



I can't define the niche but I know it when I see it: a formal link between statistical theory and the ecological niche

William Godsoe

W. Godsoe (williamgodsoe@vandals.uidaho.edu), Natl Inst. of Mathematical and Biological Syntheses, Univ. of Tennessee, Knoxville, TN 37996, USA.

The niche is one of the most important concepts in ecology. However, there has been a persistent controversy on how to define, measure, and predict the ecological niche of an organism. Here I argue that these problems arise in part because the niche is defined by the set of all possible environments, many of which do not exist in nature. A complete description of the niche would require knowledge of a large number of environments that do not exist in nature. Given this, I propose that ecologists should not focus on the niche itself but instead on determining if a particular environment is a part of the niche. I then demonstrate that such an analysis has a natural interpretation as an estimate of the probability that an environment is suitable and that either experimental investigations or analyses of presence data can estimate this quantity. Depending on the way that resources interact to shape environmental requirements, the probability that an environment is a part of the niche behaves like published descriptions of causal inferences and distribution models. However, in some cases the probability that an environment is suitable can be strongly influenced by unmeasured aspects of the environment. When this is true, experimental and distribution models have complimentary strengths and weaknesses.

One of the most important goals in ecology is to understand organism's ecological niche. Unfortunately, while the niche has been an important conceptual tool in ecology (May and MacArthur 1972, Chase and Leibold 2003) it has proven very difficult to define and measure (Colwell and Futuyma 1971, Whittaker et al. 1973, Hurlbert 1984, Peters 1991, Araujo and Guisan 2006, Hooper et al. 2008). In particular, it has been difficult to explicitly connect our abstract understanding of this concept to the particular observations we make about environmental requirements. Many studies have been motivated by different definitions of the term and use empirical methods ranging from experimental manipulations to small-scale observational studies and analyses of distribution records. Moreover, all definitions of the niche are confounded by the difficulty of interpreting our results in the face of the many aspects of the environment that we cannot measure. This has made it very difficult to explain what many studies of ecological niches actually do, and in turn to compare the strengths of different methodologies. To address these problems, I propose explicitly connecting a clear definition of the niche with statistical inference.

An example of the ambiguity in how empirical investigations make inferences about the niche of an organism is the current controversy between the use of ecological niche models and causal inferences. Several reviews have argued that models using presence data to predict environments where an organism may occur provide a cheap and effective

estimate of the niche of an organism (Pearson and Dawson 2003, Soberon and Peterson 2005, Soberon 2007). Other authors have argued that such models have some utility, but at best describe characteristics of the environments that an organism occupies (Kearney 2006). Instead they argue that the best way to make inferences about a species' niche is to develop a model of its underlying abiotic requirements (its fundamental niche). Such investigations can use either experimental methods (Hooper et al. 2008) or models that incorporate physiological mechanisms (Kearney and Porter 2004) to determine which aspects of the abiotic environment will cause an environment to be unsuitable.

For both statistical and biological reasons, this distinction is unsatisfying. Biologically, it is not clear that causal understanding is superior to inferences developed from correlative data in every case. For example, why is it necessarily more valuable to know that an organism cannot tolerate temperatures that simulate high elevations than to know that in the wild it is always absent from available high elevation sites? From a statistical perspective, even if a precise causal understanding of nature is ideal, such studies are not guaranteed to provide the most reliable inferences. Modern statisticians argue that all models are wrong, and so endeavoring to develop a precise mechanistic understanding of a particular problem may produce less reliable inferences than an approximate understanding that is well supported by the data (Peters 1991, Burnham and Anderson 2002). Perhaps the deepest problem with the claim that distribution

based niche models are suspect is that in many cases they represent the only data available.

To overcome this problem, we need a formal criterion to measure the utility of different approaches to investigate a species' niche. Here, I develop a criterion with a simple biological and statistical interpretation, by noting that either correlative or experimental studies ought to tell us if an environment is suitable to our study organism. This criterion can be mathematically formalized by stating that a good method should provide an accurate estimate of whether an environment is suitable to a species, in other words, whether the environment is a part of the species' Hutchinsonian niche (Hutchinson 1957). I then apply this criterion to statistical investigations of species' niche and determine when we should expect correlative and experimental studies to provide an accurate estimate of this probability.

Model

I develop a model of the niche primarily based on Hutchinson's (1957) ideas, with minor changes to link his set theory with statistical terminology. I start from this definition because Hutchinson's work remains one of the clearest articulations of the term "niche" (Hurlbert 1984). As we shall see, this definition also has a natural connection to probability theory and statistics. Moreover, this work has been important for developing a theoretical understanding of the relationship between the distribution of a species and its environmental requirements (Soberon and Peterson 2005, Kearney 2006, Soberon 2007). Hutchinson (1957) described the niche as "a set of points in an abstract n -dimensional N space", where each dimension represents an axis of environmental variation. Statistical theory frequently eschews the term 'point', and so I will describe each environment \underline{X} as a vector of n elements, each one of which corresponds to one environmental variable (definitions in Table 1). With this terminology we can consider any environment as a part of an n -dimensional space, i.e. \mathfrak{R}^n . Since we now know that dispersal from source populations can ensure persistence in unsuitable habitats I will follow a slightly modified version of this definition: the set of environments where population growth rate is positive, in the absence of immigration (Pulliam 1988, Holt 1996).

