Research

Ecography 33: 159–167, 2010
doi: 10.1111/j.1600-0587.2009.06074.x

Subject Editors: Núria Roura-Pascual and Miguel Araújo. Accepted 18 December 2009

Niche and area of distribution modeling: a population ecology perspective

Jorge M. Soberón

J. M. Soberón (jsoberon@ku.edu), Museum of Natural History and Dept of Ecology and Evolutionary Biology, Univ. of Kansas, Dyche Hall 1345 Jayhawk Blvd., Lawrence, KS 66045, USA.

Statistical modeling of areas of distribution of species by correlative analysis of the environmental features of known presences has become widespread. However, to a large degree, the logic and the functioning of many of these applications remain obscure, not only due to the fact that some of the modeling methods are intrinsically complex (neural networks, genetic algorithms, generalized additive models, for example), but mainly because the role of other ecological processes affecting the species distributions sometimes is not explicitly stated. Resorting to fundamental principles of population ecology, a scheme of analysis based on separation of three factors affecting species distributions (environment, biotic interactions and movements) is used to clarify some results of niche modeling exercises. The area of distribution of a virtual species which was generated by both environmental and biotic factors serves to illustrate the possibility that, at coarse resolutions, the distribution can be approximately recovered using only information about the environmental factors and ignoring the biotic interactions. Finally, information on the distribution of a butterfly species, *Baronia brevicornis*, is used to illustrate the importance of interpreting the results of niche models by including hypothesis about one class of movements. The results clarify the roles of the three factors in interpreting the results of using correlative approaches to modeling species distributions or their niches.

The sets of techniques collectively called Species Distribution Modeling (SDM) (Guisan and Zimmermann 2000, Guisan and Thuiller 2005), and the related Ecological Niche Modeling (ENM) (Peterson 2006) have experienced a tremendous growth in the last 15 yr (Thuiller et al. 2009). This growth is partly due to data availability: presence–only data are now available on-line in very large quantities, for example through the Global Biodiversity Information Facility (Edwards 2004), which currently provides access to more than 150 million records. The other class of data required to perform SDM or ENM are electronic "layers" of a specific class of environmental variables, which also have exploded in terms of availability, resolution and ease of use (Soberón and Peterson 2004).

Another reason for the acceleration in the usage of SDM and ENM techniques is simply that these techniques have proved their predictive capacity in a number of situations (Sanchez-Cordero and Martínez-Meyer 2000, Feria and Peterson 2002, Raxworthy et al. 2003). Finally, there is a widespread perception that the success of these techniques is due to the existence of a deep relation between large-scale features of the distributions of species and their environmental requirements (Pulliam 2000, Peterson 2003, Soberón and Peterson 2005, Araújo and Guisan 2006, Soberón 2007, Hirzel and Le Lay 2008). However, there are also dissenting views (Bahn and McGill 2007, Beale et al. 2008) and the fields of SDM and ENM still require conceptual clarification (Pearson and Dawson 2003, Araújo and Guisan 2006, Kearney 2006, Jiménez-Valverde et al. 2008, Elith and Graham 2009, Lobo et al. 2010). Better conceptual clarity is achieved by consideration of the causal factors that determine geographic distribution. It is possible to express in a single mathematical framework the interaction of three such major factors: environmental conditions, regulatory factors and movements. These factors underlie the presence of individuals of a species in space and time, thus providing a theoretical framework for the relation between types of species distributions and different kinds of niche concepts. I will apply this theoretical framework to the clarification of the role of two major complicating factors in ENM, namely biotic interactions and movements (Araújo and Guisan 2006, Hirzel and Le Lay 2008).

Spatially-explicit population dynamics

Concepts of niche are always defined in terms of the combinations of requirements that allow a population to survive and grow in a given place, and on the impacts that a species has on the ecological community where it lives (Chase and Leibold 2003). Since in every case there are large numbers of combinations of conditions and resources that permit a population to survive and reproduce, a multidimensional view of niche (Hutchinson 1957) becomes necessary. Areas of distribution, on the other hand, are defined by the possibility of detecting individuals of a species in a given locality (Brown et al. 1996), and this in turn depends on the demographic conduct of the species in a geographically explicit context, and most important, on the resolution of the observations. A model of population dynamics that is spatially-explicit contains the parameters that can be used to define niches. It can also be used to define areas of distribution, since the solutions of the model reveal what areas will have individuals present at different densities, or what regions of the space contain environmental parameters capable of producing positive growth rates.

Following ideas first proposed by Vandermeer (1972) and Pulliam (2000), I consider a set of *1,2,...N* species interacting in a world defined by a discrete grid of $j = 1,2,\ldots H$ cells. The cells should be small enough in relation to the biology of the species as to allow meaningful definition of their environmental parameters and movements. We denote the entire set of grid cells by **G**. The per capita growth rate of a population of species i in cell j is:

$$\frac{1}{x_{i,j}}\frac{dx_{i,j}}{dt} = r_{i,j}(\vec{e}_j) - \varphi_{i,j}(\vec{x}_j;\vec{R}_{i,j},\vec{e}_j) + \psi_{i,j}(\vec{x}_i;\mathbf{T}) \qquad (1)$$

where $x_{i,j}$ is the population density, $r_{i,j}(\vec{e}_j)$ is the intrinsic growth rate, defined only in terms of a vector of density-independent parameters $\vec{e}_j$ that characterize the environmental conditions (Begon et al. 2006) in the *j*-th cell. These parameters may be correlated and their interaction can define gradients (Elith and Leathwick 2009). The function $\varphi_{i,j}(\vec{x}_j;\vec{R}_{i,j},\vec{e}_j)$ is a density-dependent term that expresses the effects of interactions with other species and resource usage. It is defined in terms of a vector of interaction parameters $\vec{R}_{i,j}$, a vector of densities of every species $\vec{x}_j$, and by the density-independent parameters $\vec{e}_j$. Finally there is a term $\psi_{i,j}(\vec{x}_i;\mathbf{T})$ expressing movements of individuals of species *i* from other cells into *j* according to a transition matrix **T** containing the probabilities of each possible movement.

