Ecological Modelling 248 (2013) 57–70

Contents lists available at SciVerse ScienceDirect

Ecological Modelling

journal homepage: www.elsevier.com/locate/ecolmodel

Comparing the relative contributions of biotic and abiotic factors as mediators of species' distributions

Constantino González-Salazar[a,c,*], Christopher R. Stephens[b,c,1], Pablo A. Marquet[d,e,f,2]

[a] Instituto de Biología, Universidad Nacional Autónoma de México, 04510 México, D.F., Mexico
[b] Instituto de Ciencias Nucleares, Universidad Nacional Autónoma de México, 04510 México, D.F., Mexico
[c] C3 – Centro de Ciencias de la Complejidad, Universidad Nacional Autónoma de México, 04510 México, D.F., Mexico
[d] Center for Advanced Studies in Ecology and Biodiversity (CASEB) and Departamento de Ecología, Pontificia Universidad Católica de Chile, Casilla 114-D, Santiago, Chile
[e] Instituto de Ecología y Biodiversidad (IEB) Casilla 653, Santiago, Chile
[f] The Santa Fe Institute, 1399 Hyde Park Road, Santa Fe, NM 87501 USA

ARTICLE INFO

Article history:
Received 6 February 2012
Received in revised form 1 September 2012
Accepted 7 October 2012
Available online 9 November 2012

Keywords:
Biotic interaction
Data mining
Environmental data
Ecological niche
Ecological modeling
Species distribution

ABSTRACT

Models to predict species' ranges have chiefly been limited to abiotic variables. However, the full ecological niche depends on a myriad of factors, both biotic and abiotic, that often correspond to completely different data types. We applied a methodology based on data mining techniques to construct ecological niche models composed of biotic as well as abiotic variables using three quite different sets of variables: climatic layers, maps of land cover and point collections of Mexican mammals. We show how potential ecological interactions can be inferred from geographic data using co-occurrences as proxies, and generate corresponding distribution models. We consider two case studies: an insect genus (*Lutzomyia* sp.) and a mammal species (*Lynx rufus*). We show that for both examples model predictability is higher using biotic versus abiotic variables, but even higher when both variable types are integrated together. Also, by identifying those variables that are most relevant in describing the suitable (niche) and unsuitable (anti-niche) areas we can establish an ecological profile for any geographic location and quantify the relative influence of each location and its impact on species. In conclusion, we show that including both abiotic and biotic factors not only leads to a fuller more comprehensive understanding of the niche, but also leads to more accurate prediction models.

1. Introduction

One of the most important goals in ecology and biogeography studies is to identify the principal factors that constrain the range of species. An understanding of these factors and their relative impact will permit us to better understand and model both current and future distributions (Lomolino et al., 2005; Araujo and Guisan, 2006; Araujo and Luoto, 2007). Unfortunately, the number of potential factors that can affect species' distributions is enormous. However, a great deal of information associated with many of these factors is now available in online databases. Thus, the most advanced techniques for describing and predicting species' distributions are nowadays based on data mining, where large volumes of observational data are systematically explored using different mathematical models, from standard regression type models to sophisticated artificial intelligence techniques. Such methods have recently been used in various ecological applications, such as biodiversity studies, modeling biological invasions and species distributions (Stockwell and Peters, 1999; Guralnick and Pearman, 2009).

Current niche modeling has chiefly been limited to abiotic variables, due to the difficulties of incorporating information associated with biotic interactions (Guisan and Thuiller, 2005; Araujo and Guisan, 2006). Recently, however, a methodology based on data mining techniques has been developed and applied in ecogeographic studies (Sánchez-Cordero et al., 2008; Stephens et al., 2009) that naturally facilitate the incorporation of biotic factors associated with point collection data. Although the methodology is general, what was not explicitly considered in those papers was the question of how to integrate together different data types, including both abiotic and biotic factors, thus permitting a deeper insight into the relative importance of the different types of factor in determining species' distributions.

Although modeling the range of a species without further insight into what factors potentially affect the range can be a useful goal, much greater insight can be gained by determining and

* Corresponding author at: Instituto de Biología, Universidad Nacional Autónoma de México, 04510 México, D.F., Mexico. Tel.: +52 555 623 0222x47880.
E-mail addresses: cgs@ibiologia.unam.mx (C. González-Salazar), stephens@nucleares.unam.mx (C.R. Stephens), pmarquet@bio.puc.cl (P.A. Marquet).
[1] Tel.: +52 555 622 4692; fax: +52 555 622 4693.
[2] Tel.: +56 2 6862639.