Subsequent studies have distinguished between the abiotic aspects of the environment and interactions between

species (Hutchinson 1978, Leibold 1995) and between unchanging aspects of the environment and resources that may be depleted (Leibold 1995, Soberon 2007). For the sake of generality, I will simply treat each interaction as a separate axis of environmental variation (Araujo and Guisan 2006). For example, under some models, competition between our focal species and another species for two resources could be defined by a small number of environmental variables, including the supply point of each resource and the abundance of the competitor (Chase and Leibold 2003).

To model the relationship between a species' niche and its distribution, we must also consider the environments to which a species can disperse. Recent treatments of the niche (Soberon 2007) have defined the set of geographic grid cells that are accessible to the dispersal capacities of an organism as M . The purpose of this paper is to create models for determining the suitability of environments, not particular locations. For this reason, I slightly redefine M as the set of environments associated with the locations to which a species can disperse (hereafter 'accessible' locations). To connect Hutchinson's ecological niche to the actual distribution of an organism, we need a model of how processes such as environmental requirements, local extinctions and dispersal limitation interact to shape the distribution of a species. Several authors have proposed special-case models that incorporate these processes (Pulliam 2000, Tilman 2004, Bahn et al. 2008), but there is no general model of how they interact. Given this, I will add the simplifying assumption that the organism is present at each suitable location to which it can disperse (Guisan and Theurillat 2000, Pearson and Dawson 2003). Biologically, this assumption has at least two interpretations; the simplest is that when an organism can reach a location, its ability to persist is very strongly determined by whether the environment is suitable. Alternatively, a biologist may be able to use additional sources of data to distinguish sources from sinks and so only count source populations as presences. I will further define R as the set of environments associated with the cells in the geographic grid of the region that we wish to investigate.

I will assume that any study of the ecological niche of an organism will be carefully designed and executed (Guisan and Thuiller 2005). At a minimum, such a study will determine if the organism is present across a representative sample of locations in the study region R . Many current studies are based on museum specimens or other on haphazardly collected presence records, so

Table 1. List of variables used in the text.

Variable	Definition
\underline{X}	The vector of all n environmental variables that describe an individual environment
\mathfrak{R}^n	The hyper space defined by all possible aspects of the environment
N	The niche of an organism, i.e. the set of environments where the population growth rate of an organism is positive in the absence of emigration
M	the set of environments associated with the locations to which a species can disperse
R	the set of environments associated with the cells in the geographic grid of the region that we wish to investigate
L	The set of environments used in a replicated experiment. This may include field conditions, laboratory conditions, or some hybrid of the two
$\underline{X}_{\text{observed}}$	The sub-set of conditions found at environment \underline{X} that a scientist can measure and/or manipulate
$P()$	The probability of an outcome

this assumption may represent an optimistic approximation. Empirical studies can only measure a sub-set of environmental variables, defined here as $\underline{X}_{\text{observed}}$. This set consists of a number of individual environmental variables denoted $(X_1 \dots X_i)$.

I will next consider experimental studies that estimate population growth rate. Statistical analysis of such data is usually an estimate of the probability that population growth rate takes on a particular value for a particular set of conditions. (Sibly and Hone 2002, Hooper et al. 2008). I explicitly define the set of environments used for experimental manipulations as L . This set might represent a sub-set of R if the experiment is conducted in a field site in the study region. It may also represent a field site outside of R , or the set of environments available in a laboratory.

Probability that an environment is a part of the niche

To understand why the niche is so difficult to measure, it is important to consider how we sample environments. In order to achieve generality, Hutchinson defined the niche as a subset of all possible environments (the set \mathfrak{R}^n), but this would seem to pose an insurmountable problem for empirical investigations. In principle, we can ask what is the probability that an environment is a part of the niche, given all possible environments, as $P(\underline{X} \in N | \mathfrak{R}^n)$. To estimate this quantity statistically, we need to know which environments are parts of the niche in a representative sample of \mathfrak{R}^n . Since it is impossible to obtain a sample of all environments that could exist, there is no general way to estimate the probability that an environment is a part of the niche.

It is important to note that this problem is no different from that facing many other scientific concepts. For example, consider a concept as simple as the average value of a parameter. There is generally no way to measure every individual, and so there is typically no way to measure a true parameter value (Fisher 1955). Instead of endeavoring to develop an absolute estimate of the average of a variable, most biologists are content with estimating the range of possibilities that are reasonable, given the data at hand and the population of interest. This same procedure can be applied to the concept of the niche. For the idea to be useful we do not need a precise description of the entire range of the environmental tolerances of a species. Instead, we need a way to estimate when a species will meet its environmental requirements, given the set of environments that are available and the variables we can measure.