The spatiotemporal solutions of the above equations can have extremely complicated behavior, compatible with simple attractors, limit cycles and chaotic behavior, as well as the generation of spatial waves and patterns (Solé and Bascompte 2006). However for the purpose of this discussion one does not need to parameterize and solve the general equations. It is enough to notice that the most important factors that determine an area of distribution are present in eq. (1). As has been discussed by many authors (Hutchinson 1957, MacArthur 1972, Pulliam 2000, Pearson and Dawson 2003, Guisan and Thuiller 2005, Araújo and Guisan 2006), the interactions between the three sets of factors (density independent growth rate, biological interactions and metapopulation structure) can be used to define areas of distribution of various types, by determining the regions where a species may be actually or potentially present.

The equations also suggest a simplifying assumption (Soberón 2007), namely, splitting the niche definition into the factors related to the intrinsic growth rate (the Grinnellian niche) and those related to the interactions and resources related term (the Eltonian niche). Although this distinction is strictly speaking artificial, since in general the solutions of the equations depend on the interactions between all terms simultaneously, it is useful because by separating extreme classes of factors conceptual clarity is achieved, as we will see below. The task of documenting empirically to what extent, for what scales, or for what taxa the assumption is valid, remains largely open.

Operationally, the core of the distinction between Grinnellian and Eltonian factors is the assumption that there are two types of variables, as Hutchinson (1978) suggested in a seldom-quoted chapter. The ecological environment in the cells will be separated in two classes of variables: first, those that affect the fitness, but on which the species has no impact. Hutchinson (1978) called these variables scenopoetic, for the Greek roots of "setting the scene". These type of variables have been called "direct gradients" by Austin (1980) and "conditions" by Begon et al. (2006), and we symbolize them with $\vec{e}_j$, a vector of scenopoetic variables in cell *j.* Obvious examples of scenopoetic variables are climatic and topographic variables. As long as the impacts of a species on habitat-structure variables is limited, or slow in relation to its own population dynamics, habitat structure may also be regarded as a scenopoetic variable. For instance, forest structure may be regarded as relatively unaffected by the population dynamics of some species of animal dwellers (James et al. 1984). It may be the case that defining scenopoetic variables is meaningful mainly at large scales (coarse resolutions and large extents), but this mostly will depend on the biology of the species in question, and high resolution scenopoetic variables may also exist.

The second type of variables affects the fitness of the population, but can also be consumed or modified and therefore the population has an impact on them. More loosely, these are the "bionomic" variables (the term is somewhat misleading) of Hutchinson (1978), the "resources" of Begon et al. (2006) and the "resource gradients" of Austin (1980). In general, factors that affect fitness in a regulatory manner (Meszena et al. 2006), whether top-down or bottom-up (Hassell et al. 1998) should be regarded as bionomic variables. Examples are the presence of dynamically-coupled competitors, predators or mutualists, or resources. The key distinction is whether populations of a species actually have measurable impacts on the values of the variables affecting its niche (Leibold 1996). This may take place by consumption, population interactions, or habitat modification via "niche building" (Odling-Smee et al. 2003). When the impacts a population has on variables cannot be disregarded, a niche axis cannot be represented simply as a set of numbers, the way it is done with non-interactive scenopoetic variables. In view of these reasons, in this work niches (sets of environmental values) will be defined using only scenopoetic variables (Jackson and Overpeck 2000).

The BAM diagram

Since a full analysis of eq. (1) is seriously complicated, Soberón and Peterson (2005) and Soberón (2007) proposed a heuristic scheme, to which we refer as the BAM diagram (Fig. 1), to describe some of the results of the interacting factors determining a species distribution. Briefly, these authors use the diagram as an abstract representation of geographical space, subdivided by the terms of eq. (1). First, **A** is the region in geographic space where the intrinsic growth rate of a species would be positive, on the basis of the scenopoetic environment only. By hypothesis, the environments in **A** are contained in the fundamental Grinnellian niche of the species. There are literally Terabytes of scenopoetic data available. Some scenopoetic variables may only be indirectly related to the intrinsic growth rate, therefore the variables defining **A** should ideally be selected because their proximal effects on $r_{i,j}(\vec{e}_j)$, not simply because of data availability (Elith and Leathwick 2009).

Second, the region **B** represents those areas in **G** where the results of interactions and density-dependent resource consumption [as defined by the term $\varphi_{i,j}(\vec{x}_j; \vec{R}_{i,j}, \vec{e}_j)$] would produce positive instantaneous growth rates. In other words, **B** is the area where the biotic conditions are suitable for the species. This region is determined by parameters only in the most simplified models (see below). In general, estimation of these parameters would require having a dense set of observations over large spatial extents, since it is known that the results of interactions frequently change drastically over the geographic distribution of a species (Thompson 2005, Sagarin et al. 2006). Not surprisingly, few datasets about **B** are available for most species.

Finally, the region **M** corresponds to parts of **G** that the species has been able to reach during a relevant period of time (for example, since that glacial maximum). It may be possible to estimate **M** on the basis of biogeographical considerations, or by analysis of models of the dispersal of individuals. The intersection of **A**, **B** and **M** may be regarded as representing the current area of distribution of the species.

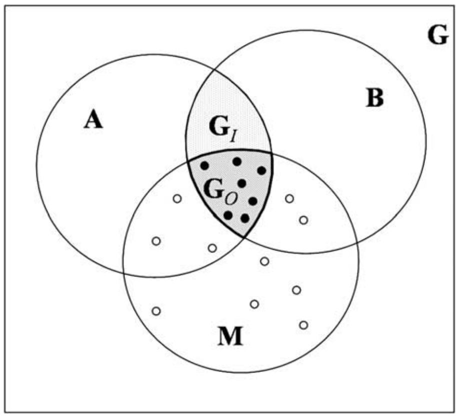

Figure 1. The BAM diagram. The region **G** represents the entire geographical region under consideration. Region **A** is the area in which scenopoetic conditions are favorable for a species. This is a potential region, probably unoccupied by the species in some sections. Region **B** is the area in which the bionomic conditions are suitable for the species. Region **M** is the area that the species has been able to reach within a given time period. The intersection $\mathbf{A}\bigcap\mathbf{B}\bigcap\mathbf{M}^C=\mathbf{G}_I$ represents a region that may potentially be invaded because both types of conditions are suitable, but that the species has not yet been able to reach. Finally $\mathbf{A}\bigcap\mathbf{B}\bigcap\mathbf{M}=\mathbf{G}_O$ represents the actual area of distribution of the species. The closed circles represent source populations. The open circles represent sink populations.

