J.C. & White P.S. (1999) The Distance Decay of Similarity in Biogeography and Ecology. *Journal of Biogeography* 26, 867-878

Noss R.F. (1990) Indicators for Monitoring Biodiversity - a Hierarchical Approach. *Conservation Biology* 4, 355-364

Palmer M.W. & Dixon P.M. Small-Scale Environmental Heterogeneity and the Analysis of Species Distributions along Gradients. Journal of Vegetation Science 1[1], 57-65. 1990.

Pettitt A.N. & Mcbratney A.B. (1993) Sampling Designs for Estimating Spatial Variance-Components. *Applied Statistics-Journal of the Royal Statistical Society Series C* 42, 185-209

Philippi T. (2005) Adaptive Cluster Sampling for Estimation of Abundances Within Local Populations of Low-Abundance Plants. *Ecology* 86, 1091-1100

Preston F.W. The Commonness, And Rarity, of Species. Ecology 29[3], 254-283. 1948.

Rosenzweig M.L. (1995) *Species Diversity in Space and Time*. Cambridge University Press, Cambridge, UK.

Savage J.M. (1995) Systematics and the Biodiversity Crisis - There Is an Urgent Need for an Accelerated Accumulation of Knowledge About Biodiversity. *Bioscience* 45, 673-679

Stohlgren T.J., Binkley D., Chong G.W., Kalkhan M.A., Schell L.D., Bull K.A., Otsuki Y., Newman G., Bashkin M. & Son Y. (1999) Exotic Plant Species Invade Hot Spots of Native Plant Diversity. *Ecological Monographs* 69, 25-46

Thomas C.D., Cameron A., Green R.E., Bakkenes M., Beaumont L.J., Collingham Y.C., Erasmus B.F.N., De Siqueira M.F., Grainger A., Hannah L., Hughes L., Huntley B., Van Jaarsveld A.S., Midgley G.F., Miles L., Ortega-Huerta M.A., Peterson A.T., Phillips O.L. & Williams S.E. (2004) Extinction Risk From Climate Change. *Nature* 427, 145-148

Thompson S.K. (1991) Adaptive Cluster Sampling - Designs With Primary and Secondary Units. *Biometrics* 47, 1103-1115

Tobin P.C. (2004) Estimation of the Spatial Autocorrelation Function: Consequences of Sampling Dynamic Populations in Space and Time. *Ecography* 27, 767-775

Urban D., Goslee S., Pierce K. & Lookingbill T. (2002) Extending Community Ecology to Landscapes. *Ecoscience* 9, 200-212

Urban D.L. (2000) Using Model Analysis to Design Monitoring Programs for Landscape Management and Impact Assessment. *Ecological Applications* 10, 1820-1832

Vitousek P.M., Mooney H.A., Lubchenco J. & Melillo J.M. (1997) Human Domination of Earth's Ecosystems. *Science* 277, 494-499

Wagner H.H. & Fortin M.J. (2005) Spatial Analysis of Landscapes: Concepts and Statistics. *Ecology* 86, 1975-1987

Weakley A.S. (2006) *Flora of the Carolinas, Virginia, Georgia, and Surrounding Areas: Working Draft of 17 January 2006*. University of North Carolina Herbarium, North Carolina Botanical Garden, Chapel Hill, NC.

White, Peter S. The flora of Great Smoky Mountains National Park: An annotated checklist of the vascular plants and a review of previous floristic work. Research/Resources Management Report SER-55. 219. 1982. Atlanta, GA, Department of the Interior, National Park Service, Southeast Regional Office.

White P.S. & Wafford B.E. (1984) Rare native Tennessee vascular plants in the flora of Great Smoky Mountain[s] National Park. *Journal of the Tennessee Academy of Science* 59, 61-64

Whittaker R.H. Vegetation of the Great Smoky Mountains. Ecological Monographs 26[1], 1-80. 1956.

Wolock D.M. & Mccabe G.J. (1995) Comparison of Single and Multiple Flow Direction Algorithms for Computing Topographic Parameters in Topmodel. *Water Resources Research* 31, 1315-1324

Yeakley J.A., Swank W.T., Swift L.W., Hornberger G.M. & Shugart H.H. (1998) Soil Moisture Gradients and Controls on a Southern Appalachian Hillslope From Drought Through Recharge. *Hydrology and Earth System Sciences* 2, 41-49

Figures

Figure 4.1 A guide for generating samples using the two-stage stratified clustered protocol.

1 - Generate the ecological zipcodes for the study area (Chapter 2)

2 - Select the ecological zipcode digits to be used in 1st and 2nd stage stratification using mantel correlograms.

3 - Assign the number of plots per cluster and the maximum cluster radius based on the output of an ecological zipcode focal variety analysis for increasing radii. (Table 4.1)

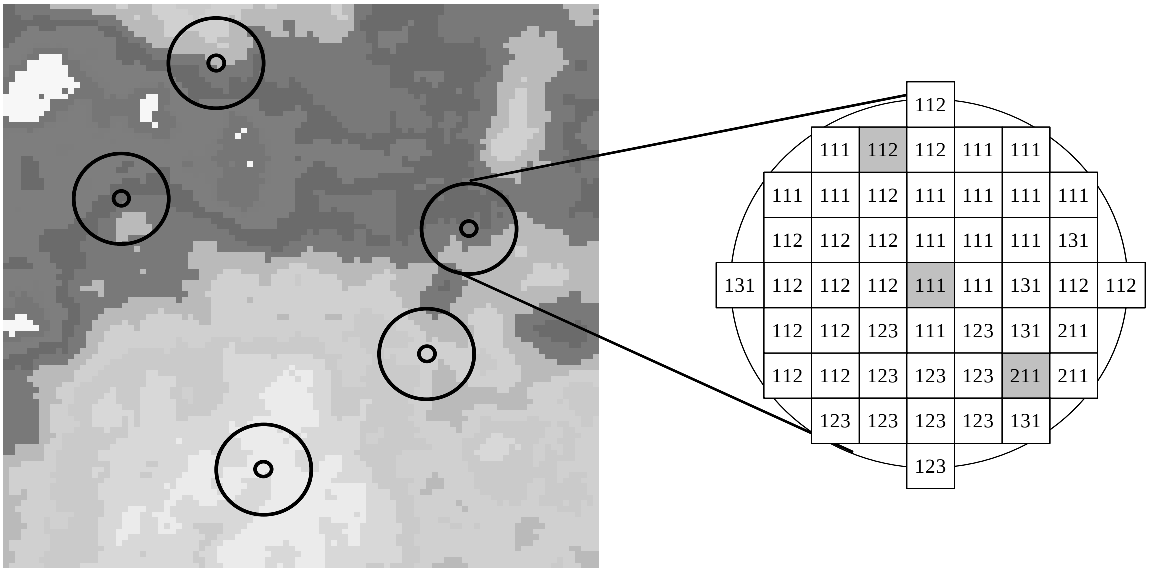
4 - Calculate the adjacency matrix of zipcode co-occurrence probabilities for the selected maximum cluster radius (Arc AML code available on request)

5 - Generate an accessibility model for the study area (Chapter 3)

6 - Test the bias introduced by accessibility weighting against the unweighted case

7 - Generate the final sample using the determined weighting.

Figure 4.2 Diagram of the two-stage stratified clustered sampling process using the ecological zipcodes. Cluster centers (left, small circles) are randomly selected for each ecological zipcode (left, shades of gray). Additional sites within the cluster are randomly selected from all sites within the maximum cluster radius (right, selected sites in gray), with the limitation the each site within a cluster be of a unique zipcode.



				112				
		111	112	112	111	111		
	111	111	112	111	111	111	111	
	112	112	112	111	111	111	131	
131	112	112	112	111	111	131	112	112
	112	112	123	111	123	131	211	
	112	112	123	123	123	211	211	
		123	123	123	123	131		
				123				

Figure 4.3 Spatial autocorrelation functions for the three variables used to stratify samples in the protocol: a) elevation, b) hillshade, and c) relative wetness (TCI). Observed autocorrelation (solid lines), lower confidence limits (dotted lines), and upper confidence limits (dashed lines) (α=0.95) for each variable are shown. Since variations in the elevation gradient occur at such large scales (>10km), this variable is not appropriate for stratification within clusters of sample sites. Relative wetness and hillshade, on the other hand, exhibit variation in spatial structure at scales appropriate for stratification within clustered sample sites (~200-400km).