What niche models measure

It is relatively easy to integrate Hutchinson's definition of the niche into a statistical framework. When an organism occurs only in environments that are suitable, and accessible, it will be present in the set of environments given by $N \cap M$. We can then define the probability that a species is present in a particular environment found in the region R . This is the probability that an organism has access to a particular environment, and that the environment is a part of its niche, given that the environment is in the set of environments present in R (Fig. 1a), symbolically:

$$P(\underline{X} \in N \cap M | R) = \frac{P(N \cap M \cap R)}{P(R)} \quad (1)$$

Furthermore, we only estimate the effects of a subset of the n environmental variables, and so it is only possible to estimate the probability that a species is present conditioned on the set of environmental variables examined, $\underline{X}_{\text{observed}}$

$$P(\underline{X} \in N \cap M | R, \underline{X}_{\text{observed}}) = \frac{P(N \cap M \cap R | \underline{X}_{\text{observed}})}{P(R | \underline{X}_{\text{observed}})} \quad (2)$$

To link the set theoretic notation above with statistical terminology, we may replace the term $N \cap M$ with an equivalent indicator function Y_k , which has a value of 1 when the species is present in the environment k , and zero when this species is absent.

$$P(Y_k = 1 | R, \underline{X}_{\text{observed}}) = \frac{P(N \cap M \cap R | \underline{X}_{\text{observed}})}{P(R | \underline{X}_{\text{observed}})} \quad (3)$$

Equation 3 has a simple statistical interpretation: the probability that a species is present, conditioned on the environments that can be sampled and the variables measured. This can also be expressed as the probability that a species is present, conditioned on the data. Analyses of a random sample of distribution records from R using parametric and semi-parametric algorithms can estimate this probability, though they make different modeling assumptions. For example, methods such as generalized linear models, generalized additive models, naive Bayesian classification and boosted regression trees model estimate $P(Y_k = 1 | R, \underline{X}_{\text{observed}})$ directly with presence/absence records (Hastie and Tibshirani 1990, Friedman et al. 2000, Hand and Yu 2001). Methods using presence data and background points (pseudoabsences) including maxent, must be interpreted with more caution. These models do not estimate the probability that a species is present, but many such algorithms do estimate a function that increases monotonically with this probability (Phillips et al. 2009).

In general, a model of presence data may be strongly influenced by dispersal limitation, the niche, or both, depending on the environments available and the capacity of the organism to disperse within the region studied (Fig. 1a). In some cases, though, a niche model can be even more informative. When an organism can disperse to all of the environments within R (i.e. $R \subseteq M$), then the model provides an estimate for:

$$P(\underline{X} \in N | L, \underline{X}_{\text{observed}}) = \frac{P(N \cap M \cap L | \underline{X}_{\text{observed}})}{P(L | \underline{X}_{\text{observed}})} \quad (4)$$

(Appendix 1)

Equation 4 indicates that when a species is at pseudoequilibrium and we restrict our inferences to habitats to which our organism could disperse, an estimate of the probability that a species is present is also an estimate of the probability that an environment is a part of the niche (Fig. 1b).

What experimental studies measure

First, I analyze experiments used to determine if the intrinsic rate of growth of an organism was greater than

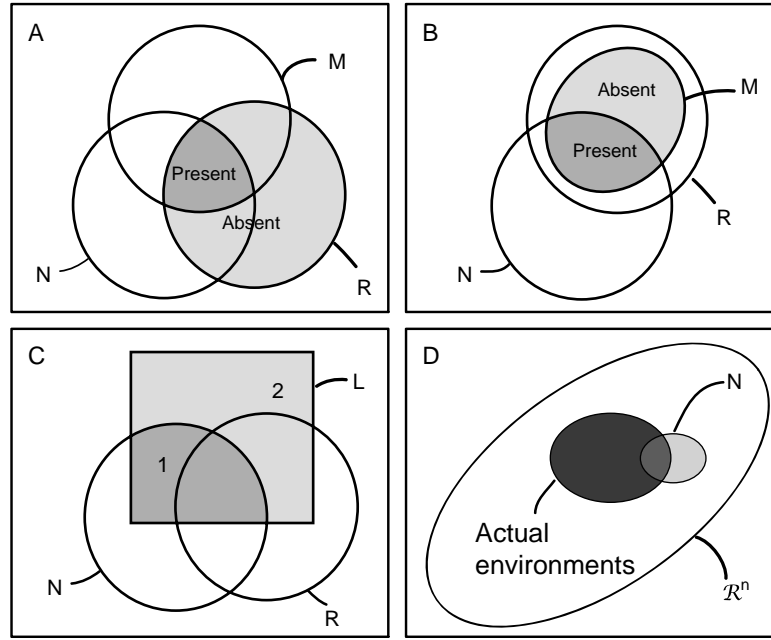


Figure 1. (A) Conceptual model of presence data. We may observe our study organism in the set of environments given by: $N \cap M \cap R$. From this we can infer that an environment occupied by the organism is a part of the niche, while a species may be absent because an environment is unsuitable, or unavailable because of dispersal limitation, or both. (B) A model of presence data when the organism can disperse to every environment studied. In this case, presences correspond to environments that are a part of the niche of the organism, and absences correspond to environments that are not a part of the niche of the organism. (C) A conceptual model of experimental investigations of the niche. A scientist investigates the set of manipulated conditions, L . With such a design, it is possible to determine which environments are a part of the niche (region 1), and which are not (region 2). Most commonly, L is not a representative sample of environments in R , and instead represents some amalgam of natural and artificial environments. (D) An illustration of why the niche sensu Hutchinson is difficult to study empirically. Hutchinson defined the niche (N) in terms of the set of all possible environments \mathfrak{R}^n , however the set of environments that actually exist are only a tiny fraction of \mathfrak{R}^n . Thus, to actually measure a Hutchinsonian niche we would need knowledge of many environments that are a part of N , but have never existed.

zero (Hooper et al. 2008). I then use probability theory to analyze the utility of causal inferences in the face of incomplete data.