SDM and ENM are statistical procedures aiming to identify different regions in Fig. 1 (or their corresponding environmental sets of variables, or ''niches'') on the basis of observations of presences, and sometimes presences and absences. This is not a simple process, as I discuss in the following sections.

Estimation of Grinnellian niches by ecological niche modeling

A number of correlative techniques of ENM allow transferring from a set of observed presences to some undefined geographical region, and correspondingly, from the observed environmental space, to a larger one (Brotons et al. 2004, Guisan and Thuiller 2005, Soberón and Peterson 2005, Araújo and Guisan 2006, Elith et al. 2006, Peterson 2006) representing some niche. When unbiased absence data are available, correlative methods can be used to estimate the probability $P(Y=1|\vec{e}_j)$ of the species being present ($Y=1$), conditioned to environmental information $\vec{e}_j$ (Guisan and Zimmermann 2000, Keating and Cherry 2004, Phillips et al. 2009). ''Absence data'' is not a straightforward concept (Jiménez-Valverde et al. 2008, Lobo et al. 2010). Here, I use the term to mean ''contingency'' and ''environmental'' absences, sensu Lobo et al. (2010). This provides a direct and rigorous estimate of $\mathbf{G}_O$ (I suggest that these are the SDM sensu stricto). Unfortunately, unless absence data are included, correlative methods alone cannot estimate $\mathbf{G}_O$ (Ward et al. 2009). Presence–only ENM estimates sets of environmental variables that are ''similar'', in ways dependent on the algorithm used, to those in the presence localities. Therefore, without ancillary information (on effects of other species and dispersal restrictions), their geographical transferring may be somewhere between the actual area of distribution and the entire region with potentially favorable scenopoetic environments (Jiménez-Valverde et al. 2008).

There are more than 15 different correlative methods currently applied for SDM or for ENM (Kriticos and Randall 2001, Elith et al. 2006). Mechanistic methods (Sutherst and Maywald 1985, Kearney and Porter 2009), which are extremely important because they allow direct estimation of the Grinnellian fundamental niche (and thus of **A)** on the basis of physiological and biophysical first-principles, will not be discussed here.

Since the different classes of correlative methods calculate very different – albeit related – theoretical objects, and different configurations of the BAM diagram in Fig. 1 represent quite different biological scenarios (Soberón and Peterson 2005), it should be clear that ENM cannot be properly interpreted without careful consideration of the features of the algorithm used (Hirzel and Le Lay 2008, Jiménez-Valverde et al. 2008, Soberón and Nakamura 2009) and of the ecological setting (Austin 2002).

In particular, the roles of **B** and **M** must be assessed carefully. In the following sections I will discuss briefly these problems.

Mean field competition

ENM and SDM seem able to capture a significant amount of ecological signature despite the fact that biotic data are seldom included (Elith and Leathwick 2009). How can this apparent puzzle be explained? Several authors have proposed the hypothesis that at large scales (coarse resolutions and large extents), the biotic component may be dominated by the environmental (Shmida and Wilson 1985, Pearson and Dawson 2003, Soberón 2007), although there is mounting evidence that, at least in certain cases, biotic interactions are significant at large extents (Bullock et al. 2000, Leathwick and Austin 2001). Below, I explore this question using one simple but illuminating special case of eq. (1). When competition is via a ''mean field'' that aggregates the effect of competitors via a single parameter (Molofsky et al. 1999, Solé and Bascompte 2006), local dominance of a single species is guaranteed and yet global coexistence is feasible.

For two species and no migration eq. (1) reduce to:

$$\frac{1}{x_{i,j}}\frac{dx_{i,j}}{dt} = r_{i,j}(\vec{e_j}) - a_{i,j}(x_{1,j} + x_{2,j})$$

where $a_{i,j}$ is the mean field effect of competitive effects on species i, in cell j. The very old theorem of competitive exclusion (Volterra 1931) states that locally (in cell j), with no migration, species 1 wins over species 2 if $(a_{2,j}r_{1,j} - a_{1,j}r_{2,j}) > 0$. This equation allows modeling a couple of competing virtual species. I used a realistic grid of 76 137 cells of 30 seconds of arc resolution (30″), centered on the Balsas Basin of Mexico. To model Grinnellian factors, I used a single scenopoetic variable per cell, the annual mean temperature T, obtained from the WorldClim (Hijmans et al. 2005) database, at the coarse resolution of 8 minutes of arc (8′). The intrinsic growth rates for species 1 and 2 were simple bell-shaped curves [$r_{1,j} = e^{-(T-27)^2/10000} - 0.95$ and $r_{2,j} = e^{-(T-24)^2/20000} - 0.95$], which means that species one prefers a narrower range of higher temperatures. Since they are based on the mean temperature layer at 8′ resolution, the values of the intrinsic growth rates are also coarse grained. The set of all 30″ cells for which $r_{1,\mathrm{j}} > 0$ corresponds to the **A** circle in the BAM diagram and is displayed in Fig. 2, in aggregated form (at 8′), as the cells with thick outlines.

The set of cells for which $a_{2,j}r_{1,j} - a_{1,j}r_{2},j > 0$ is the region **A**$\bigcap$**B**, where both Grinnellian and Eltonian factors are favorable to species 1 (it is assumed that the entire region is accessible to the two species, which means that **M** contains **A** and **B**), and therefore $\mathbf{G}_{O1} = \mathbf{A}\bigcap\mathbf{B}$. The parameters were calibrated in such a way that **A** for species 1 measures 35% of the total grid, and 68% of this region overlaps with the competitor species 2. To model the Eltonian factors, the bionomic, mean field competition parameters were simulated by randomly assigning values in the interval (0,1) to the high resolution 76 137 pairs of coefficients $a_{1,j}$ and $a_{2,j}$, using a bell-shaped beta distribution. Two scenarios were simulated: one in which the Grinnellian and the Eltonian parameters are spatially independent and therefore the values of both $a_{1,j}$ and $a_{2,j}$ are independent of the mean temperature, and another in which mean temperature and $a_{2,j}$ are negatively associated (simulating stronger competitive effects in the regions in which r_1 is higher). The above scheme ensures that in every 30″ cell, at most one species will be present. In other words, the fine-grained bionomic effect is very strong and presence depends simultaneously on the scenopoetic and bionomic variables.