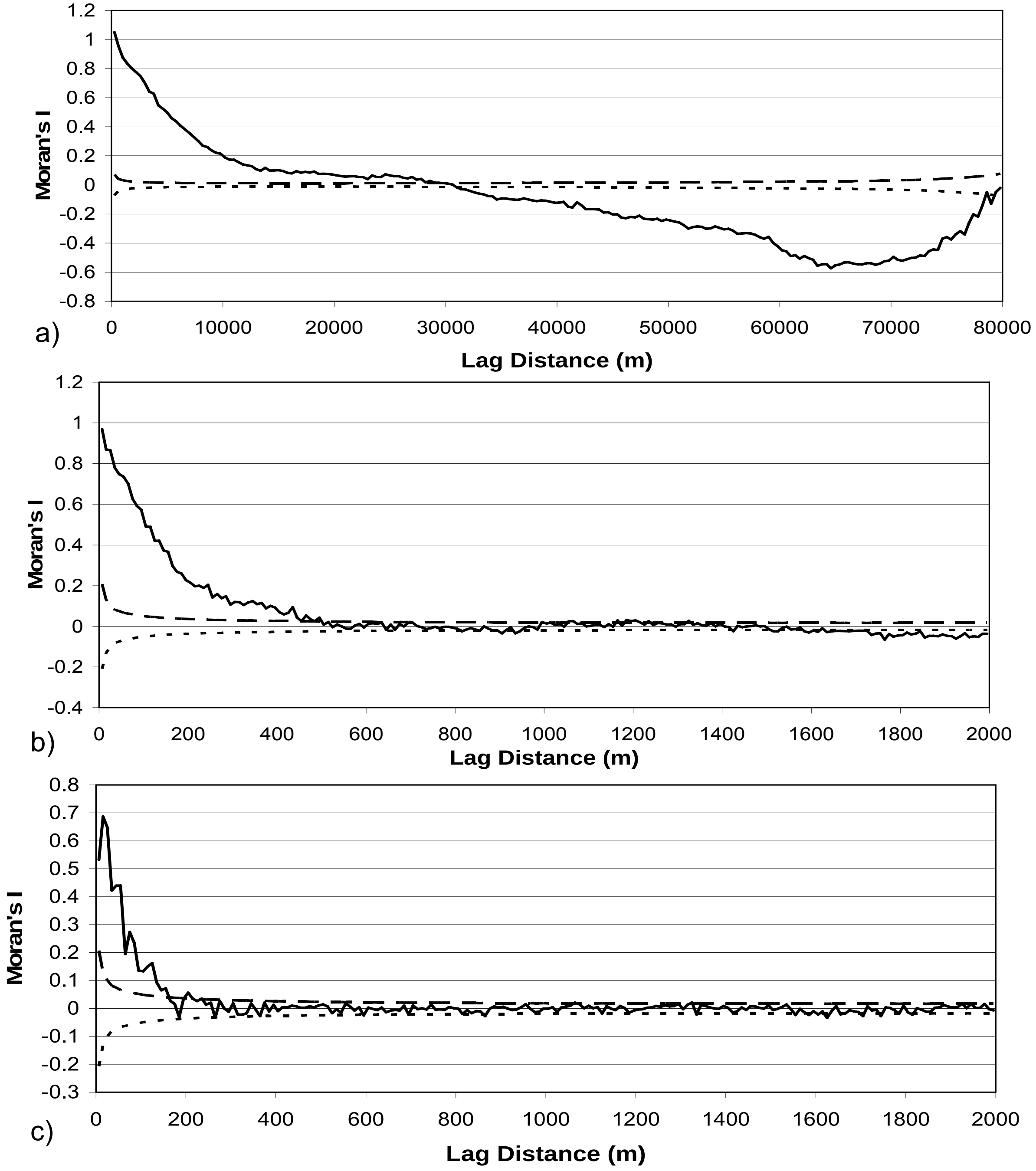


Figure 4.4 Two samples that use the two-stage stratified-clustered protocol for GSMNP. The first sample is unweighted for accessibility (open circles with dot). The second is weighted with z=16 in Eq. 1. (open circles with X). Though the joint inclusion probabilities of ecological zipcodes for the second sample differ significantly from the Park as a whole, its accessibility makes it more appealing.

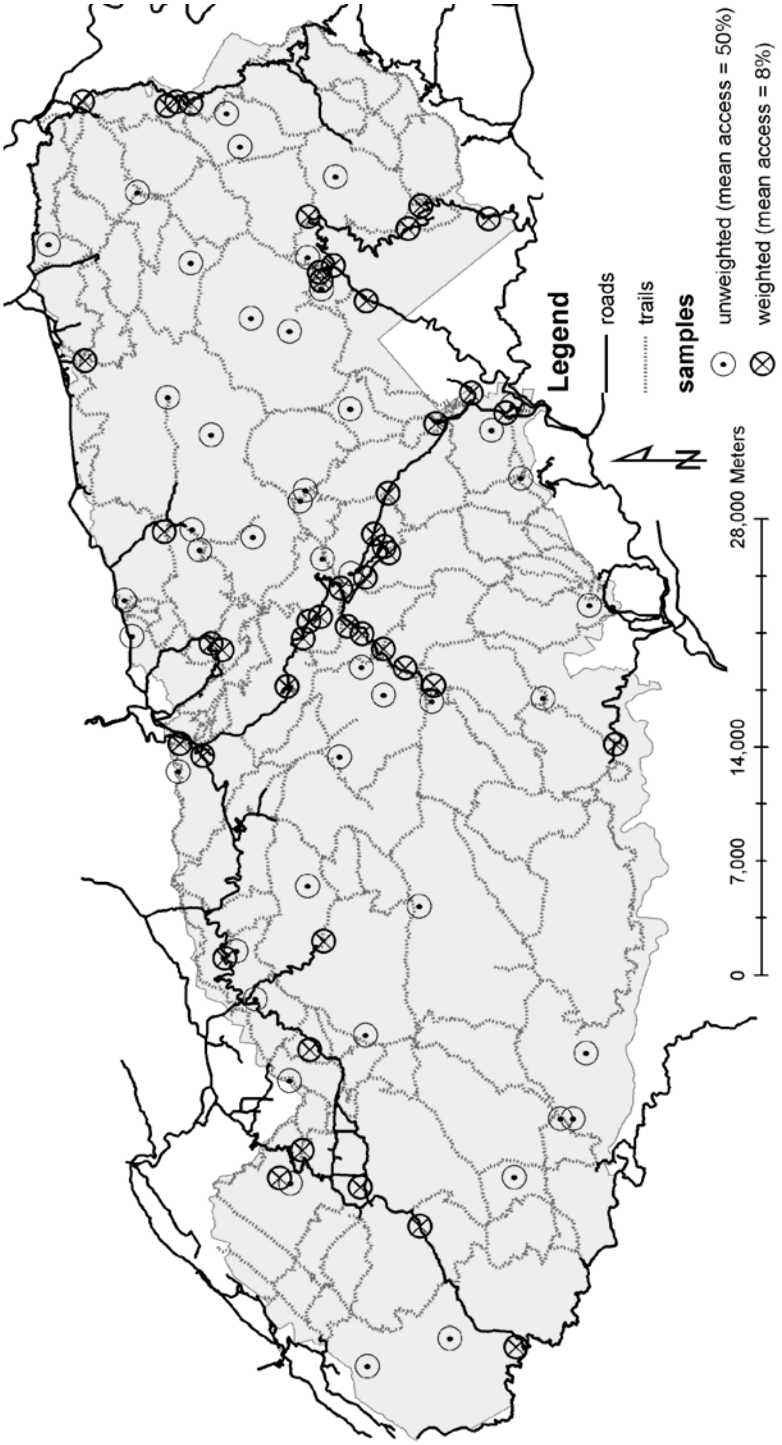


Figure 4.5 Selection probability for sites according to their accessibility via a simple accessibility weighting function (Eq. 1). Probabilities for different weighting exponents (*z* in Eq. 1) are shown. Mean accessibility for samples using different weighting exponent are given in Table 4.2.

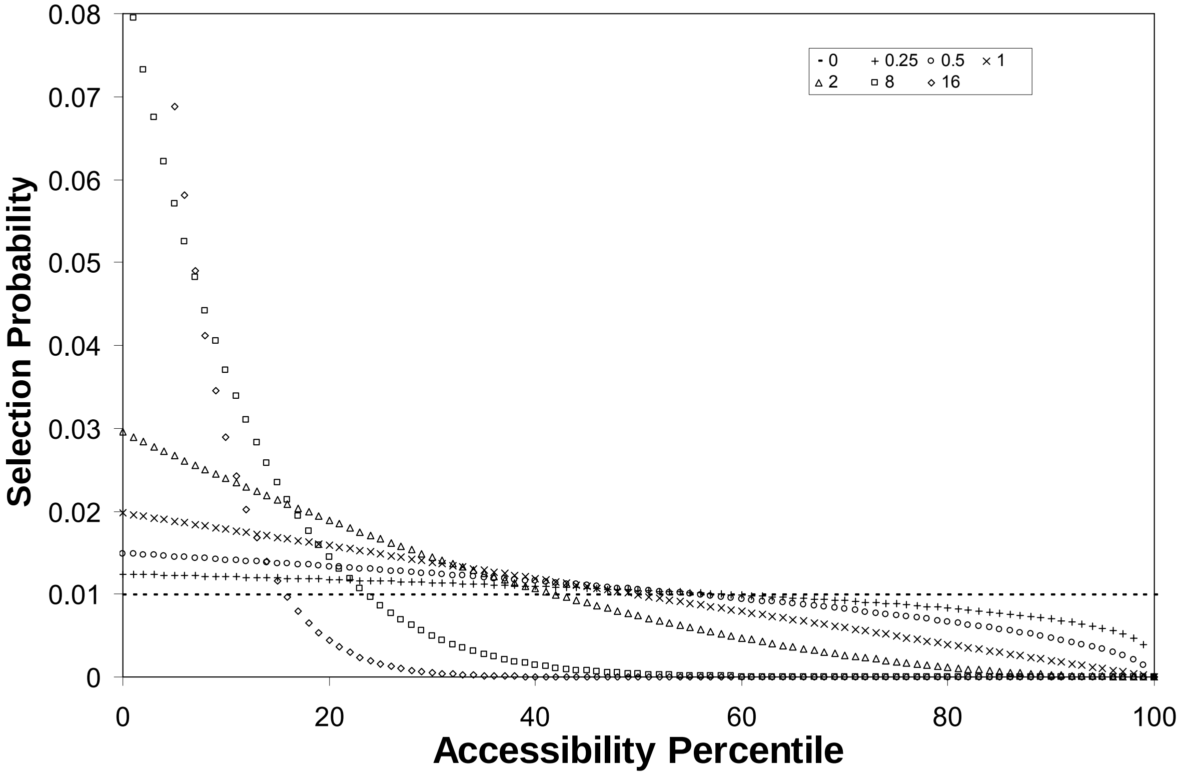


Figure 4.6 Distributions of the number of vegetation communities captured using different sample designs. Each distribution is based on 1000 random samples of 135 sites. The low number of vegetation communities captured using for random clustered sampling illustrate the weakness of clustered sampling designs. Stratification is a necessity for clustered sampling design. Further, selecting the appropriate stratification variables is also vital. Variety in vegetation communities is limited in the stratified samples, because the ecological zipcodes do not necessarily reflect subtle changes in vegetation communities.