A factorial experiment may be thought of as a probabilistic estimate, much like a niche model. An experiment samples from the set of environments, L . In this set, some environments are a part of the niche ($N \cap L$). The job of an investigator is then to determine what variables cause an environment to be a part of the niche. This is done by manipulating a subset of variables $\underline{X}_{\text{observed}}$, while randomly assigning treatments across the environments in L . As with a niche model, we can derive the probability that an environment is a part of the niche, given the environments we have sampled.

$$P(\underline{X} \in N | L, \underline{X}_{\text{observed}}) = \frac{P(N \cap L | \underline{X}_{\text{observed}})}{P(L | \underline{X}_{\text{observed}})} \quad (5)$$

Equation 5 represents the probability that an environment is a part of the niche conditioned on the environmental variables manipulated, and the set of environments examined (Fig. 1c). A standard statistical analysis, such as a generalized linear model, will estimate this probability. Notably, such a study estimates the probability that an environment in L is a part of the niche, and we must make further assumptions to apply such knowledge directly to the set of environments present in the entire study region (R).

If there are only a small number of resources it is relatively easy to use probability theory to determine how unmeasured variables can affect the interpretation of our models. In Supplementary material Appendix 2, I investigate our ability to extrapolate from incomplete local measurements of different resource types when we are just as likely to encounter any combination of the two resources that define a species' niche.

Hutchinson described a relatively simple model of environmental requirements, where an organism will survive when the amount of each resource is higher than a minimum threshold and lower than a maximum threshold (Fig. 2A). If we only determine the species' requirements for one resource, in one environment, then we can correctly identify many unsuitable environments. We may however identify as suitable some environments that are unsuitable (Fig. 2B–C).

Another popular model involves two resources, one of which can be substituted for the other (Tilman 1980) (Fig. 2D). If we determine species' requirements for one such resource in a single environment, it appears that concentrations of this resource above a threshold result in an environment being suitable (Fig. 2E). However, across the entire range of available environments, the amount of the resource necessary can vary a great deal (Fig. 2, Supplementary material Appendix 2). Moreover, a model based on