0304-3800/$ – see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.ecolmodel.2012.10.007

understanding which factors, barriers or biotic interactions are important for a particular species in a particular geographical location (Brown, 1995; Brown et al., 1996; Arntzen and Themudo, 2008). In other words, to better understand the relation between the geographical distribution of the species and its niche. Such considerations, for instance, could help to cover the current gap between local ecological observations and other regional processes (Guisan and Zimmermann, 2000; Pearson and Dawson, 2003; Guisan et al., 2006).

The main contribution of this article is to present a modeling framework in which both abiotic and biotic factors associated with different data types can be integrated together into models that lead to more accurate predicted distributions and to a fuller understanding of the corresponding niches of a given taxon. The methodology is an extension of that proposed in Stephens et al. (2009), where only biotic variables were included. We will show how to use such models to achieve the above stated goals of predicting taxa' distributions using a more complete set of potential niche variables, as well as quantifying the relative importance not only of biotic versus abiotic factors in explaining the presence of a taxon in a particular geographic location, but also of specifying exactly which factors are the most important.

Although our approach is quite general, specific results, such as the distribution of particular taxon and the associated niche variables, depend on the taxon being studied. We will therefore illustrate the method by determining the predictive power of regional variables in predicting the distribution and quantifying and characterizing the ecological niches of two very different taxa, the genus *Lutzomyia* (sandfly) and the species *Lynx rufus* (bobcat), describing their ecological profiles and the role played by different biotic and abiotic factors as constraints on their distributions. Note that, instead of choosing two distinct species here as our examples, we consider a species and a genus, thereby showing that the modeling framework we present is applicable to different taxonomic levels. One might, of course, ask whether or not it is appropriate to talk of predicting the distribution and the associated niche of a taxon other than a species. For instance, if the niches of different species within a genus are quite distinct, it may well be that the genus niche is so smeared out it has no characterizing features and subsequently the corresponding distribution model is only weakly predictive. The very fact that our results are positive – leading to an accurate predictive model and a characteristic niche for the genus Lutzomyia – show that, in this case at least, it makes perfect ecological sense to model at the genus level. This does not, of course, guarantee that this will always be the case.

So, in the first case study we consider data associated with actual and potential reservoirs of an important emerging disease, Leishmaniasis, a disease widely distributed in tropical regions that is transmitted by sandflies. As Leishmaniasis is a zoonotic tropical disease, sylvan reservoirs are crucial to the maintenance of the parasite in ecological communities and, further, are intimately associated with human transmission (Wolfe et al., 2007). For our second case study, we consider the effect of abiotic and biotic factors on the distribution of the carnivore *Lynx rufus*, given that it has not been collected south of the Isthmus of Tehuantepec in Mexico, in spite of the fact that there are geographic regions below the Isthmus that coincide with the ecological requirements associated with its fundamental niche (Sánchez-Cordero et al., 2008).

By using two such contrasting taxa (an insect and a mammal) we can illustrate the scope of this methodology in inferring different types of inter-specific interactions (e.g., competition, mutualism, commensalism), and their relative importance for taxa distributions. Additionally, we can directly compare and contrast the role of climatic factors to that of biotic interactions to determine which variables most affect the presence or absence of a species and in what way the prediction of species' ranges could be improved (Guisan et al., 2006; Heikkinen et al., 2007).

2. Materials and methods

In applying our methodology we used a class of biotic variables – collection data for Mexican mammals – as well as abiotic (climatic) variables (e.g. temperature, precipitation) and land cover. Thus, models were built using four sets of explanatory variables: (1) abiotic variables only; (2) mammals species only; (3) land cover only; and (4) abiotic variables, mammals and land cover together.

2.1. Data types

2.1.1. Abiotic variables

2.1.1.1. Climatic data. Nineteen bioclimatic variables were used as environmental layers (Table 1) obtained from WorldClim with a spatial resolution of 30″ (http://www.worldclim.org/current.htm; Hijmans et al., 2005). They represent annual trends (e.g., annual mean temperature and precipitation), seasonality (e.g. annual temperature and precipitation ranges), environmental extremes (e.g., highest and lowest values of temperature for the warmest and coolest months) of temperature and precipitation, and characterize the dimensions of climate considered particularly relevant in determining species distributions (Waltari and Guralnick, 2009).