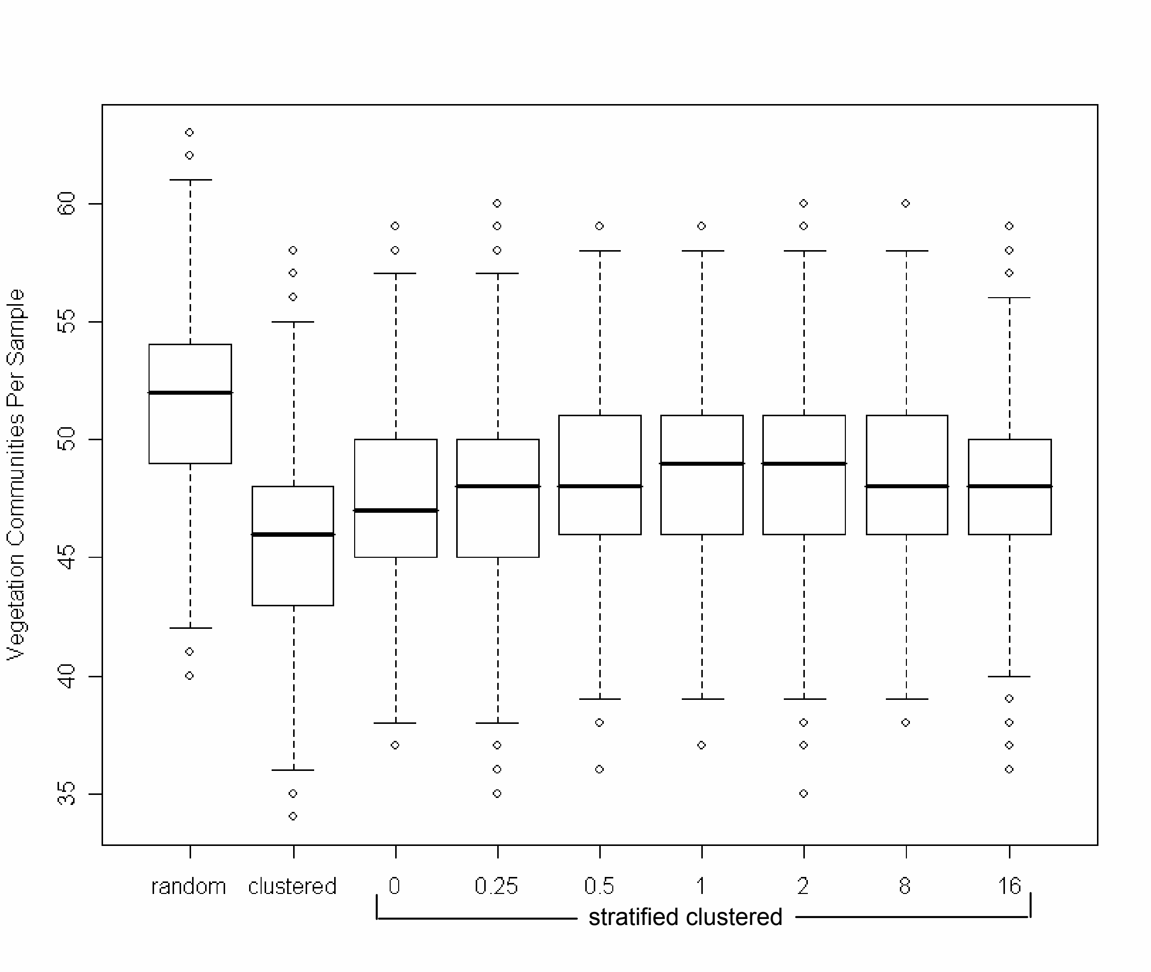
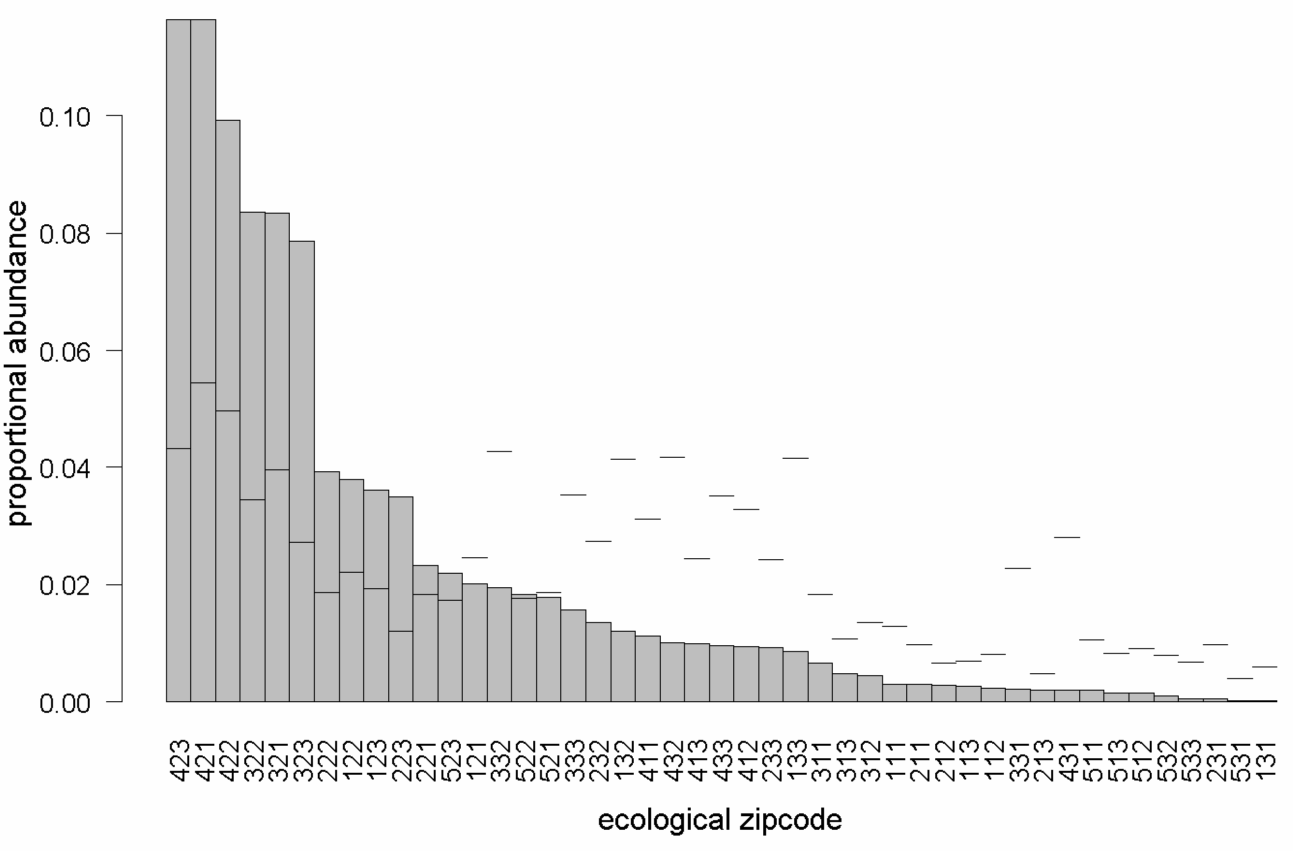


Figure 4.7 Observed and expected abundance of hemlock for each ecological zipcode. Expected values are for a random sample of hemlock. Observed values are the mean abundance from 1000 samples using the protocol (unweighted for accessibility). Rare hemlock environments are more frequently observed than would be expected by random, because of the stratification in the protocol. Since stratification occurs across ecologically important variables rather than just broad vegetation patterns, the protocol captures a greater diversity of environments for any given species.



Tables

Table 4.1 The proportion of sites within GSMNP having a variety of unique ecological zipcodes greater than or equal to a given number within circles of increasing radius. Variations in elevation within the zipcodes are not considered in these calculations because of the large distances over which elevation gradient is spatially autocorrelated. Window radius and zipcode variety pairs that encompass at least 95% of the Park (gray shading) are candidates for the cluster size and maximum cluster radius of the sampling protocol. Based on these results, I selected three sites per cluster and with a 200 m maximum cluster radius (shown in bold) for generating samples.

	P(Z>=N Zipcodes)								
	Circular window radius (m)								
N Zipcodes	100	150	200	250	300	350	400	450	500
2	0.936	0.984	0.993	0.995	0.995	0.995	0.996	0.996	0.996
3	0.737	0.919	**0.968**	0.984	0.987	0.988	0.989	0.990	0.990
4	0.453	0.766	0.896	0.956	0.972	0.979	0.982	0.984	0.985
5	0.200	0.526	0.738	0.878	0.928	0.951	0.964	0.970	0.974
6	0.058	0.281	0.510	0.726	0.832	0.892	0.928	0.947	0.957
7	0.009	0.094	0.253	0.483	0.642	0.755	0.839	0.889	0.918
8	0.001	0.017	0.071	0.196	0.327	0.457	0.591	0.698	0.776
9	0.000	0.001	0.004	0.019	0.041	0.074	0.124	0.184	0.248

Table 4.2 χ2 goodness-of-fit (GOF) test results for the joint probability distributions of ecological zipcodes and the distributions of zipcodes frequencies for samples with different accessibility weightings. The GOF test tests the null hypothesis that weighted samples derive from a population indistinguishable from that of unweighted samples. For even relatively small weighting (mean accessibility < 30%), this null hypothesis is rejected, illustrating the fact that biasing for accessibility introduces strong bias in environmental representativeness.

z	**Mean**	**GOF Joint Prob. Dist.**		**GOF Zipcode Freq.**	
	Access %	**X-sq**	**p**	**X-sq**	**p**
		(df = 989)	**(*** < 0.01)**	**(df = 44)**	**(*** < 0.01)**
0	49.70	----	----	----	----
0.25	45.06	869	0.997	68	0.011 ***
0.5	41.56	958	0.758	64	0.026 ***
1	35.99	1032	0.166	89	0.001 ***
2	28.57	1260	0.000 ***	250	0.000 ***
8	13.04	3033	0.000 ***	728	0.000 ***
16	8.47	4653	0.000 ***	806	0.000 ***

Chapter 5

Estimating Landscape-scale Species Richness: Reconciling Frequency- and Turnover-based Approaches

Abstract

One hypothesis for why estimators of species richness tend to underestimate total richness is that they do not explicitly account for increases in species richness due to spatial or environmental turnover in species composition. I analyze the similarity of a dataset of native trees in Great Smoky Mountains National Park, and assess the robustness of these estimators against recently developed ones that incorporate turnover explicitly: the total species accumulation method (T-S) and a method based on the distance decay of similarity. I show that the T-S estimator can give reliable estimates of species richness, given an appropriate grouping of sites. The estimator based on distance decay of similarity performed poorly. The results suggest that separating the biases associated with small sample sizes, while controlling for environmental heterogeneity can improve the richness estimates based on the turnover of species composition with distance.

Keywords

Beta-diversity, species richness estimation, distance-decay of similarity

Introduction

The assessment of species richness over landscapes requires ever-increasing sample effort to capture the ever rarer species (Fisher *et al.* 1943;Fisher *et al.* 1943;Preston 1948). Unless a landscape is completely surveyed, all samples are likely to fail to record a certain proportion of rare species. Consequently, in order to estimate total species richness (S_{true}) at a landscape-scale (10^3-10^6 ha), we are forced to extrapolate from incomplete surveys of total richness. There is a long history in ecology of estimating total species richness (S_{est}) from sample data, and a wide variety of techniques for doing so (Chao 2004). Numerous papers have assessed the bias and precision of any of a number of estimators for a given taxonomic group and location (for reviews see Cao *et al.* 2004;Walther & Moore 2005).

Species richness estimators rely upon the relationship between species richness and the accumulation of sample effort or area to estimate the total number of species for some unmeasured amount of sample effort either as time, area, or number of individuals sampled. The relationship between species richness and sample effort is summarized as a species accumulation curve where the x-axis is increasing sample effort or number of individuals, and the y-axis is increasing species richness. The order in which sample effort is accumulated profoundly affects the shape of the species accumulation curve. As a consequence, richness estimators rely upon a rarefied species accumulation curve which describes increases in species richness when sample effort is randomly accumulated (Gotelli & Colwell 2001)

Species richness estimators fall into two-broad categories: parametric and non-parametric. The former fit a function (typically a Michaelis-Menten) to the species accumulation curve (e.g. Jimenez-Valverde *et al.* 2006;Plotkin *et al.* 2000). The latter estimate species richness based on the frequency distribution of either species among sites (incidence-based) or the number of individuals of each species (abundance-based). These non-parametric estimators have more accurate estimates at small sample sizes than parametric ones and are typically preferred (Colwell & Coddington 1994).