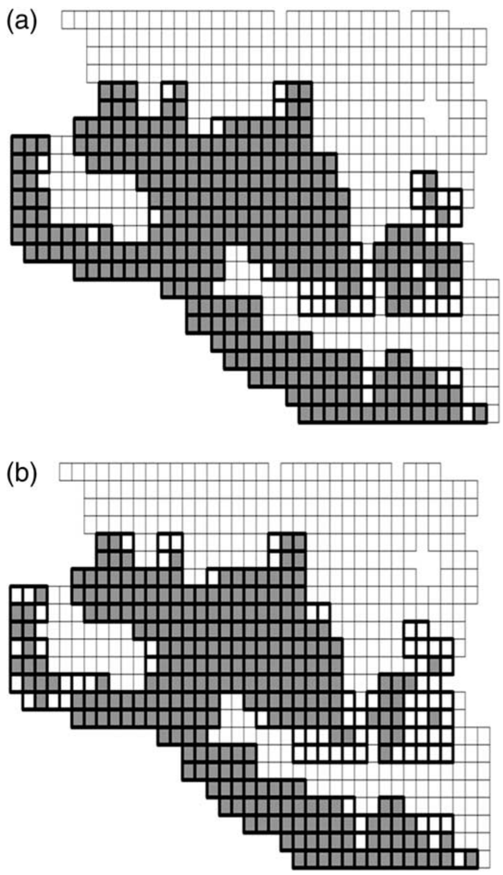

Figure 2. Results of the simulation of a virtual species. The thin lines represent the universe **G** in the BAM diagram, at the coarsest resolution (8′). The thick-outlined cells in the grid represent the region **A** where the species has a positive intrinsic growth rate (339 cells), calculated using a function of the scenopoetic variable (temperature). The gray squares contains at least one 30″ cell where the species survives the effects of: a random competitor (a), or of a competitor with effects correlated with the scenopoetic preferences of the focal species (b). At the resolution shown, there is a reduction of the original 100% scenopoetic area (39 and 72 out of 339 cells in the random competition field, and in the correlated competition field, respectively).

To explore the relative importance of the scenopoetic and the bionomic parameters, logistic regressions were performed on the area of distribution of species 1 ($\mathbf{G}o_1$), using SAM 3.0 software (Rangel et al. 2006). Spatial autocorrelation effects represent a significant problem in SDM (Dormann et al. 2007). I used a linear Trend Surface Analysis (TSA) on the latitude and longitude to

Table 1. Multiple regression models for presence of virtual species *1* as a function of bionomic (a_1, a_2) and scenopoetic (mean annual temperature) variables, at three spatial resolutions (see text). The smallest AIC value for each pair of models is in bold characters. The term TSA represents the spatial trend effect.

Model	Predictors	30″ (n = 7700)		4′ (n = 2554)		8′ (n = 674)	
		p	AIC	p	AIC	p	AIC
Temperature, a_1 and a_2 independent							
Full model	Mean temperature	<0.001	**311.742**	<0.001	892.001	<0.001	257.561
	a_1	<0.001		0.057		0.099	
	a_2	<0.001		0.727		0.92	
	TSA	0.609		0.127		0.094	
Scenopoetic model	Mean temperature	<0.001	1110.83	<0.001	**891.871**	<0.001	**256.332**
	TSA	0.596		0.088		0.062	
Temperature and a_2 correlated							
Full model	Mean temperature	<0.001	**467.476**	<0.001	934.112	<0.001	271.994
	a_1	<0.001		0.626		0.741	
	a_2	<0.001		0.952		0.08	
	TSA	0.92		0.586		0.067	
Scenopoetic model	Mean temperature	<0.001	1374.03	<0.001	**930.353**	<0.001	**271.318**
	TSA	0.495		0.566		0.081	

produce a spatial variable that was used as a predictor in the regressions (Rangel et al. 2006). Results appear in Table 1. At the three resolutions, two models were fitted: one using the scenopoetic and bionomic ($a_{1,j}$ and $a_{2,j}$) terms, and the spatial term (full model), and another using only the scenopoetic and the spatial term (scenopoetic model). Each of the above was repeated for uncorrelated and correlated $a_{2,j}$ (Table 1). The values of temperature and the bionomic parameters in the coarser cells were averaged over every contained smaller-resolution cell. At the highest resolution there are 76 137 cells, too many for the spatial analysis, therefore for this resolution I used a 10% random sample of the data. In all models, inspection of Moran's I correlograms show that the residual terms lack spatial correlation almost at all spatial lags. This suggests that the autocorrelation effect was successfully removed by the TSA term. As shown in Table 1, the effect of the bionomic terms is significant only at the highest resolution. At this resolution the Akaike information criterion (AIC) indicates better performance of the full models (Δ AIC = 799.09 and 906.55 for the uncorrelated and the correlated models, respectively). However, at lower resolutions the bionomic parameters are not significant, and the AIC indicates that the simpler, scenopoetic models performed as well or slightly better than the full models.

The above simple example is useful to illustrate how bionomic effects, essentially intermingled with a scenopoetic factor in determining local presences or absences, may become much less important when modeling distributions at coarser resolutions. This effect crucially will depend on the relative spatial structure of the two types of factors (Whittaker et al. 2001, Pearson and Dawson 2003). Therefore, in cases in which the Grinnellian and the Eltonian factors are uncorrelated in their spatial structure, coarse grained correlative niche modeling may still capture a strong biological signal. The example of this section highlights the need to document empirically the spatial structure of scenopoetic and bionomic factors.

The niche and distribution of *Baronia brevicornis*

As stated before, lack of true-absence data prevents direct estimation of the probability of presence conditioned to the environment, $P(Y = 1|\vec{e}_j)$, and thus hinders the simplest approach to estimating $\mathbf{G}_O$. Nevertheless, presences only data supplemented with hypothesis about **M** still may allow estimation of $\mathbf{G}_O$, as I will illustrate with an example. The butterfly *B. brevicornis* (Lepidoptera: Papilionidae), endemic to Mexico (Pérez-Ruiz 1977, León-Cortés 2004), feeds on legume species (mostly *Acacia cochliacantha*). *Baronia* is a species associated with the tropical deciduous forest. Although the principal food plant, the habitat and the right climatic conditions of the species are widespread in Mexico, the species is only found in the Balsas (*B. brevicornis brevicornis*) and central Chiapas (*B. brevicornis rufodiscalis*) basins, suggesting a limitation of dispersal power across the Neovolcanic mountainous belt of Mexico due to its limited flying capabilities (Pérez-Ruiz 1977). In what follows, I will not make a distinction between the two subspecies, because doing so will complicate the analysis, which is presented mainly as an illustration of a method.