2.1.2. Biotic variables

2.1.2.1. Land cover. We used the Inventario Nacional Forestal (INF) 2000 (Palacio et al., 2000) as a base for current land use and vegetation types in México. INF 2000 is based on both LandSat satellite imagery interpretation and ground field validation of the main vegetation types and land use in Mexico and scaled at 1:250,000. It is jointly produced by the Instituto de Geografía of the Universidad Nacional Autónoma de México (www.igeograf.unam.mx), and the Instituto Nacional de Estadística y Geografía (www.inegi.gob.mx). This layer included 77 types of vegetation in México.

2.1.2.2. Species occurrence data. The data set consisted of point collection data associated with one Class, Mammalia, including the species – *Lynx rufus* – and one genus – *Lutzomyia* – of the class Insecta. The mammal data set contains 37,297 unique point collections from georeferenced localities for 427 terrestrial mammals occurring in Mexico. Data are based on museum voucher specimens from national and international collections, public electronic databases (GBIF; www.gbif.org, and CONABIO; www.conabio.gob.mx), and published records (Hall, 1981; Guevara-Chumacero et al., 2001). For *L. rufus* there were 220 collections points. For *Lutzomyia*, there were 270 collections points belonging to 11 species (see supplementary material) taken from published literature and from national collections: Instituto de Diagnóstico y Referencia Epidemiológica (InDRE, Mexico City), the Colección Entomológica Regional Universidad Autónoma de Yucatán (UADY, Mérida) and the Laboratorio de Medicina Tropical at the Universidad Nacional Autónoma de México (UNAM, Mexico City). For all data sets, each locality was georeferenced to the nearest 0.01 degrees of latitude and longitude using 1:250,000 topographic maps (INEGI; www.inegi.gob.mx; Instituto de Geografía, Universidad Nacional Autónoma de México; www.igeograf.unam.mx). Point collection data was, of course, not collected in order to provide and unbiased sampling of underlying species abundance and therefore must be considered carefully to understand potential statistical biases that might be present. The utility and limitations of point collection data have been amply discussed in (Ponder et al., 2001; Graham et al., 2004).

With respect to the data set for Mexican mammals, this data has been collected over a period of more than 100 years with a

Table 1
Bioclimatic variables from WorldClim: BIO1 = annual mean temperature; BIO2 = mean diurnal range (mean of monthly (max temp–min temp)); BIO3 = isothermality [((BIO2/BIO7) × 100)]; BIO4 = temperature seasonality (standard deviation × 100); BIO5 = max temperature of warmest month; BIO6 = min temperature of coldest month; BIO7 = temperature annual range (BIO5–BIO6); BIO8 = mean temperature of wettest quarter; BIO9 = mean temperature of driest quarter; BIO10 = mean temperature of warmest quarter; BIO11 = mean temperature of coldest quarter; BIO12 = annual precipitation; BIO13 = precipitation of wettest month; BIO14 = precipitation of driest month; BIO15 = precipitation seasonality (coefficient of variation); BIO16 = precipitation of wettest quarter; BIO17 = precipitation of driest quarter; BIO18 = precipitation of warmest quarter; BIO19 = precipitation of coldest quarter. These bioclimatic variables were derived from the average monthly mean temperature (°C × 10), average monthly minimum temperature (°C × 10), average monthly maximum temperature (°C × 10) and average monthly precipitation (mm) (Hijmans et al., 2005).