An implicit assumption of non-parametric estimators is that sites are spatially homogeneous (Chazdon *et al.* 1998) and that the population from which individuals or species are drawn is stationary. As such, these estimators are typically used to estimate alpha-diversity (Whittaker 1972),which is local species richness where environment and other factors that control the species distributions are relatively constant. Recent evidence suggests, however, that these estimators are robust to spatial heterogeneity as long as sample coverage, the proportion of species observed relative to total species, is high (Brose *et al.* 2003;O'dea *et al.* 2006). In fact, many studies that assess the performance (as measured by bias, precision, and/or accuracy) of these richness estimators were based on samples that could be considered heterogeneous (e.g. Palmer 1990;Chiarucci *et al.* 2001).

In spite of their popularity, available non-parametric estimators typically underestimate species richness (Colwell & Coddington 1994;Chao 1984). One hypothesis for why non-parametric estimators tend to underestimate total richness is that they do not explicitly account for increases in species richness due to turnover in species composition across gradients. Such turnover of species composition across gradients, or beta-diversity (Whittaker 1972), is responsible for greater species richness of large areas than would be suggested by extrapolating from small areas and is the driving force behind changes of richness with scale (Rosenzweig 1995). Beta-diversity holds the key to understanding how richness scales from small areas that can be completely surveyed to large areas that cannot. Two recently published richness estimators (Harte *et al.* 1999;Ugland *et al.* 2003) explicitly use beta-diversity in calculating S_{est}. The T-S estimator of (Ugland *et al.* 2003) relies on grouping sites into ecologically meaningful subsets and integrating species richness estimates across different combination of these groupings. The estimator developed by Harte *et al.* (1999) builds upon the log-log relationship between richness and area (Arrhenius 1921) to calculate S_{est} based on the distance decay of compositional similarity (Nekola & White 1999). This method has received attention in the literature for estimating the richness of micro-organisms (Green *et al.* 2004;Horner-Devine *et al.* 2004) and landscape-scale vegetation (Krishnamani *et al.* 2004)

Here, I assess the robustness of these species richness estimators that explicitly incorporate species turnover relative to other, more popular estimators that do not. I demonstrate the sensitivity of turnover-based estimators to sample size, and sample coverage. I analyze the relationships between the similarity of environments, geographic locations, and species composition. Finally, I show how estimators based explicitly on turnover and those based on the frequency distribution of richness among sites actually estimate different processes on the same ecological template. I offer an alternative approach to species richness estimation that combines the benefits of frequency-based and turnover-based estimators.

Methods

Study Site

The study site uses the Great Smoky Mountains National Park (GSMNP, TN-NC, US) (Table 5.1) to illustrate and evaluate methods for estimating total species richness. The Park area is a little over 2000 km^2. 95% of the Park is forested. The Park runs E-W along a main ridgeline of the Southern Appalachian Mountains. Forests range from high-elevation Red Spruce-Frasier Fir (*Picea rubens-Abies fraseri)* and Northern Hardwood forest dominated by red maple, American beech, and yellow birch (*Acer saccharum*, *Fagus grandifolia*, and *Betula alleghaniensis*), to eastern hemlock (*Tsuga canadensis*) and pine-oak (*Pinus* spp.-*Quercus* spp.) forests on mesic and dry sites, respectively. At lower elevations, rich cove forests dominated by tulip poplar, American basswood and red maple (*Liriodendron tulipifera*, *Tilia americana var heterophylla*, and *Acer rubrum var rubrum*) are present. 129 native tree species (S_{true}) have been documented within the Park (165 including exotic species). The tree list on which these numbers are based is taken from a database on the vascular plants of GSMNP (Peter White, unpublished which was originally based on White (1982), and updated by Peter White and Jason Fridley (unpublished) with the help of GSMNP botanist Janet Rock). Nomenclature follows Weakley (2006)).

Data

The dataset used for this analysis is a compilation of vegetation studies conducted in GSMNP spanning roughly 30 years (Table 5.1). Though each study in the compilation had its own research questions, they all record the presence or absence of every vascular plant in an area of 1000m^2. I have limited this analysis to trees, because the actual number of species in the Park (S_{true}=129) is known to within a few species. I have further limited the species list to only native trees because of the rapidly changing richness of exotic species in the flora. 103 native tree species had recorded observations in the dataset. The final dataset consisted of 805 plots after removing those lacking native trees.

Analysis

I generated incidence-based species accumulation curves (S_{obs}) (Colwell *et al.* 2004) for the Park and parametric and non-parametric species richness estimates using the software package EstimateS (Colwell 2005). There are a large number of incidence-based richness estimators, so I limited this analysis to those that have been reported in the literature to perform best. Among the parametric equations, I used a fitted Michaelis-Menten (M-M)(Raaijmakers 1987;Colwell *et al.* 2004). The non-parametric estimators were the incidence coverage estimator (ICE) (Chao *et al.* 2000;Chazdon *et al.* 1998), Chao's incidence-based estimator ($Chao_2$) (Chao 1984;Chao 1987), and the second-order jack-knife estimator ($Jack_2$) (Burnham & Overton 1979;Burnham & Overton 1978;Smith & Vanbelle 1984;Palmer 1991). All these estimators were calculated based on rarefied species accumulation curves (S_{obs}) (Colwell *et al.* 2004).

Four different measures of site similarity were generated for the set of plots. For all pairs of sites, I calculated the Jaccard and Sorenson similarity of species composition, and the Euclidean distance and Euclidean distance of normalized environment. The environmental variables included in the similarity analysis were elevation, hillshade (azimuth 135, altitude 45), and relative wetness (as measured by the topographic convergence index (Moore *et al.* 1991;Wolock & Mccabe 1995;Yeakley *et al.* 1998). These variables taken together correspond to the important ecological gradients of energy flux, temperature, and radiation (Chapter 2). Since Bray-Curtis similarity is a rank-order measure, each environmental variable had equal weighting.

To analyze the effect that various gradients had on species richness accumulation, I generated species accumulation curves assembled by maximum dissimilarity. I used the three similarity measures outlined above for composition, geographic and environmental distance. The species accumulation curves were based on the mean of 100 randomizations. An initial site was selected randomly. Then, the site with the greatest dissimilarity was added sequentially to create a single randomization. For comparison, I also assembled sites based on maximum complementarity. Beginning with the richest site, I sequentially added the site with the most new species. Finally, I created the actual species accumulation curve. Sites were ordered by the date they were surveyed.

The two turnover-based estimators used in this analysis are the T-S estimator (Ugland *et al.* 2003) and the method of Harte *et al.* (1999). The T-S estimator relies upon groupings of similar sites. Given *n* groups, mean species accumulation curves are generated for all combinations of 1, 2...*n* groups. Each combination has a mean maximum richness. These maximum values are then fit to a log-linear species-area model. From this equation the total richness for a given area is calculated. I generated 10 groups three different ways: by species composition, environment, and by geographic distance. I used the method of partitioning around medoids (pam) (Kaufman & Rousseeuw 1990), a more robust version of k-means clustering, to assign group membership for each grouping variable.

The method of Harte *et al.* (1999) relies on the distance-decay of compositional similarity (Nekola & White 1999) to estimate species richness. In the absence of a predefined abbreviation, hereafter I refer to this method as DDS. The theory behind DDS builds upon Harte and Kinzig (1997). Beginning with the Arrhenius (1921), power-law species-area relationship (S=cA^z) where S is the number of species a is area and z is the slope of the log-log relationship, they derive the hypothesis that z is related to the slope of a log-log distance decay of similarity (Sorenson similarity) by the function z=-2d, where d is the slope of the log-log distance-decay. The slope of the log-log species-area relationship (z) is scale-dependent (Rosenzweig 1995), so the method of Harte *et al*. (1999) is only applicable across scales in which z is constant.

Analyzing the effect of sample size on the bias of the DDS estimator is not as straight-forward as that for other richness estimators. Any random subset of sites can have a unique geographic extent. The DDS estimator relies on samples whose extent is at least as great as the square root of the area to which the extrapolation is made. To correct for this I generated smaller samples by first selecting a pair of plots randomly whose distance were at least 40 km (roughly the square root of the Park area). Additional sites were added randomly up to the desired sample size. This gave a random subset whose extent was fixed. The DDS estimator for the Park area could be calculated on these subsets.

Results

Similarity in species composition decreased with increasing distance between sites. Since distance measures are strongly influenced by edge effects, the smallest linear extent of the Park (the N-S extent) set the maximum distance for comparison among sites (40km; Figure 5.1a). Similarity in species composition shows a more direct correlation with environmental similarity than distance (Figure 5.1b). The relationship between environment and compositional similarity seems to be explained well by a log compositional similarity and linear environmental similarity. Distance and environment show a log-linear relationship out to distances of about 20km, at which point environment and distance seem uncorrelated (Figure 5.1c).

Accumulating sites by maximal dissimilarity revealed some interesting trends (Figure 5.2). First, the actual accumulation order was more species-poor initially than a random curve would be. This is not particularly surprising considering that each project used in assembling the dataset had its own, typically community-specific, research question. The curve assembled by maximal compositional dissimilarity was also more species-poor, initially, than a random curve. After accumulating 10 sites, however, assembly by maximum compositional dissimilarity added species very quickly, and became the richest accumulation curve by 50 sites. For the first few sites, accumulating species by distance yielded the greatest species accumulation. From 5 sites to 50 sites, accumulation by maximum environmental distance produced the highest richness. After adding 100 sites by maximum Euclidean distance, additional sites followed the random curve.

Since the correlation between environmental distance and geographic distance begins to increase at distances greater than 40km, the DDS estimator can only be applied to those distance less than 40km (Figure 5.1d). The Park area is roughly $(45\text{km})^2$, so the DDS method can still be applied to estimate richness for the entire park. Harte *et al*. (1999) suggest a correction for rectangular areas that involves increasing the value of *z*. Since S_{est} from the DDS method were actually much larger than S_{true}, this correction was not applied.

The residuals of regressing log Sorenson similarity against log geographic distance out to 40km suggest that errors are not independent, and thus violate one of the important assumptions of linear regression (Figure 5.1d). This also corroborates other evidence suggesting that the distance decay of compositional similarity is log-linear as opposed to log-log (Nekola & White 1999).

The classic parametric and non-parametric estimators of species richness underestimated native tree species richness in GSMNP by about 20 percent on average (Table 5.2). The similarity-based estimators performed better or overestimated species richness. Contrary to other results in the literature (Ugland *et al.* 2003;O'dea *et al.* 2006) the T-S estimators performed the best out of all the estimators. The DDS estimator was actually the poorest performer of all the estimators, over estimating species richness by 35%. The overestimation of species richness was actually worse at small sample sizes, the complete opposite of all other estimators (Figure 5.3).