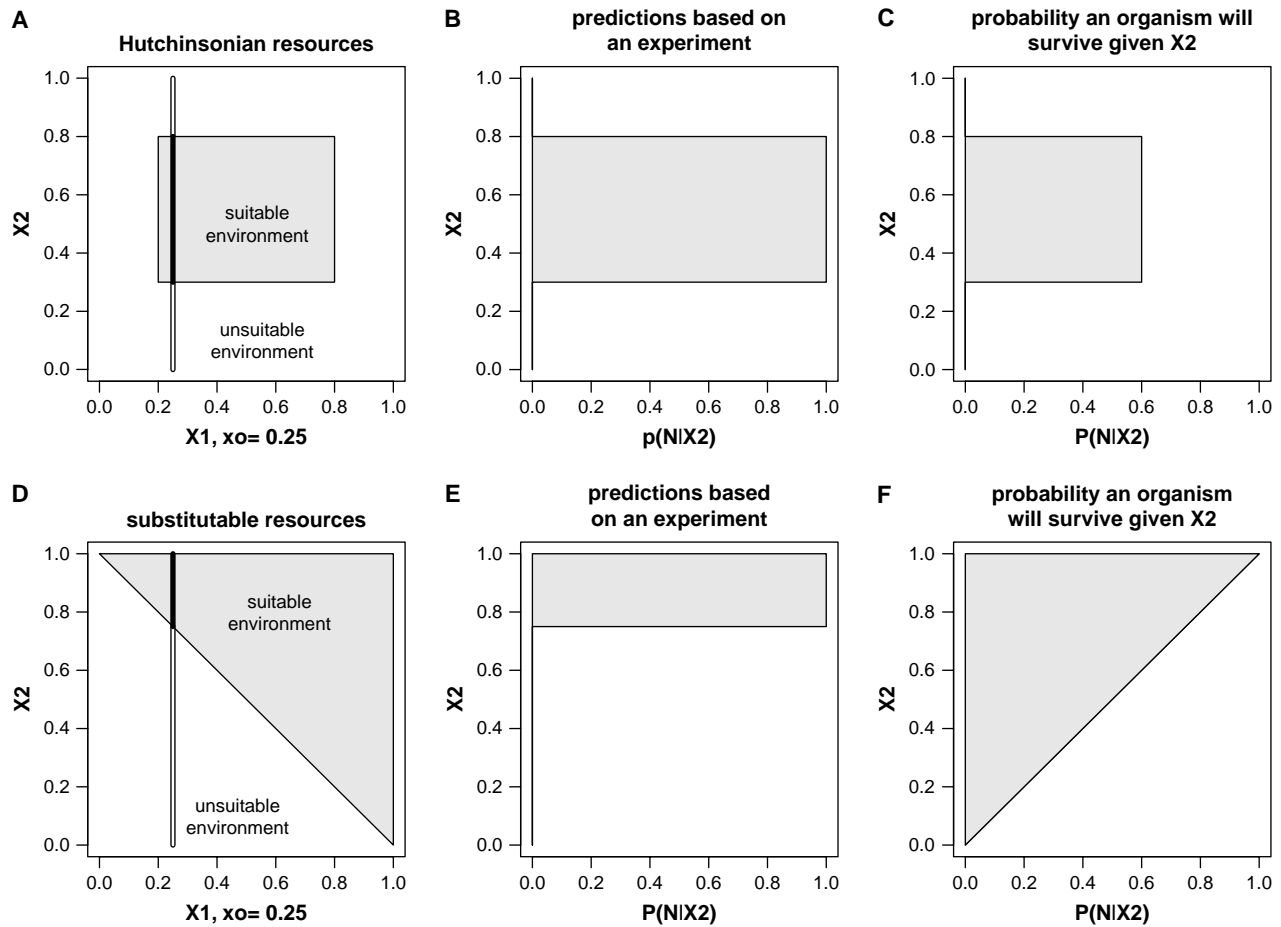


Figure 2. A comparison of local and regional descriptions of a species niche, when scientists can only measure a subset of relevant variables. (A) A plot of the effects of two resources as described by Hutchinson. Organisms will survive in environments with intermediate values of variables x_1 and x_2 but die otherwise. The vertical line through the plot represents the effect of x_2 in a particular environment (i.e. where x_1 has a single value, 0.25 in this case). (B) The probability that an environment is suitable, given x_2 when variable x_1 is equal to 0.25. An experimental study conducted in a local environment with a background value of x_1 would infer this relationship between the value of x_2 and environmental suitability. (C) The regional probability that an environment is suitable given x_2 . This probability corresponds to the proportion of environments that are suitable at a given value of x_2 . Note that a local study provides an estimate of the probability that a species is present that is similar to the regional probability that a species is present. (D) The effects of two substitutable resources, with a vertical line to represent the effects of x_2 in a particular environment. (E) The probability that an environment is suitable, given x_2 when variable x_1 is equal to 0.25. (F) The regional probability that an environment is suitable given x_2 . Note that a local estimate of the probability that a species is present is quite different from the regional probability that a species is present across the entire study region. See Supplementary material Appendix 2 for a formal derivation of these results.

experimental data may falsely infer that environments are either suitable or unsuitable (Fig. 2F).

Discussion

The results of this paper indicate that many commonly employed statistical methods provide a great deal of information about the niche of an organism. Nevertheless, we cannot interpret the results of such analyses without considering the environments that are of interest and the variables that we are unable to measure.

While the niche has been defined in terms of all possible environments, we know that planet Earth only contains some fraction of these environments (Fig. 1d). Previous authors have described this problem and proposed that changes in the availability of habitat within the niches of

individual species may have important consequences for the long-term stability of ecological communities (Jackson and Overpeck 2000). Nevertheless, it has remained difficult to connect an expansive definition of the niche with inferences based on the finite set of observations available. We can phrase this problem statistically by noting that we cannot define the niche of an organism unless we are able to make inferences from environments that have never existed. We can, however, be reasonably confident that a particular environment is a part of the niche or not – that is, we will know the niche when we see it. By developing such a conditional interpretation of the niche we can directly connect the concept of the niche to statistical investigations, while clarifying the best way to use these methods.

When investigating presence data, it is crucial to know which environments are accessible to a species (Guisan and Thuiller 2005). If a species can disperse to all background

environments considered, a model would estimate the probability that an environment is a part of the niche. Otherwise, models will only estimate the probability that a species is present (Fig. 1a). The clearest examples of this problem involve introduced species, such as *Lythrum salicaria* (Lythraceae). Two hundred years ago it would have been conceivable to create niche models for *L. salicaria* that classified its European range as presences, and North America as absences, but we now know that such an exercise would have been meaningless. Subsequent observations have demonstrated that North America may have been inaccessible to *L. salicaria*, but that environments in a great deal of that continent are a part of its niche (Blossey et al. 2001). It has already been noted that the selection of background environment can dramatically change the resulting niche model. Several studies have considered the statistical basis for picking background points (Thuiller et al. 2004, Chefaoui and Lobo 2008) and it has previously been suggested that background points should be selected from locations to which a species can disperse (Soberon 2007). This study is the first to rigorously derive this result.

Niche models will also give misleading answers if the assumption of pseudo-equilibrium is violated by frequent local extinctions. Indeed, the interplay of extinction and dispersal will clearly be a problem for many modeling approaches (Kearney and Porter 2009). Regardless of its role, we must still interpret our models in the face of incomplete measurements of highly dimensional data. I have endeavored here to produce an analytic solution to this one problem that can be generalized across different models of how environmental variables shape a species' requirements, and across different modeling methods, but I have done so at the expense of developing an explicit model of stochasticity. It is beyond the scope of this paper to model this process explicitly, but it is possible to make a few general observations on this problem. Some environments are more prone to extinction than others. In particular, small patches of marginal habitat tend to be more prone to extinction than large patches of suitable habitat. Similarly, some patches of habitat are less likely to be repopulated after an extinction event (Moilanen and Hanski 1998). A correlative model will underestimate the probability that these environments are a part of the niche because the organism may be absent even if the environment is suitable. Nonetheless, the results presented here demonstrate that we can develop valid inferences for any set of environments R , when the assumption of pseudo-equilibrium holds and the organism can disperse to all of R . Even if local extinction is common in one region, or at one scale, it may still be possible to develop niche models for the same species at a different scale, or in regions where local extinctions are rare.