To hypothesize the area of distribution ($\mathbf{G}_O$) of *B. brevicornis* using a presence only method, **A** is estimated using ENM, and hypothesis about the shapes of **B** and **M** are required. I downloaded 17 presence-data from the Inst. de Biología Insect Collection of Mexico, via the CONABIO (<www.conabio.gob.mx/remib_ingles/doctos/remib_ing.html>) web site. For the purpose of illustration I defined as **G** a large geographical area, depicted in Fig. 3c and 3d, and comprising from southern United States to northwestern South America. I used six (annual precipitation, precipitation of the wettest and the driest months, mean yearly temperature, mean temperatures of the coldest and the warmest months) bioclimatic variables out of 19 available in the WorldClim database (<www.worldclim.org/current.htm>). The variables were obtained at 2.5′ resolution (about 4.6 km of side). Figure 3a shows in black

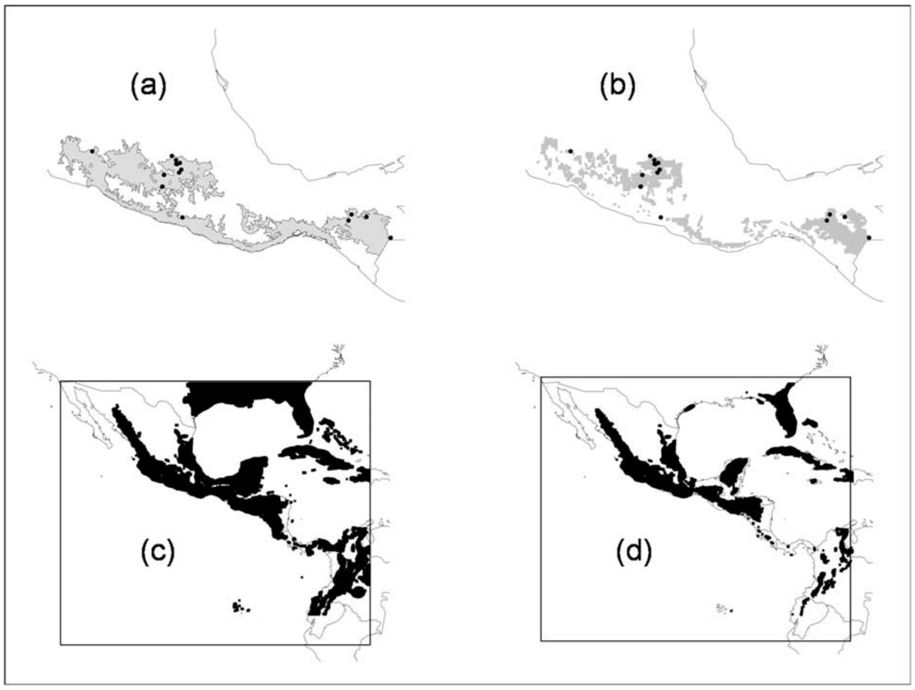

Figure 3. Data points and models for *Baronia brevicornis*. In (a) the presence points for *B. brevicornis* are displayed against a grey zone that represents the ecoregions of Mexico in which the butterfly has been reported, or **M** (see text). In (b) the intersection between **M** and the niche model using six variables is presented, with the presence localities as a reference. Results of a two (mean annual temperature and yearly precipitation) variables (c), and six variables (d) BIOCLIM algorithms.

dots the 17 presence data, together with a hypothesis about the region **M**, which for this example consists of all the ecoregions (CCA 1997) of Mexico in which both subspecies of *B. brevicornis* have been observed. Ecoregions are partially defined by biogeographical considerations and reflect the history of the changing distributions of entire biotas, therefore a butterfly of an old lineage found within an ecoregion may have historically experienced its entire area. Other hypothesis for **M** may be used (for instance river basins, or areas surrounded by mountains) but *B. brevicornis* is an old species and it seems appropriate to resort to historical criteria to define its accessibility region **M**.

For the purpose of illustration, I resort to the simplest and most transparent ENM algorithm: BIOCLIM (Busby et al. 1991). BIOCLIM simply draws a hypercube in **E**-space around a given percentile of datapoints, and then a GIS searches for spatial localities with a $\vec{e}_j$ vector within the defined region. This predicted region is symbolized by $\hat{\mathbf{G}}$. I used the DIVA-GIS (Hijmans et al. 2001) software to perform the operations, and ArcMap 9.3 for post-processing. Predictions at 100% point inclusion using every combination of 1 to six predictor variables were obtained for a total of 63 models of increasing complexity. Figure 3c shows one of the predictions using only two variables (mean temperature and yearly precipitation), and Fig. 3d the combination using the six variables. The very logic of BIOCLIM implies that every new variable added to a model can only reduce the size of the predicted region (Beaumont et al. 2005), but this happens in a non-linear way, converging to a bottom value that in this case is in the order of 50 000 cells. It is not immediately apparent to what region in the BAM $\hat{\mathbf{G}}$ corresponds. In order to interpret $\hat{\mathbf{G}}$ correctly more information is needed. Notice that although the geographical extents of the environments displayed in Fig. 3c, d are quite contrasting, in both cases the areas expressed contain regions where *B. brevicornis* has never been observed. This feature of presence-only methods to "overpredict" has caused much confusion, in part due to lack of explicit agreement about what exactly is $\hat{\mathbf{G}}$ estimating. In itself, $\hat{\mathbf{G}}$ can only be considered a hypothesis about $\mathbf{G}_O$ under very specific cases. In general $\hat{\mathbf{G}}$ should be regarded as an estimate of an area bounded by $\mathbf{G}_O$ and **A**: $\mathbf{G}_O \subseteq \hat{\mathbf{G}} \subseteq \mathbf{A}$. As such, the simple, unprocessed output of the BIOCLIM (in general, of presence–only methods) is best interpreted as an area potentially suitable for the species ($\mathbf{G}_O \bigcup \mathbf{G}_I$ in Fig. 1). Lacking absence information, however, it is still possible to use $\hat{\mathbf{G}}$ to model $\mathbf{G}_O$. This is done by reducing $\hat{\mathbf{G}}$ using knowledge or hypothesis about **M** and **B**. It is worth stressing that if unbiased true-absences were available, $\mathbf{G}_O$ could be modeled directly without need to postulate hypothesis about **M** and **B**.