Discussion

Similarity in species composition, environment, and location

Given the strong relationship between compositional similarity and environmental similarity, it would seem that the distances at which they diverge from a strong relationship with distance would match. The fact that species composition remains correlated with distance beyond that which environment explains, suggests that either there exists important, yet unmeasured environmental variables that are spatially autocorrelated at distances greater than 20km or the signal of dispersal limitation in trees is present beyond 20km. More likely, there is a significant interaction between the two in the form of disturbance history causing this pattern.

This breakdown of the environmental gradient relationship with distance also illustrates the scale dependence of the species-area relationship at large scales as well as small scale (Rosenzweig1995). That is, that range of areas over which the slope of the log-log species area relationship (*z*) is constant has an upper bound as well as a lower bound. If the derivation of Harte *et al.* (1999) is correct, *z* is not constant from $1x10^3m^2$ to $2x10^5m^2$. This makes the method of Harte *et al.* (1999) even more restricted. This limitation is overcome in practice by successively integrating over small changes in area, where changes in *z* are small (Hortal *et al.* 2006). Species estimates derived from extrapolation between the plot and some larger area (smaller than the landscape) are used in the calculation of S_{true}. This procedure has two shortcomings. First, since the parameter being estimated (*z*) is exponentially related to species number, small errors in estimating z yield drastic errors in estimating species number. Second, these errors in the estimation of *z* are multiplicative when applied sequentially from small areas to large areas so that the cumulative error in S_{est} is much larger than that for any single extrapolation of S_{obs}.

Accumulation order matters

Richness estimators usually take into consideration only the randomized species accumulation curve. Site order does not matter in calculation. In practice, however, the randomized species accumulation curve can change drastically depending on the sample size and the accumulation order (Figure 5.2). For instance, if 100 sites of the original 805 were chosen based on environmental dissimilarity in the Park, richness would accumulate much faster than if sites were chosen randomly. The species accumulation curve for this subset would be steep compared to the random case, even after rarefaction. The species richness estimators would be higher (Gotelli & Colwell 2001). This occurs mainly because more environments would have been sampled, and the effective S_{true} would be much larger. Thus, sampling design matters hugely for species richness estimation so that even though site order is randomized for the richness estimators, the variety of sites in the sample still has a major impact on richness estimates.

For maximizing the gain in species richness in a sample, environmental dissimilarity among sites rather than distance among sites should be maximized (Figure 5.2). Though environment and distance covary, changes in environmental similarity become uncorrelated with distance at distances greater than 20km for

GSMNP (Figure 5.1b). This roughly corresponds to the extent of a single watershed within GSMNP. For GSMNP accumulating sites of maximally different environments within a watershed is the best course of action.

It is interesting that assembling sites by maximal compositional dissimilarity does not result in a steeper species accumulation curve, initially, than assembly by maximal environmental or Euclidean distance (Figure 5.2). This illustrates the weakness of compositional similarity measures and their application to beta-diversity. Sites with maximal dissimilarity are likely to be species poor. Stated another way, the probability of have no species in common is greater for sites with only 2 species than sites with 30 species. This bias has a profound impact on measures of similarity and the species richness estimators based on them. I address these impacts below.

Turnover-based estimators

The classic estimators of species richness (M-M, ICE, $Chao_2$, Jack2, and their abundance-based counterparts) have their origin in methods for extrapolating true population size from mark-recapture sampling (Chao 1984). These estimators attempt to estimate the number of unobserved species in an unknown *stationary* population (of species). As sample coverage (the proportion of the entire pool of species observed in the sample) increases, the accuracy of the richness estimator increases. Brose et al (2003) have shown that these metrics are relatively insensitive to environmental heterogeneity and spatial autocorrelation, so should perform well with samples that include a lot of heterogeneity and spatial autocorrelation, but they are relatively sensitive to sample coverage. In practice, though, increasing sample heterogeneity by adding new sites increases the species pool. Sample coverage is also decreased because the species pool grows faster than the proportion of species captured in the sample. So, while gradients may not affect estimator performance directly, they affect their performance indirectly by making the universe of species bigger as dissimilar sites are added.

Turnover-based estimators are plagued by the same problems, but in a different way. The T-S estimator performed better than expected based on the results of other studies. In previous studies, the estimator always overestimated S_{true} by a substantial amount. The errors in species estimation are mainly due to that fact that choosing the number of groups and the membership in each group is somewhat arbitrary. Group membership, in particular is important because all of the members are assumed to have the same species pool. In previous studies, group membership was decided based on either making equal-interval divisions across an ordination axis, or an environmental gradient, or across categorical habitat types. None of these methods ask the data which groupings are appropriate. Assigning group membership by non-hierarchical clustering (such as pam) allows natural groupings of similar sites based on the dataset. This is probably the reason for the better performance of this estimator. O'Dae and others (2006) suggest that T-S estimator is unnecessary because the species-area relationship is implicit in estimators of species richness. Nevertheless, richness estimators not based on turnover always underestimate richness due to turnover between sites. The key to incorporating compositional turnover explicitly in species richness estimation lies in separating the difference in species composition between sites that are due to environmental or ecological turnover, from those that are influences because sample coverage is too small.

The impact of sampling constraints on similarity

The increase of species with area beginning from the smallest scales and moving upward is a function of two processes. The first is ecology, that is the sum total of dispersal limitation, environmental heterogeneity, and competition. The second is sampling constraints. That is, richness at small scale is constrained by the number of individuals that can fit in a given area (Fisher *et al.* 1943). As area increases, the dominance of sampling constraints becomes less and the ecological forces become greater. Since both ecological and sampling processes covary with grain, increasing sample grain is not equivalent to increasing sample size, especially for plants. The more individuals that are sampled, the more environmental heterogeneity is present and the greater the species pool. The solution proposed by Harte *et al*. (1999) is to increase the sample grain until the log-log relationship of species and area is constant. This area is likely to be quite large for trees (much greater than 1ha), though for smaller organisms sample area is not as constraining (e.g. Green *et al.* 2004;Horner-Devine *et al.* 2004).

Similarity measures are also sensitive to sample coverage. Sample coverage is itself constrained by sample size. Smaller samples will systematically exhibit lower similarities than populations that have large sample sizes. Thus similarity measured in small samples is a biased estimator of the true similarity of two sites. More importantly, this bias is more pronounced for sites whose true similarity is high (Figure 5.1a). As an example, consider two sites of 10 species each. Each site displays complete evenness, so the selection probability of species is equal. They have all species in common, so their actual Sorenson similarity is 1. The mean Sorenson similarity of many random draws of one individual from each site would be 0.1 because the probability of drawing two individuals of the same species is 1/10. This occurs purely because sample size is too low. Now consider the opposite case in which no species overlap between the two sites. The mean Sorenson similarity for many random draws of a single individual will be the true similarity between sites – 0. Sites with high similarity and low sample size exhibit greater bias toward low similarities than sites whose similarity is actually low (Figure 5.4a).

The fact that estimates of similarity based on small samples show greater bias for sites of high similarity than low similarity has an important implication for the distance-decay of similarity relationship (Figure 5.4b). Since sites that are similar are more likely to be underestimated than sites that are very dissimilar, the effect of increasing the numbers of individuals per site would be to actually increase the rate of distance decay. As numbers of individuals per site increased, bias would decrease and similarity would increase. This increase would be greater for neighboring sites whose similarities are high, than for distant sites, whose similarities are low.

If the area of sites were increased, the slope of distance-decay of similarity would also increase. Thus, if sample size increases, then the DDS estimator should actually become worse. This is not necessarily the case, however, because increases in sample size at a particular location necessitate an increase in the environmental heterogeneity of the site, especially for plants. As discussed above, increasing sample size can actually decrease sample coverage because the number of species that could occupy a site, all else being equal, increases faster than the rate at which species are captured by the sample. As sample coverage decreases with increasing area, the similarity bias associated with small sample sizes returns. One solution to this problem might be to sum species numbers for a site through time (Fridley *et al. In Press*) but sites through time are subject to the same assumption of stationarity as sites through space. Namely, disturbance or the shifting mosaic of landscape patches can cause the species pools for any given site to change through time. Below, I describe an alternative to understanding the distance decay of similarity relationship and similarity-based richness estimators that accounts for the small sample effect without increasing the species pool?

Incorporating turnover in species richness estimators

The alternative to incorporating turnover into richness estimators involves combining the approaches of the point-estimators with turnover-based estimators. Point-estimators of stationary populations need to be used at scales and locations where they are appropriate (i.e. within relatively homogeneous sites). Employed in this way with abundance data, the true or asymptotic similarity between two sites can be estimated. Chao and others (2005) have developed just such a series of asymptotic similarity estimators that are analogous to the ones currently used in distance-decay of similarity analysis. This removes, or at least removes the estimated effect, of low sample size. Then, using these similarity estimates for each site, one could apply the approach of Harte *et al.* (1999), which relates the distance-decay of similarity to the accumulation of species with area. Unfortunately, that would only increase overestimation of S_{true} because the distance-decay of similarity relationships would become steeper. The Harte *et al.* (1999) model is based on a fundamentally-flawed power-law relationship between species, area and distance decay of similarity. Log-linear approaches relating species accumulation and distance-decay of similarity are likely alternatives, but derivations from first principles are not yet available.

Conclusion

Even in a world where species were all ecologically equivalent, their spatial distributions would form a complex mosaic on the landscape (Hubbell 2001). Understanding how patterns such as distance decay of similarity influence species richness patterns, can only serve to improve our understanding and estimates of richness at scales too large to be exhaustively sampled. My results suggest that estimators that incorporate compositional turnover can provide reasonable estimates of species richness. Estimators that separate sampling processes from ecological ones offer the most potential for advances in estimating species richness, since estimators that do not explicitly include ecological processes tend consistently underestimate species number. Further empirical and theoretical studies are needed to shed light on the interactions between similarity, richness, and sampling processes.

References

Arrhenius O. Species and Area. The Journal of Ecology 9[1], 95-99. 1921.

Brose U., Martinez N.D. & Williams R.J. (2003) Estimating Species Richness: Sensitivity to Sample Coverage and Insensitivity to Spatial Patterns. *Ecology* 84, 2364-2377

Burnham K.P. & Overton W.S. (1978) Estimation of Size of a Closed Population When Capture Probabilities Vary Among Animals. *Biometrika* 65, 625-633

Burnham K.P. & Overton W.S. (1979) Robust Estimation of Population-Size When Capture Probabilities Vary Among Animals. *Ecology* 60, 927-936

Cao Y., Larsen D.P. & White D. (2004) Estimating Regional Species Richness Using a Limited Number of Survey Units. *Ecoscience* 11, 23-35

Chao A. (1984) Nonparametric-Estimation of the Number of Classes in a Population. *Scandinavian Journal of Statistics* 11, 265-270

Chao A. (1987) Estimating the Population-Size for Capture Recapture Data With Unequal Catchability. *Biometrics* 43, 783-791

Chao A. (2004) Species richness estimation. In: *Encyclopedia of Statistical Sciences* (eds Balakrishnan N., Read C.B. & Vidakovic B.) Wiley, New York.