A related limitation of the model is that it estimates only the probability that population growth rate is positive. Previous authors have modeled the probability that a population will grow at a specified rate given the available environmental conditions (Sibly and Hone 2002, Hooper et al. 2008). It is my hope that the probabilistic approach I have outlined here can be extended to consider both extinction and variation in population growth rate. However, it may be difficult to derive analytic results for existing models (Pulliam 2000, Tilman 2004).

It is difficult to understand the merits of experimental approaches without considering our limited ability to conduct physiological studies in different background environments. Small-scale experiments, may provide strong estimates in one context i.e. $P(\underline{X} \in N | L, \underline{X}_{\text{observed}})$, but do not necessarily provide estimates from a representative sample of naturally occurring environmental conditions. If these environments do not constitute a representative sample, experiments may give a misleading picture of whether an environment is a part of the niche, or an optimistic assessment of the errors associated with our inferences. One well-known example of this problem involves the role of carbon and phosphorus in promoting the growth of algal blooms. Experiments in the 1960s concluded that carbon was a limiting nutrient for the phytoplankton that cause blooms of blue green algae (Lange 1967). However, correlative data and subsequent large-scale experiments concluded that carbon was more or less irrelevant in real environments, and that previous assertions to the contrary were artifacts of the laboratory setting in which the experiments were conducted (Schindler 1971). Correlative studies have an advantage when it is easier to sample the set of naturally occurring environments than it is to model every aspect of environmental variation. However, any statistical inference is only strictly valid for the population sampled. Thus, the results presented here do not directly bear on the ability of distribution models or experimental studies to predict species distributions in novel environments (environments we were not able to sample).

There are some similarities between the results presented here and current interpretations of mechanistic investigations. The claim that it is easier to understand uncertainty in correlative models because we may sample environments at random is perhaps a formalization of existing ideas of the relative merits of mechanistic and correlative models. Kearney and Porter (2009) for example claim that it is "easier to incorporate geographical variation [in distribution models] because it is indirectly represented in the occurrence data". Under Hutchinson's model of environmental requirements a causal inference can correctly identify many unsuitable areas but will not necessarily identify all suitable regions. This result is reminiscent of the way unmeasured variables are interpreted in biophysical niche models. For example, (Kearney et al. 2008) state, "because we can never capture all factors constraining the fundamental niche, our strongest inference is on the identification of areas outside the fundamental niche". Interestingly, the results presented here indicate that causal inferences can erroneously infer that areas are unsuitable when resources are substitutable. This said, mechanistic models are typically not conceptualized as statistical inferences (Kearney and Porter 2009) and so it is not clear that such models can be considered an estimate of the probability that an environment is a part of a species' niche.

Previous authors have tackled our limited ability to make inferences about a species' niche by attempting to define different aspects of the niche, such as defining the niche in terms of our knowledge of mechanisms (Kearney 2006), biotic and abiotic variables (Kearney and Porter 2004, Soberon and Peterson 2005) or depletable versus non-depletable resources (Soberon 2007). Doubtless it is important to recognize that different environmental variables

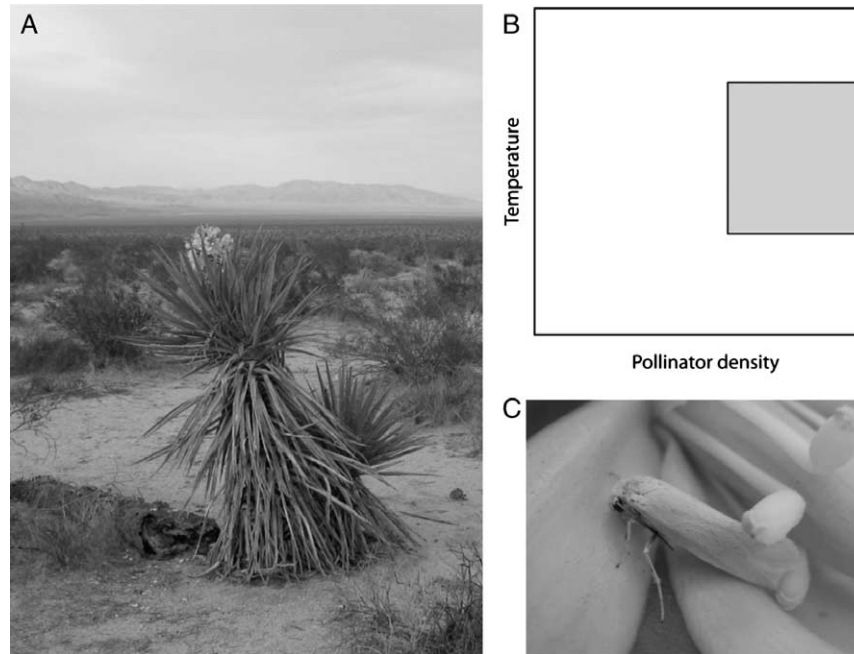


Figure 3. Environmental requirements in the face of strong biotic interactions. (A) *Yucca schidigera* a perennial plant from the American southwest can only reproduce with the assistance of its pollinator *Tegeticula* spp. (B, C). It is possible to determine this species' physiological requirements such as the temperatures in which it will persist. Note that below some pollinator density the *Y. schidigera* cannot persist (B) and so this species does not have a fundamental niche.