Since the observations of presences come, by hypothesis, from the $\mathbf{G}_O$ area (we assume no sink populations were sampled) then they should include the effects of **B** and **M**. First consider **B**. The butterfly has no known competitors, and only generalist predators (Pentatomid bugs and Asiilid flies) or parasitoids (Trichogrammatidae wasps) (Pérez-Ruiz 1971). Moreover, its food plants, *Acacia* spp. are widely distributed over most arid tropical regions of Mexico covering a large portion (not shown) of the region depicted in Fig. 3c, d. This information is consistent with the hypothesis that $\mathbf{B} \subseteq \mathbf{A}$ and therefore the BIOCLIM output $\hat{\mathbf{G}}$ displayed in Fig. 3c, d (based on six variables) represent a hypothesis about a region contained in **A**. Checking this hypothesis would require per force either experimental work or mechanistic modeling.

Regardless of whether the areas in Fig. 3c, d are small or large, the BAM diagram indicates that an **M** region should be included in the analysis. Although this is seldom done explicitly in the literature, when attempting to model an actual area of distribution using presence–only methods, there is simply no way of avoiding reducing the potential region $\hat{\mathbf{G}}$ from a potential to some form of actual distribution. One way is by intersecting $\hat{\mathbf{G}}$ and **M**. Figure 3a depicts **M** estimated using ecoregions where *B. brevicornis* has been observed. The intersection of **M**, and the niche estimation using the six-variables BIOCLIM (Fig. 3d) produces the region $\hat{\mathbf{G}} \cap \mathbf{M}$ (of 6905 cells of size) shown in Fig. 3b. Essentially, the method found a region within **M** with similar environmental combinations to those where the species has been observed. Intersection with **M** drastically reduces the ''overprediction'' obtained by the Bioclim algorithm, as can be seen comparing Fig. 3b and d. It is very important to notice the different interpretation due to inclusion of **M**. Without it, we have a BIOCLIM prediction of an area where environments are potentially favorable to the species. Intersection with **M** reduces that potential region to one hypothesized to be both suitable from a scenopoetic point of view and available to dispersal by the species.

Is $\hat{\mathbf{G}} \cap \mathbf{M}$ then an estimate of $\mathbf{G}_O$? This conclusion would depend on the truth of the three ancillary hypotheses: 1) bionomic suitability is spatially contained within scenopoetic suitability ($\mathbf{B} \subseteq \mathbf{A}$), 2) that the ecoregions estimate of **M** really reflects dispersal restrictions and 3), that the observed occurrences encompass an unbiased sample of the relevant environmental space.

Presence–only methods do not go beyond this. These methods, which are attractive due to the massive availability of presence-only data, can only be assumed to estimate areas $\hat{\mathbf{G}}$ such that $\mathbf{G}_O \subseteq \hat{\mathbf{G}} \subseteq \mathbf{A}$, and ancillary hypothesis and post-processing of the results, as exemplified by the *B. brevicornis* example, may yield approximations to $\mathbf{G}_O$. It is only unbiased presence-absence correlative methods that can estimate directly $\mathbf{G}_O$, and thus the realized niche. And it is only mechanistic methods (Kearney and Porter 2009) that can provide direct estimates of the fundamental niche, and therefore of its geographic projection **A**. However, use of the BAM diagram facilitates clarifying otherwise implicit assumptions, and the interpretation of results of presence-only ecological niche modeling. In the case of *B. brevicornis*, the available data would suggest a simplified BAM configuration where **A** contains **B** and a much smaller region **M** is contained in both, and therefore $\hat{\mathbf{G}} \cap \mathbf{M} \approx \mathbf{G}_O$ For this particular BAM configuration, the area $\mathbf{G}_O$ is hypothesized to be much smaller than **A**, even when using six environmental variables, and mostly determined by limitations in the movements of the butterfly. In principle, transplant experiments can be used to check this type of predictions (Angert and Schemske 2005). It may well be the case that other unknown biological factors are also acting to prevent a larger occupation of the favorable area (Leathwick 1998), but the point of this section is that without explicit mention of the main factors that determine a distribution, correlative ENM produce very poorly defined outputs, which are difficult to interpret.

Conclusions

Predicting the area of distribution of a species, and estimating the set of environmental conditions present in it are very important problems in ecology, macroecology and biogeography. It is unfortunate that to a certain extent the literature is still mired in semantic arguments about whether those sets should be called niches or not, (Kearney 2006, Jiménez-Valverde et al. 2008), or the precise nature of the objects modeled by SDM and ENM. Using symbols, diagrams and terminology suggested by the population processes causing the area of distribution help significantly to clarify those semantic discussions. Austin (2002) has called for explicit inclusion of ecological knowledge in statistical modeling of species distributions. The scheme presented in the BAM diagram is a move in this direction because it represents explicit hypothesis relating three types of biological factors that determine the area of distribution (two levels of niche, and dispersal). The underlying logic is based on fundamental population processes that suggest rigorous definitions of all the terms, and clarification of nagging questions, like what is it specifically that correlative ENM calculates. I gave examples of how specific hypothesis about the spatial structure of the **B** and **M** factors allow, even presence–only methods, to estimate the area of distribution $\mathbf{G}_O$. When those hypotheses fail, ENM estimates other, less well defined niches and the corresponding regions in the map.

The scheme relies on a separation between what I have called the Grinnellian and the Eltonian factors. At the risk of repetition, I stress that both the mathematics of spatially-explicit population dynamics (eq. 1) and the empirical evidence available (Grinnell 1914, Leathwick 1998, Bullock et al. 2000, Leathwick and Austin 2001, Mackey and Lindenmayer 2001, Coudun et al. 2006, Heikkinen et al. 2007) show that it is the simultaneous interplay of all factors that, strictly speaking, determines population dynamics and therefore where individuals of a species can be observed. The distinction between scenopoetic and bionomic therefore, is just a simplification. However, ENM has been, in practice, based mostly on scenopoetic variables, related only to the circle **A** in the BAM diagram. This suggests that, at the spatial scale at which ENM mostly is used (very large extents and coarse resolutions relative to the size and movements of individuals), the Grinnellian factors can successfully capture a significant amount of biological signal, as the example using the virtual species illustrated. How much, and at what scales and for what groups remains a matter of empirical research.