Chao A., Chazdon R.L., Colwell R.K. & Shen T.J. (2005) A New Statistical Approach for Assessing Similarity of Species Composition With Incidence and Abundance Data. *Ecology Letters* 8, 148-159

Chao A., Hwang W.H., Chen Y.C. & Kuo C.Y. (2000) Estimating the Number of Shared Species in Two Communities. *Statistica Sinica* 10, 227-246

Chazdon R.L., Colwell R.K., Denslow J. S. & Guariguata M.R. (1998) Statistical methods for estimating species richness of woody regeneration in primary and secondary rain forests of NE Costa Rica. In: *Forest biodiversity research, monitoring and modeling: Conceptual background and Old World case studies* (eds Dallmeier F. & Comiskey J. A.), pp. 285-309. Parthenon Publishing, Paris.

Chiarucci A., Maccherini S. & De Dominicis V. (2001) Evaluation and Monitoring of the Flora in a Nature Reserve by Estimation Methods. *Biological Conservation* 101, 305-314

EstimateS: Statistical estimation of species richness and shared species from samples. Colwell, R. K. 2005.

Colwell R.K. & Coddington J.A. (1994) Estimating Terrestrial Biodiversity Through

Extrapolation. *Philosophical Transactions of the Royal Society of London Series B-Biological Sciences* 345, 101-118

Colwell R.K., Mao C.X. & Chang J. (2004) Interpolating, Extrapolating, and Comparing Incidence-Based Species Accumulation Curves. *Ecology* 85, 2717-2727

Fisher R.A., Corbet A.S. & Williams C.B. The Relation Between the Number of Species and the Number of Individuals in a Random Sample of an Animal Population. The Journal of Animal Ecology 12[1], 42-58. 1943.

Fridley J.D., Peet R.K., van der Maarel E. & Willems J.H. (*In Press*) Integration of Local and Regional Species-Area Relationships from Space-Time Species Accumulation. *The American Naturalist*

Gotelli N.J. & Colwell R.K. (2001) Quantifying Biodiversity: Procedures and Pitfalls in the Measurement and Comparison of Species Richness. *Ecology Letters* 4, 379-391

Green J.L., Holmes A.J., Westoby M., Oliver I., Briscoe D., Dangerfield M., Gillings M. & Beattie A.J. (2004) Spatial Scaling of Microbial Eukaryote Diversity. *Nature* 432, 747-750

Harte J. & Kinzig A.P. (1997) On the Implications of Species-Area Relationships for Endemism, Spatial Turnover, and Food Web Patterns. *Oikos* 80, 417-427

Harte J., Mccarthy S., Taylor K., Kinzig A. & Fischer M.L. (1999) Estimating Species-Area Relationships From Plot to Landscape Scale Using Species Spatial-Turnover Data. *Oikos* 86, 45-54

Horner-Devine M.C., Lage M., Hughes J.B. & Bohannan B.J.M. (2004) A Taxa-Area Relationship for Bacteria. *Nature* 432, 750-753

Hortal J., Borges P.A.V. & Gaspar C. (2006) Evaluating the Performance of Species Richness Estimators: Sensitivity to Sample Grain Size. *Journal of Animal Ecology* 75, 274-287

Hubbell S.P. (2001) *The Unified Neutral Theory of Biodiversity and Biogeography*. Princenton University Press, Princeton, NJ.

Jimenez-Valverde A., Mendoza S.J., Cano J.M. & Munguira M.L. (2006) Comparing Relative Model Fit of Several Species-Accumulation Functions to Local Papilionoidea and Hesperioidea Butterfly Inventories of Mediterranean Habitats. *Biodiversity and Conservation* 15, 177-190

Kaufman L. & Rousseeuw P.J. (1990) *Finding Groups in Data: An Introduction to Cluster Analysis.* John Wiley & Sons, New York.

Krishnamani R., Kumar A. & Harte J. (2004) Estimating Species Richness at Large Spatial Scales Using Data From Small Discrete Plots. *Ecography* 27, 637-642

Moore I.D., Grayson R.B. & Ladson A.R. (1991) Digital Terrain Modeling - a Review of Hydrological, Geomorphological, and Biological Applications. *Hydrological Processes* 5, 3-30

Nekola J.C. & White P.S. (1999) The Distance Decay of Similarity in Biogeography and Ecology. *Journal of Biogeography* 26, 867-878

O'dea N., Whittaker R.J. & Ugland K.I. (2006) Using Spatial Heterogeneity to Extrapolate Species Richness: a New Method Tested on Ecuadorian Cloud Forest Birds. *Journal of Applied Ecology* 43, 189-198

Palmer M.W. (1990) The Estimation of Species Richness by Extrapolation. *Ecology* 71, 1195-1198

Palmer M.W. (1991) Estimating Species Richness - the 2nd-Order Jackknife Reconsidered. *Ecology* 72, 1512-1513

Plotkin J.B., Potts M.D., Yu D.W., Bunyavejchewin S., Condit R., Foster R., Hubbell S., Lafrankie J., Manokaran N., Seng L.H., Sukumar R., Nowak M.A. & Ashton P.S. (2000) Predicting Species Diversity in Tropical Forests. *Proceedings of the National Academy of Sciences of the United States of America* 97, 10850-10854

Preston F.W. The Commonness, And Rarity, of Species. Ecology 29[3], 254-283. 1948.

Raaijmakers J.G.W. (1987) Statistical-Analysis of the Michaelis-Menten Equation. *Biometrics* 43, 793-803

Rosenzweig M.L. (1995) *Species Diversity in Space and Time*. Cambridge University Press, Cambridge, UK.

Smith E.P. & Vanbelle G. (1984) Nonparametric-Estimation of Species Richness. *Biometrics* 40, 119-129

Ugland K.I., Gray J.S. & Ellingsen K.E. (2003) The Species-Accumulation Curve and Estimation of Species Richness. *Journal of Animal Ecology* 72, 888-897

Walther B.A. & Moore J.L. (2005) The Concepts of Bias, Precision and Accuracy, and Their Use in Testing the Performance of Species Richness Estimators, With a Literature Review of Estimator Performance. *Ecography* 28 , 815-829

Weakley A.S. (2006) *Flora of the Carolinas, Virginia, Georgia, and Surrounding Areas: Working Draft of 17 January 2006*. University of North Carolina Herbarium, North Carolina Botanical Garden, Chapel Hill, NC.

White, Peter S. The flora of Great Smoky Mountains National Park: An annotated checklist of the vascular plants and a review of previous floristic work. Research/Resources Management Report SER-55. 219. 1982. Atlanta, GA, Department of the Interior, National Park Service, Southeast Regional Office.

Whittaker R.H. Evolution and Measurement of Species Diversity. Taxon 21[2/3], 213-251. 1972.

Wolock D.M. & Mccabe G.J. (1995) Comparison of Single and Multiple Flow Direction Algorithms for Computing Topographic Parameters in Topmodel. *Water Resources Research* 31, 1315-1324

Yeakley J.A., Swank W.T., Swift L.W., Hornberger G.M. & Shugart H.H. (1998) Soil Moisture Gradients and Controls on a Southern Appalachian Hillslope From Drought Through Recharge. *Hydrology and Earth System Sciences* 2, 41-49

Figures

Figure 5.1 Distance decay of similarity for species composition and environment for 1000m^2 vegetation plots in GSMNP. Circles are the mean value for each of 10 equal-sized groups of distances along the abscissa. Bars show the 1 standard deviation above and below the mean for each group. Comparisons shown are: a) log of compositional similarity (Jaccard's) by linear distance. This is the standard distance decay of similarity plot (*sensu* (Nekola & White 1999)); b) log of compositional similarity by the log of environmental similarity (Bray distance); c) linear environmental similarity versus linear distance; And d) log of Sorenson similarity versus the log of linear distance, whose linearly regressed slope is equal to -2*z* where *z* is the exponent of the Arrhenius (1921) species-area function (*sensu* (Harte *et al.* 1999)).

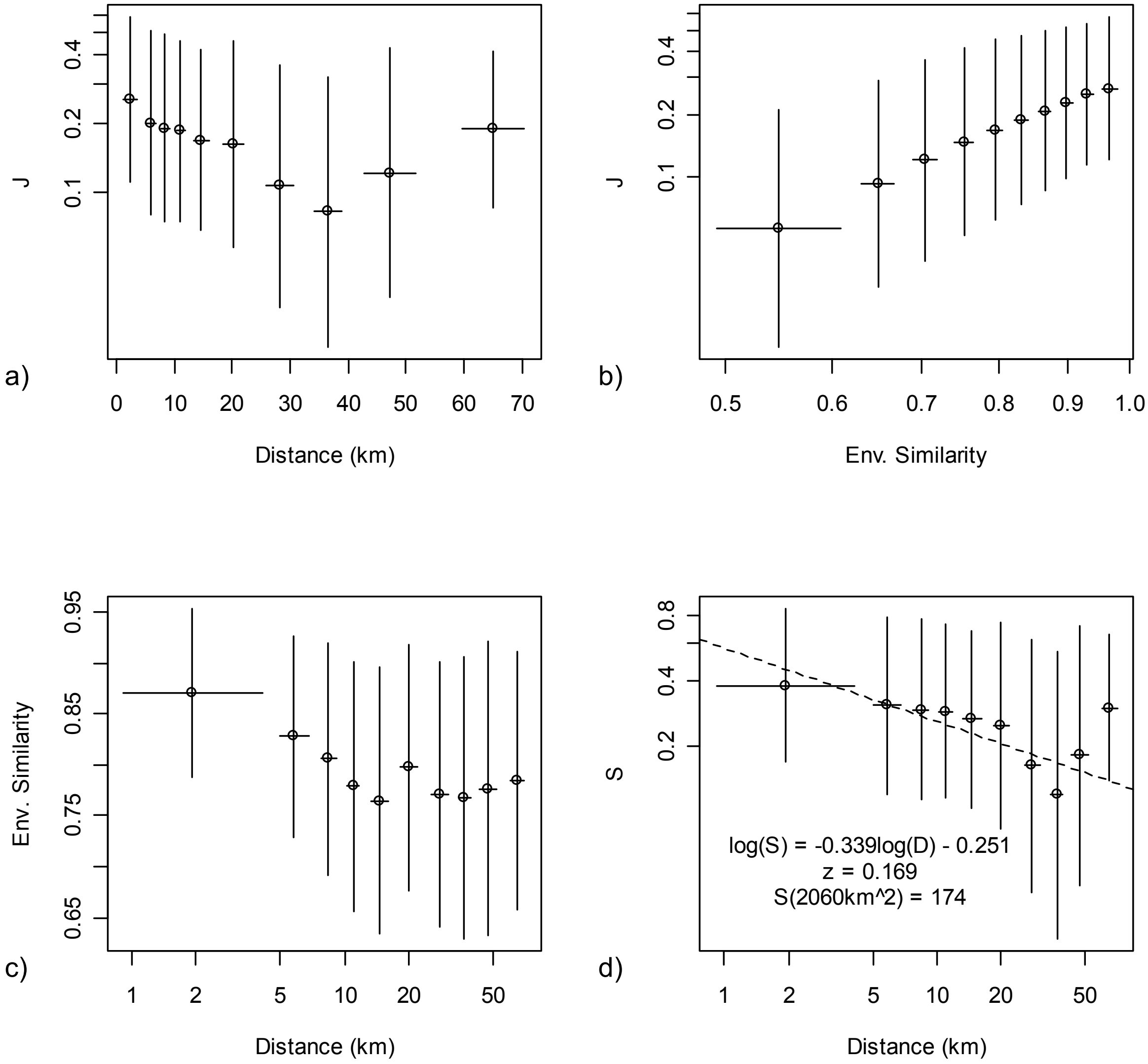


Figure 5.2 Species accumulation curves for sites accumulated by maximum dissimilarity in species composition (Jaccard's), Euclidean distance, and environment (Bray-Curtis). For comparison, the random case (standard species accumulation curve), the maximum case in which sites are accumulated by number of species they add to the total richness, and the actual accumulation order of plots through time are shown.