act in different ways, but it is frequently impossible to statistically disentangle the effects of different variables (Myers 1990, Ross 1997). We should thus be suspicious of claims that any method can distinguish one component of a species' niche from another.

Obligate mutualisms provide some of the clearest examples of how endeavors to separate different components of species niches can obscure the interpretation of empirical observations. Plants such as the succulent *Yucca schidigera* can only reproduce with the assistance of obligate pollinators (Pellmyr 2003). Physiologists have measured the climatic tolerances of this species and used this information to make inferences about which abiotic environments are suitable (Loik et al. 2000). Paradoxically, such a description is more useful than an estimate of *Y. schidigera*'s fundamental niche (Fig. 3). Following Kearney and Porters' (2004) definition of the fundamental niche as the "set of conditions and resources that allow a given organism to survive and reproduce in the absence of biotic interactions," *Y. schidigera* has no fundamental niche since it cannot reproduce in the absence of its pollinators. In this case, and estimate of the probability that an environment is suitable, given the species' abiotic requirements and the fact that pollinators are present could be much more useful than an estimate of the species' fundamental niche.

A probabilistic interpretation of the niche provides a way to conceptualize our models when we are unable or unwilling to identify the separate effects of every possible variable. We can recognize, for example, that a mechanistic model provides insights on *Y. schidigera*'s abiotic requirements, conditioned on the presence of favorable biotic interactions. Likewise, distribution models tell us about whether an environment will be suitable conditioned on the values of some measurements of the abiotic environment,

even if we cannot determine if the organisms require these abiotic environments or have a distribution that results from associated interactions like competition and facilitation. If we are confident that we have correctly measured the effects of every relevant aspect of the environment this perspective is of limited utility. When we cannot be certain that unmeasured aspects of the environment are irrelevant or that we have perfectly partitioned the effects of different variables then probability theory provides a way to interpret our models and to identify the relative strengths of different methods.

Acknowledgements – Luke Harmon, Jeffrey Evans, Melanie Murphy, Ben Ridenhour, Olle Pellmyr, Jeremy B. Yoder, Paul Joyce, Christopher I. Smith, Stacey Dunn, Lydia Gentry and Zaid Abdo provided helpful comments and suggestions. Financial support was provided by National Science Foundation grant DEB-0516841, a pre-doctoral fellowship from the Canadian National Science and Engineering Research Council and the Foundation for Orchid Research and Conservation.

References

- Araujo, M. B. and Guisan, A. 2006. Five (or so) challenges for species distribution modeling. – *J. Biogeogr.* 33: 1677–1688.
- Bahn, V. et al. 2008. Dispersal leads to spatial autocorrelation in species distributions: a simulation model. – *Ecol. Modell.* 213: 285–292.
- Blossey, B. et al. 2001. Impact and management of purple loosestrife (*Lythrum salicaria*) in North America. – *Biodiv. Conserv.* 10: 1787–1807.
- Burnham, K. P. and Anderson, R. D. 2002. Model selection and multimodel inference: a practical information-theoretic approach. – Springer.