If nothing else, the scheme of the BAM diagram and the hierarchical distinction between Grinnellian and Eltonian factors allows straightforward terminology that may enable clearer communication of concepts and less muddled debates. However, I suggest that the true merit of this simplified scheme is that it places the focus firmly on the fundamental ecological factors that determine species distributions, rather than on the statistical technicalities of algorithms used to estimate poorly-defined objects. Even when absence data are available, correlative methods are statistical exercises that require ecological interpretation. One way of providing this is by resorting to the underlying population dynamics processes that determine the spatial location of individuals, and thus the area of distribution of the species. Although, generally speaking, the equations of the dynamics may be impossible to solve, using them as heuristic devices clarifies the meaning of the outputs of ENM algorithms, and suggest testable explanations based on causal factors. In the end, the only satisfactory way of testing the very useful predictions of correlative models would be by field observation and experiments about the causal factors, and not by statistically juggling subsets of data.

Acknowledgements – I would like to thank Town Peterson, Alberto Jiménez and Jorge Lobo for many discussions on the topic of niches, Grinnellian or not. Yoshinori Nakazawa and Andrés Lira helped me deal with GIS problems, and A. Lira helped me with the figures and analysis of *Baronia brevicornis* data. T. Rangel clarified the use of SAM 3.0. I received very useful comments from many contributors to this symposium, including Mary Warren, Darren Kriticos, Michael Krabbe, Gosta Nachman, Joaquín Hortal and Jorge Lobo, and from three anonymous referees. Microsoft Research Grant # KUCR 47780 partially supported this work.

References

Angert, A. L. and Schemske, D. W. 2005. The evolution of species distributions: reciprocal transplants across the elevation ranges of *Mimulus cardinalis* and *M. lewisii*. – Evolution 59: 1671–1684.

Araújo, M. B. and Guisan, A. 2006. Five (or so) challenges for species distribution modelling. – Global Ecol. Biogeogr. 33: 1677–1688.

Austin, M. P. 1980. Searching for a model to use in vegetation analysis. – Vegetatio 42: 11–21.

Austin, M. P. 2002. Spatial prediction of species distribution: an interface between ecological theory and statistical modelling. – Ecol. Model. 157: 101–118.

Bahn, V. and McGill, B. J. 2007. Can niche-based distribution models outperform spatial interpolation? – Global Ecol. Biogeogr. 16: 733–742.

Beale, C. et al. 2008. Opening the climate envelope reveals no macroscale associations with climate in European birds. – Proc. Nat. Acad. Sci. USA 105: 14908–14912.

Beaumont, L. J. et al. 2005. Predicting species distributions: use of climatic parameters in BIOCLIM and its impact on predictions of species' current and future distributions. – Ecol. Model. 186: 250–269.

Begon, M. et al. 2006. Ecology: from individuals to ecosystems. – Blackwell.

Brotons, L. et al. 2004. Presence–absence versus presence–only modelling methods for predicting bird habitat suitability. – Ecography 27: 437–448.

Brown, J. H. et al. 1996. The geographic range: size, shape, boundaries and internal structure. – Annu. Rev. Ecol. Syst. 27: 597–623.

Bullock, J. M. et al. 2000. Geographical separation of two *Ulex* species at three spatial scales: does competition limit species' ranges? – Ecography 23: 257–271.

Busby, J. R. et al. 1991. BIOCLIM – a bioclimate analysis and prediction system. – In: Margules, C. R. and Austin, M. P. (eds), Nature conservation. Cost Effective Biological Surveys and Data Analysis, Canberra, Australia.

CCA 1997. Ecological regions of North America. – Comisión de Cooperación Ambiental de América del Norte, Montreal.

Chase, J. M. and Leibold, M. 2003. Ecological niches: linking classical and contemporary approaches. – Univ. of Chicago Press.

Coudun, C. et al. 2006. Soil nutritional factors improve models of plant species distribution: an illustration with *Acer campestre* (L.) in France. – J. Biogeogr. 33: 1750–1763.

Dormann, C. F. et al. 2007. Methods to account for spatial autocorrelation in the analysis of species distributional data: a review. – Ecography 30: 609–628.

Edwards, J. 2004. Research and societal benefits of the global biodiversity information facility. – BioScience 54: 485–486.

Elith, J. and Graham, C. 2009. Do they? How do they? WHY do they differ? On finding reasons for differing performances of species distribution models. – Ecography 32: 66–77.

Elith, J. and Leathwick, J. 2009. Species distribution models: ecological explanation and predictionacross space and time. – Annu. Rev. Ecol. Evol. Syst. 40: 677–697.

Elith, J. et al. 2006. Novel methods improve prediction of species' distributions from occurrence data. – Ecography 29: 129–151.

Feria, P. and Peterson, A. T. 2002. Prediction of bird community composition based on point-occurrence data and inferential algorithms: a valuable tool in biodiversity assessments. – Divers. Distrib. 8: 49–56.

Grinnell, J. 1914. Barriers to distribution as regards birds and mammals. – Am. Nat. 48: 248–254.

Guisan, A. and Zimmermann, N. 2000. Predictive habitat distribution models in ecology. – Ecol. Model. 135: 147–186.

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

Hassell, M. et al. 1998. Top-down versus bottom-up and the Ruritanian bean bug. – Proc. Nat. Acad. Sci. USA 95: 10661–10664.

Heikkinen, R. K. et al. 2007. Biotic interactions improve prediction of boreal bird distributions at macroscales. – Global Ecol. Biogeogr. 16: 754–763.

Hijmans, R. J. et al. 2001. Computer tools for spatial analysis of plant genetic resources data: DIVA-GIS. – Plant Genet. Resour. Newslett. 127: 15–19.

Hijmans, R. J. et al. 2005. Very high resolution interpolated climate surfaces for global land areas. – Int. J. Climatol. 25: 1965–1978.

Hirzel, A. H. and Le Lay, G. 2008. Habitat suitability modelling and niche theory. – J. Appl. Ecol. 45: 1372–1381.