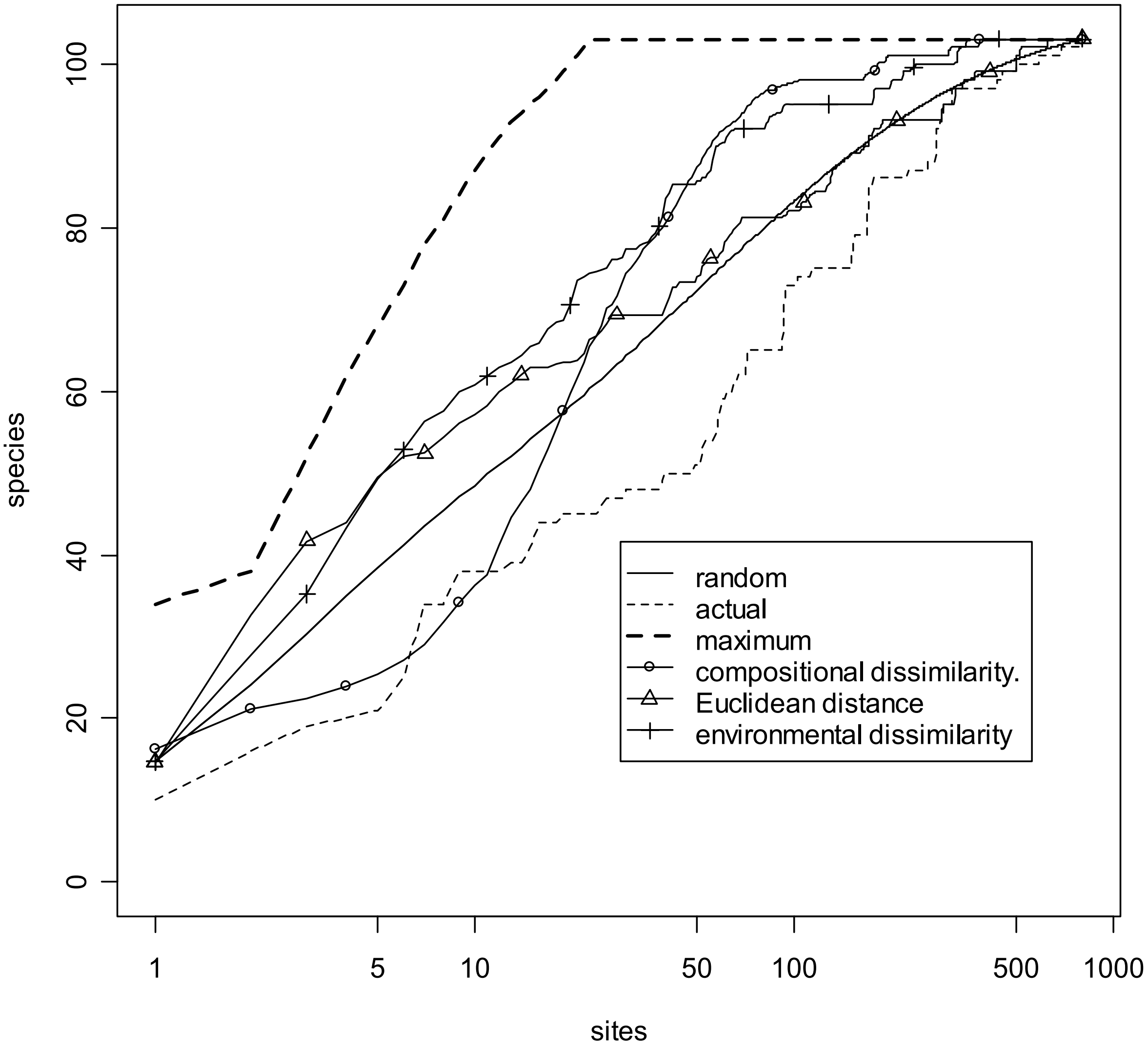


Figure 5.3 Estimated species richness of trees in GSMNP with increasing sample size for a) parametric and non-parametric estimators and b) the turnover-based estimator of Harte *et al*. (1999). All estimators decrease bias with sample size, but the turnover based estimator tends to overestimate richness at small sample sizes, while other estimators underestimate richness.

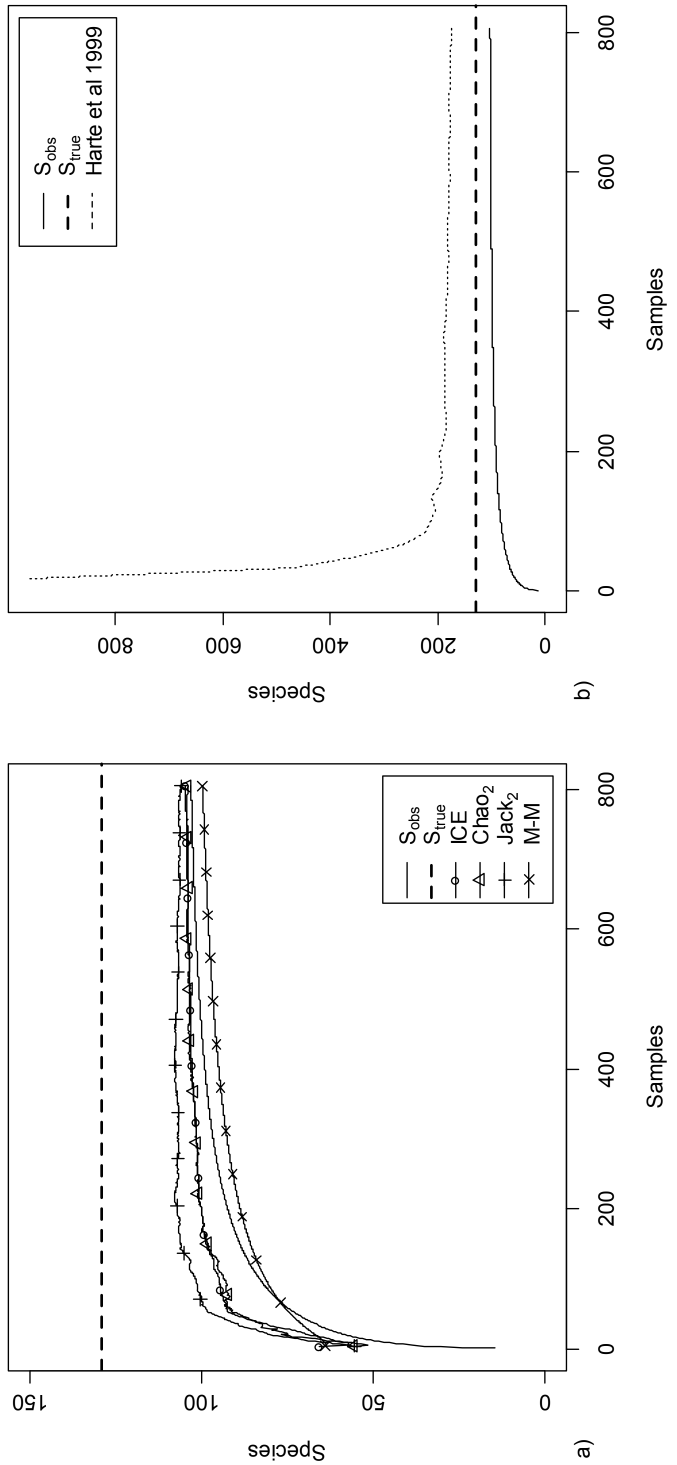
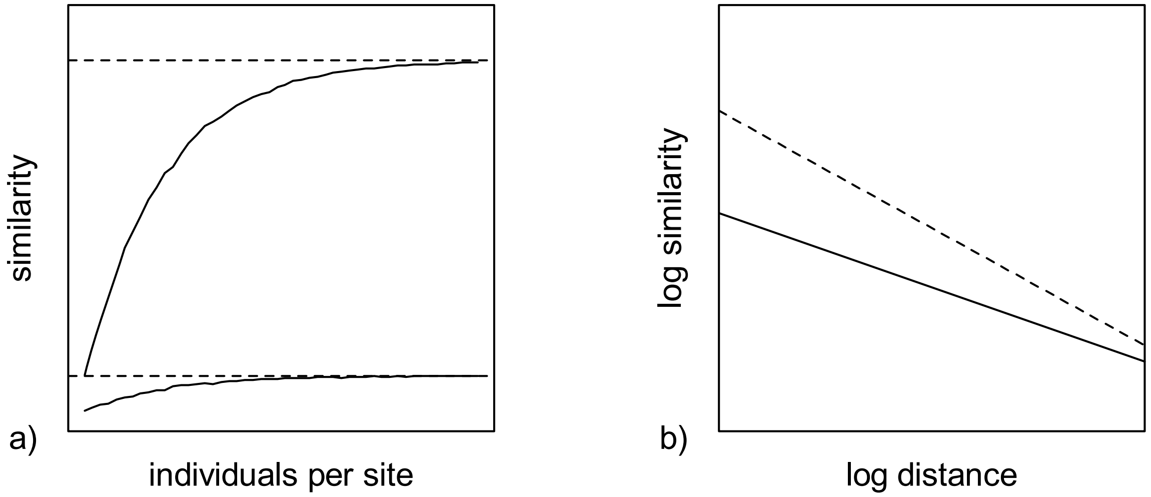


Figure 5.4 A hypothetical example showing how bias in similarity measures at small sample sizes influences the distance decay of similarity. a) Two sites are shown: one with high actual similarity and one with low actual similarity (dashed lines). Estimated similarities based on samples of individuals from each species pool (dark lines) are biased toward lower similarities with lower sample size. This bias at small sample sizes is greater for sites with high similarity than low similarity. b) The implication for distance-decay relationship is that as number of individuals per sample increases similarity increases faster for similar sites than dissimilar ones. This results in a steeper distance-decay relationship (dashed lines) than observed for samples with fewer numbers of individuals per site (dark lines).