- Chase, J. M. and Leibold, M. A. 2003. Ecological niches – linking classical and contemporary approaches. – Univ. of Chicago Press.
- Chefaoui, R. M. and Lobo, J. M. 2008. Assessing the effects of pseudo-absences on predictive distribution model performance. – *Ecol. Modell.* 210: 478–486.
- Colwell, R. K. and Futuyma, D. J. 1971. On the measurement of niche breadth and overlap. – *Ecology* 52: 567–576.
- Fisher, R. 1955. Statistical methods and scientific induction. – *J. R. Stat. Soc. B* 17: 69–78.
- Friedman, J. et al. 2000. Additive logistic regression: a statistical view of boosting. – *Ann. Stat.* 28: 337–407.
- Guisan, A. and Theurillat, J.-P. 2000. Equilibrium modeling of alpine plant distribution: how far can we go? – *Phytocoenologia* 30: 353–384.
- Guisan, A. and Thuiller, W. 2005. Predicting species distribution: offering more than simple habitat models. – *Ecol. Lett.* 8: 993–1009.
- Hand, D. J. and Yu, K. 2001. Idiot's Bayes: not so stupid after all. – *Int. Stat. Rev.* 69: 385–398.
- Hastie, T. and Tibshirani, R. 1990. Generalized additive models. – Chapman and Hall.
- Holt, R. D. 1996. Adaptive evolution in source-sink environments: direct and indirect effects of density-dependence on niche evolution. – *Oikos* 75: 182–192.
- Hooper, H. L. et al. 2008. The ecological niche of *Daphnia magna* characterized using population growth rate. – *Ecology* 89: 1015–1022.
- Hurlbert, S. H. 1984. A gentle depilation of the niche: dicean resource sets in resource hyperspace. – *Evol. Theory* 5: 177–184.
- Hutchinson, G. E. 1957. Concluding remarks. – In: Cold Spring Harbor Symp. on Quantitative Biology, pp. 415–427.
- Hutchinson, G. E. 1978. An introduction to population ecology. – Yale Univ. Press.
- Jackson, S. T. and Overpeck, J. T. 2000. Responses of plant populations and communities to environmental changes of the late Quaternary. – *Paleobiology* 26: 194–220.
- Kearney, M. 2006. Habitat, environment and niche: what are we modeling. – *Oikos* 115: 186–191.
- Kearney, M. et al. 2008. Modelling species distributions without using species distributions: the cane toad in Australia under current and future climates. – *Ecography* 31: 423–434.
- Kearney, M. and Porter, W. P. 2004. Mapping the fundamental niche: physiology, climate and the distribution of a nocturnal lizard. – *Ecology* 85: 3119–3131.
- Kearney, M. and Porter, W. P. 2009. Mechanistic niche modelling: combining physiological and spatial data to predict species ranges. – *Ecol. Lett.* 12: 334–350.
- Lange, W. 1967. Effect of carbohydrates on the symbiotic growth of planktonic blue-green algae with bacteria. – *Nature* 215: 1277–1278.
- Leibold, M. A. 1995. The niche concept revisited: mechanistic models and community context. – *Ecology* 76: 1371–1382.
- Loik, M. E. et al. 2000. Low temperature tolerance and cold acclimation for seedlings of three Mojave Desert *Yucca* species exposed to elevated CO₂. – *J. Arid Environ.* 46: 43–56.
- May, R. M. and MacArthur, R. H. 1972. Niche overlap as a function of environmental variability. – *Proc. Natl Acad. Sci.* 69: 1109–1113.
- Molander, A. and Hanski, I. 1998. Metapopulation dynamics effects of habitat quality and landscape structure. – *Ecology* 79: 2503–2515.
- Myers, R. H. 1990. Classical and modern regression with applications. – Brooks/Cole.
- Pearson, R. G. and Dawson, T. P. 2003. Predicting the impacts of climate change on the distribution of species: are bioclimate envelope models useful. – *Global Ecol. Biogeogr.* 12: 361–371.
- Pellmyr, O. 2003. Yuccas, yucca moths, and coevolution: a review. – *Ann. Miss. Bot. Gard.* 90: 35–55.
- Peters, R. H. 1991. A critique for ecology. – Cambridge Univ. Press.
- Phillips, S. J. et al. 2009. Sample selection bias and presence-only distribution models: implications for background and pseudo-absence data. – *Ecol. Appl.* 19: 181–197.
- Pulliam, R. 1988. Sources, sinks and population regulation. – *Am. Nat.* 132: 652–661.
- Pulliam, R. 2000. On the relationship between niche and distribution. – *Ecol. Lett.* 3: 349–361.
- Ross, S. 1997. A first course in probability. – Prentice Hall.
- Schindler, D. W. 1971. Carbon, nitrogen and phosphorus and the eutrophication of freshwater lakes. – *J. Phycol.* 7: 321–329.
- Sibly, R. and Hone, J. 2002. Population growth rate and its determinants: an overview. – *Philos. Trans. R. Soc. Lond. B* 357: 1153–1170.
- Soberon, J. 2007. Grinnellian and Eltonian niches and geographic distributions of species. – *Ecol. Lett.* 10: 1115–1123.
- Soberon, J. and Peterson, A. T. 2005. Interpretation of models of fundamental ecological niches and species' distributional areas. – *Biodiv. Inf.* 2: 1–10.
- Thuiller, W. et al. 2004. Effects of restricting environmental range of data to project current and future species distributions. – *Ecography* 27: 165–172.
- Tilman, D. 1980. A graphical-mechanistic approach to competition and predation. – *Am. Nat.* 116: 362–393.
- Tilman, D. 2004. Niche tradeoffs, neutrality, and community structure: a stochastic theory of resource competition, invasion, and community assembly. – *Proc. Natl Acad. Sci.* 101: 10854–10861.
- Whittaker, R. H. et al. 1973. Niche, habitat and ecotope. – *Am. Nat.* 107: 321–338.

Supplementary material (available online as Appendix O17630 at <www.oikos.ekol.lu.se/appendix>). Appendix 2

Appendix 1. Proof that when an organism can disperse to all of R, a niche model estimates the probability that an environment is a part of the niche.

if $R \subseteq M$ then $M \cap R = R$

then

$$\begin{aligned} P(\underline{X} \in N \cap M | R, \underline{X}_{\text{observed}}) &= \frac{P(N \cap M \cap R | \underline{X}_{\text{observed}})}{P(R | \underline{X}_{\text{observed}})} \\ &= \frac{P(N \cap R | \underline{X}_{\text{observed}})}{P(R | \underline{X}_{\text{observed}})} \\ &= P(\underline{X} \in N | R, \underline{X}_{\text{observed}}) \end{aligned}$$