Hutchinson, G. E. 1957. Concluding remarks. – Cold Spring Harbor Symp. Quant. Biol. 22: 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 (Suppl.): 194–220.

James, F. C. et al. 1984. The Grinnellian niche of the wood thrush. – Am. Nat. 124: 17–47.

Jiménez-Valverde, A. et al. 2008. Not as good as they seem: the importance of concept in species distribution modelling. – Divers. Distrib. 14: 885–890.

Kearney, M. 2006. Habitat, environment and niche: what are we modelling? – Oikos 115: 186–191.

Kearney, M. and Porter, W. P. 2009. Mechanistic niche modelling: combining physiological and spatial data to predict species' ranges. – Ecol. Lett. 12: 334–350.

Keating, K. A. and Cherry, S. 2004. Use and interpretation of logistic regression in habitat-selection studies. – J. Wildl. Manage. 68: 774–7789.

Kriticos, D. and Randall, R. P. 2001. A comparison of systems to analyze potential weed distributions. – In: Groves, R. H. et al. (eds), Weed risk assessment. CSIRO Publ.

Leathwick, J. 1998. Are New Zealand's *Nothofagus* species in equilibrium with their environment? – J. Veg. Sci. 9: 719–732.

Leathwick, J. R. and Austin, M. 2001. Competitive interactions between tree species in New Zealand's old-growth indigenous forest. – Ecology 82: 2560–2573.

Leibold, M. 1996. The niche concept revisited: mechanistic models and community context. – Ecology 76: 1371–1382.

León-Cortés, J. 2004. Complex habitat requirements and conservation needs of the only extant Baroniinae swallowtail butterfly. – Anim. Conserv. 7: 241–250.

Lobo, J. M. et al. 2010. The uncertain nature of absences and their importance in species distribution modelling. – Ecography 33: 103–114.

MacArthur, R. 1972. Geographical ecology. – Harper and Row.

Mackey, B. and Lindenmayer, D. B. 2001. Towards a hierarchical framework for modelling the spatial distribution of animals. – J. Biogeogr. 28: 1147–1166.

Meszena, G. et al. 2006. Competitive exclusion and limiting similarity: a unified theory. – Theor. Popul. Biol. 69: 68–87.

Molofsky, J. et al. 1999. Local frequency dependence and global coexistence. – Theor. Popul. Biol. 55: 270–282.

Odling-Smee, F. J. et al. 2003. Niche construction. The neglected process in evolution. – Princeton Univ. Press.

Pearson, R. G. and Dawson, T. P. 2003. Predicting the impacts of climate change on the distribution of species: are bioclimatic envelopes useful? – Global Ecol. Biogeogr. 12: 361–371.

Pérez-Ruiz, H. 1971. Algunas consideraciones sobre la población de *Baronia brevicornis* (Slv. (Lepidoptera, baroniinae)) en la región de Mezcala, Guerrero. – Anales del Inst. de Biol., U.N.A.M. 42 Ser. Zool. 1: 63–72.

Pérez-Ruiz, H. 1977. Distribución geográfica y estructura poblacional de *Baronia brevicornis* Salv. (Lepidoptera, Papilionidae, Baroniinae) en la República Mexicana. – Anales del Inst. de Biol., U.N.A.M. 48: 151–164.

Peterson, A. T. 2003. Predicting the geography of species' invasions via ecological niche modeling. – Q. Rev. Biol. 78: 419–433.

Peterson, A. T. 2006. Uses and requirements of ecological niche models and related distributional models. – Biodivers. Inform. 3: 59–72.

Phillips, S. 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. 2000. On the relationship between niche and distribution. – Ecol. Lett. 3: 349–361.

Rangel, T. F. et al. 2006. Towards and integrated computational tool for spatial analysis in macroecology and biogeography. – Global Ecol. Biogeogr. 15: 321–327.

Raxworthy, C. J. et al. 2003. Predicting distributions of known and unknown reptile species in Madagascar. – Nature 426: 837–841.

Sagarin, R. et al. 2006. Moving beyond assumptions to understand abundance distributions across ranges of species. – Trends Ecol. Evol. 21: 524–530.

Sanchez-Cordero, V. and Martínez-Meyer, E. 2000. Museum specimen data predict crop damage by tropical rodents. – Proc. Nat. Acad. Sci. USA 97: 7074–7077.

Shmida, A. and Wilson, M. V. 1985. Biological determinants of species diversity. – J. Biogeogr. 12: 1–20.

Soberón, J. 2007. Grinnellian and Eltonian niches and geographic distributions of species. – Ecol. Lett. 10: 1115–1123.

Soberón, J. and Peterson, A. T. 2004. Biodiversity informatics: managing and applying primary biodiversity data. – Phil. Trans. R. Soc. B 35: 689–698.

Soberón, J. and Peterson, A. T. 2005. Interpretation of models of fundamental ecological niches and species' distributional areas. – Biodivers. Inform. 2: 1–10.

Soberón, J. and Nakamura, M. 2009. Niches and distributional areas: concepts, methods and assumptions. – Proc. Nat. Acad. Sci. USA 106: 19644–19650.

Solé, R. and Bascompte, J. 2006. Self-organization in complex ecosystems. – Princeton Univ. Press.

Sutherst, R. W. and Maywald, G. F. 1985. A computerised system for matching climates in Ecology. – Agric. Ecosyst. Environ. 13: 281–299.

Thompson, J. N. 2005. The geographic mosaic of coevolution. – Chicago Univ. Press.

Thuiller, W. et al. 2009. BIOMOD – a platform for ensamble forecasting of species distributions. – Ecography 32: 369–373.

Vandermeer, J. 1972. Niche theory. – Annu. Rev. Ecol. Syst. 3: 107–132.

Volterra, V. 1931. Variations and fluctuations of the number of individuals in animal species living together (translation from the Italian original). – In: Whittaker, R. H. and Levin, S. A. (eds), Niche: theory and applications. Dowden, Hutchinson and Ross, Stroudsburg, PA.

Ward, G. et al. 2009. Presence–only data and the EM algorithm. – Biometrics 65: 554–563.

Whittaker, R. J. et al. 2001. Scale and richness: towards a general hierarchical theory of species diversity. – J. Biogeogr. 28: 453–470.