Tables

Table 5.1 Description of the study area and dataset used to generate species richness estimates.

Study Area			
Great Smoky Mountains National Park			
Location	**35°35' N**	Extent	
	83°33' W	N-S	**39 km**
Area	**2062 km^2**	E-W	**87 km**
Native Tree Richness (S_{true}) **129 Species**			
Sample			
Time Period	**1976-2004**	Inter-plot distances	
Number of Plots	**805**	Mean	**23.9 km**
Area Per Plot	**1000m^2**	Median	**17.4 km**
Richness Per Plot		Maximum	**86 km**
Mean	**14.7 Species**		
Maximum	**34 Species**		
Observed Native Tree Richness (S_{obs}) **103 Species**			

Table 5.2 Richness estimates and performance for a variety of estimators. S_{true} is 129 species. Contrary to previous studies, T-S estimators out-performed any other estimator. The method of Harte *et al*. (Harte *et al.* 1999) was the poorest performer.

Estimator	S_{est}	S_{est}/S_{true}	Model
S_{obs}	103	0.80	
Parametric			
Michaelis-Menten	97	0.75	
Non-parametric			
ICE	105	0.81	
$Chao_2$	104	0.81	
$Jack_2$			
Similarity-based			
T-S 1) Environmental Distance	121	0.94	14.3ln(a)+15.7
2) Geographic Distance	125	0.97	12.8ln(a)+24.2
3) Compositional Similarity	122	0.95	13.7ln(a)+16.5
Harte et al. 1999	174	1.35	$(14.7)(a/0.001)^{0.17}$

Chapter 6

Conclusion: Bridging the gap between pattern, process, and sampling

Fundamental contributions

From fundamental property to unknown quantity, the concept of species richness is a treasure trove for ecological inquiry. The complexities arising from species responding individualistically to ecological gradients and the multiple spatial and temporal scales at which pattern is exhibited result in an ecological concept that will provide research potential for years to come. Beyond my study of species richness patterns, basic questions remain, such as "How many species are there?", or "What is the relative importance of environment, history, and species interactions in creating and maintaining patterns of species richness?" While we are unlikely to find precise answers to these questions, my work has provided methodology and important insights towards a more complete understanding of species richness patterns. My work helps close the gaps between ecological processes and richness patterns, between human influences and richness patterns, between sampled and true richness patterns, and between local richness patterns and landscape ones.

First, my work offers a framework for with respect to assess the relationship between species richness and environment. Many different environmental variables can impact species distributions. Distilling this multivariate environmental space into units that are both ecologically important and coarse enough to stratify samples provides a vital and parsimonious link between the spatial distributions of species and the ecological processes that create them. I have presented an objective and (perhaps most importantly) iterative strategy for classifying environment, and I have shown how this classification can be used to develop a sampling design to assess and monitor species richness patterns. This strategy begins from ecological first principles and can be iteratively improved as our knowledge of how species relate to environment increases. This strategy lends an ecological rigor to biodiversity assessment. It also lends flexibility to assessments; allowing broader scope and application of collected data.

Second, my results have shown that landscape structure controls not only patterns of richness, but also patterns of human interaction with that richness. This has consequences for ecological inference, since all inferences about species richness are based on sample data. I developed a model of human accessibility that estimates the energetic cost of walking through a landscape. I have shown a strong bias for accessibility to be present in vegetation survey data spanning 40-years in Great Smoky Mountains National Park. While this, in and of itself, is no cause for alarm, I have also shown the more disturbing result that important communities in the Park are under-sampled relative to their abundance because they are inaccessible. These inaccessible communities may have a distinctive species composition because of their distance from sampled sites as well as decreased human disturbance.

Third, I have shown that sample data with even small biases, such as those for accessibility, can suggest patterns of richness that are markedly different from the true landscape patterns of richness. The poor representation that comes from biased samples can be alleviated, however, if samples are stratified by important ecological variables. Of course, choosing the appropriate variable for stratification is paramount to capturing variation in species composition. I have shown that stratification of samples using ecological zipcodes is effective in capturing a wide variety of environments within broad community patterns. I have re-emphasized the importance of clustered sampling for the assessment of spatial covariation of environment and species distributions as well as separation of environmental effects from other spatially structured processes such as dispersal. My results, however, show that clustered samples must be paired with an appropriate stratification scheme in order to capture the equivalent breadth of representation that random samples offer. Finally, I show that while efficiency is increased by weighting samples toward more accessible locations, sampling more frequently in fewer areas may offer the best balance between sampling efficiency and environmental representativeness.

Finally, I have established an important link between estimates of species richness that focus on stationary, local populations and those that incorporate information about species turnover for landscape-scale estimates. My results conflict with those of other studies. An estimator based on grouping sites of high similarity outperformed classic estimators relying upon stationarity assumptions and based on distance decay of compositional similarity. I showed that, while theoretically distance decay of similarity measures should provide reasonable estimates, compositional similarity is strongly biased by sample size. This bias is more pronounced with increasing similarity and decreasing distance. Further, this bias is not removed by just increasing sample size, mainly because environmental heterogeneity and sample size covary. I suggest alternatives that incorporate asymptotic compositional similarity estimates at local scales.

Future Directions

In addition to contributing to our understanding of species richness, the results I present also create an opportunity to address new research questions. This dissertation has the potential to stimulate new directions for researching the interactions between species richness patterns, the ecological processes that create and maintain these patterns, and the samples that help us connect pattern to process. Here, I outline some of these possibilities focusing on 5 major questions stemming from my research. 1) Given their restricted abundance, how can rare species be efficiently included in biodiversity monitoring programs? 2) Given the iterative nature of the biodiversity assessment strategy I have presented, what is the best way to incorporate lessons learned from previous studies into current sampling? 3) Is the model of human accessibility I have developed an accurate portrayal of how people move on landscapes? 4) Can we use information gained from models of accessibility to separate effects of human disturbance on species richness patterns from other spatially autocorrelated effects? And, 5) what are the theoretical and empirical connections between the species area relationship and distance decay of compositional similarity.

First, the results from my study suggest that while samples stratified by environment capture broad patterns of vegetation, they may not capture rare species and communities more effectively than random samples. If the habitat for a rare species is also rare on the landscape then stratification by environment would tend to capture these species. However many species are rare in spite of the area of associated habitat on the landscape. These would not be captured effectively by environmental stratification. This is an obvious problem for both biodiversity monitoring and species richness estimation. The difficulty in capturing rare species arises from the fact that they are often absent from seemingly suitable habitats. Thus, rare species cannot be efficiently observed merely through stratification by important ecological variables. Whether it is dispersal limitation, life history, or range reduction from human disturbance, the causes of rarity often lie in processes not easily assessed with available data. Given these limitation, researchers are left with two choices for capturing rare species in biodiversity assessments: build upon historical observations of rare species, or make samples adaptable so that when rare species are encountered more samples can be taken at those locations. In Chapter 3, I suggested that incorporating adaptive clustering into stratified-clustered designs is a good solution to capturing rare species. The more we understand about both the factors that predict rarity and sampling designs that maximize both abundant and rare species, the more effective biodiversity monitoring programs will be.

Second, I presented in chapter 2 a protocol for assessing biodiversity that can be iteratively improved through time. Deciding the location and frequency of additional samples, however, is not simple. One approach is to use classification and regression trees (CART) to determine which portions of an environmental gradient exhibit the greatest changes in species composition. It is unclear, however, how to proceed when the set of important variables change through time as our understanding of what controls patterns of species richness improves. Further, when sample sites are added, the balance between representation and sampling efficiency must be re-evaluated. With the progression of mobile technology, it is possible that soon these re-evaluations and iterative sample improvement can take place "on-the-fly" and in the field. Decisions about efficiency and representation then become even more pronounced. Finally, improving samples through time reveals a tradeoff that exists between biodiversity assessment and biodiversity monitoring. That is, with given a constrained amount of sample effort, one can allocate that effort between resurvey of established sites and the addition of sites. An appropriate balance between these two options must be explored through simulation and experimentation.

Third, I have not yet validated the accuracy of the accessibility model presented in Chapter 3 using field tests. Currently, the only estimates of the energetic costs associated with crossing vegetation come from a 30 year-old study that poorly defines many aspects of landscape structure. By selecting locations of varying accessibilities, traveling to those sites, and recording oxygen consumption, I could improve the reliability of the model. While general estimates of accessibility offered by the model are applicable for illuminating stark contrasts, such as the accessibility bias present in samples, improved accuracy as a result of field validation would be helpful in more subtle contrasts such as the presence or absence of rare species.

Though I have here applied the model of human accessibility to the distribution of samples, and vegetation communities, the applications of this model stretch far beyond those I have addressed. This model can be used to measure the intensity of human disturbances that are correlated with accessibility. Of particular importance is the harvesting of rare species. From butterfly collectors to ginseng hunters, anthropogenic harvesting can have a drastic impact on population viability and species distributions. This complicates the prediction of species distributions, species richness estimates, and understanding compositional turnover. By correlating harvest species distributions with an accurate model of accessibility, I can estimate the strength of human harvesting on these species. By incorporating those effects with other important environmental variables, I can also make accurate distribution predictions for them that include historic effects of harvesting.

Finally, in Chapter 5 I have outlined a new strategy for incorporating beta-diversity into landscape-scale estimates of species richness. This estimate relies upon the asymptotic similarity of species composition between sites and the log-linear relationships between composition, area, and distance. This estimate was not implemented here for two reasons. The first lies in limitation of animal and plant surveys. It is difficult to distinguish individuals in plant surveys, so the relative abundance distributions of plants are non-intuitive. For animals, species-area relationships are non-intuitive because of their vagility. I plan to address this problem for plants by using cover estimates and for animals by using species whose home range can be captured with a single sample. The second reason the estimate was not implemented here is that theoretical connections between log-linear species area relationships and log-linear distance decay of similarity do not exist yet. My results suggest that power-law relationships between species, area, and distance, do not explain natural patterns jointly as log-linear relationships could. Thus, new theory is needed that connects local scale species-area relationships with turnover in an exponential framework. This is an area for future research that offers significant potential for advances in ecological theory.

Conclusion

There is perhaps no more compelling reason to study patterns of species richness than the fact that, on a global scale, richness is decreasing at a phenomenal rate. Global climate change, exotic species invasions, and landscape fragmentation are just a few of the anthropogenic changes to our planet that are impacting species richness. This dissertation adds a body of theoretical and empirical tools to our search for understanding of the ecological processes that cause species richness patterns and the methods we use to uncover these relationships. It is my hope that these advances will move ecologists and conservationists closer to an understanding of species richness that spans from theory into practice. Only when we understand these connections between biodiversity patterns and ecological processes will we be able to effectively conserve them.