Range	BIO1	BIO2	BIO3	BIO4	BIO5	BIO6	BIO7
R1	−27–5	73–97	37–44	210–984	38–76	−98–65	115–166
R2	6–37	98–108	45–48	985–1759	77–114	−64–32	167–189
R3	38–70	109–119	49–51	1760–2534	115–152	−31–1	190–214
R4	71–102	120–130	52–55	2535–3309	153–190	2–34	215–238
R5	103–135	131–141	56–60	3310–4084	191–229	35–67	239–262
R6	136–167	142–153	61–64	4085–4859	230–267	68–100	263–284
R7	168–199	154–164	65–67	4860–5634	268–305	101–133	285–306
R8	200–232	165–174	68–71	5635–6409	306–343	134–166	307–329
R9	233–264	175–184	72–76	6410–7184	344–381	167–199	330–355
R10	265–297	185–207	77–84	7185–7959	382–420	200–232	356–392

	BIO8	BIO9	BIO10	BIO11	BIO12	BIO13	BIO14
R1	−22–11	−35–2	−20–14	−36–4	42–507	8–84	0–12
R2	12–45	−1–31	15–48	−3–28	508–973	85–161	13–25
R3	46–79	32–64	49–82	29–60	974–1439	162–237	26–37
R4	80–113	65–97	83–117	61–92	1440–1905	238–314	38–50
R5	114–147	98–131	118–151	93–125	1906–2371	315–391	51–63
R6	148–181	132–164	152–185	126–157	2372–2836	392–467	64–75
R7	182–215	165–197	186–220	158–189	2837–3302	468–544	76–88
R8	216–249	198–230	221–254	190–221	3303–3768	545–620	89–100
R9	250–283	231–263	255–288	222–253	3769–4234	621–697	101–113
R10	284–317	264–297	289–323	254–286	4235–4700	698–774	114–126

	BIO15	BIO16	BIO17	BIO18	BIO19
R1	37–45	18–218	0–43	1–125	0–95
R2	46–54	219–418	44–87	126–249	96–191
R3	55–63	419–618	88–131	250–373	192–287
R4	64–72	619–818	132–175	374–497	288–383
R5	73–81	819–1018	176–219	498–622	384–479
R6	82–89	1019–1218	220–262	623–746	480–575
R7	90–98	1219–1418	263–306	747–870	576–671
R8	99–107	1419–1618	307–350	871–994	672–767
R9	108–116	1619–1818	351–394	995–1118	768–1016
R10	117–125	1819–2019	395–438	1119–1243	1017–1927

consequently large number of collectors. Hence, although the data has not been collected systematically, it has probably led to an adequate sampling. Additionally, mammals are the best known and collected group in Mexico. In the case of *Lutzomyia* the coverage is less but still represents the best available.

2.2. Data integration

In Stephens et al. (2009) attention was restricted to only biotic factors. When considering the effects of both biotic and abiotic factors together subtleties arise when trying to integrate the different data types into a predictive model. One of the principal problems to face is associated with the question of to what extent they can be compared statistically. An important data class representing biotic factors is that of point collection data, which generally translates into discrete Boolean presence/no presence type variables. On the other hand, abiotic factors, such as temperature, are generally continuous in nature. How does one compare the relative effects of such different variable types in a predictive model, given that they have such different natural scales of variation and spatial resolution?

There are two fundamental characteristics of the data – the data type and its spatial–temporal resolution – that need to be made compatible within a given modeling framework. As far as variable type is concerned, abiotic factors are typically real numbers, whereas point collection data are naturally modeled as Boolean variables. There are at least two possible ways to make a fairer comparison and integrate them together: One is to extrapolate the biotic factors to become continuous variables, while, the other, is to convert the abiotic variables to Boolean type. The first of these alternatives would be basically equivalent to what is done in the case of the climatic variables themselves: a discrete set of point measurements are converted into a continuous distribution by assuming a model that interpolates from one to the other. The question there would be: How is the interpolation done? A better way, that requires less model bias, is to convert the abiotic variables to a discrete set of values. There are several ways to do this.[3] One way is to coarse grain the value of the continuous variable into a finite set of bins. We can then consider the variable to take Boolean values, as in a given region the question as to whether the variable takes a particular value in a bin has a yes/no answer, e.g., mean annual temperature is in the range 13–18°.

The next question concerns the relevant spatial scale for the variables. First, we divide up a geographic region of interest into spatial cells, x_α, such as a grid of uniform square cells. For a pair of Boolean variables we can then count the co-occurrences of those variables in a given cell and test the frequency of those co-occurrences against some null hypothesis. The question is: What is an appropriate resolution, an appropriate cell size? This is known in geography as the "modifiable areal unit problem" (MAUP)

[3] We have repeated our analysis using different coarse graining and found no qualitative change in our results. There is more discussion of this point in supplementary material.

(Openshaw, 1983). In terms of forming a spatial grid, there are at least two important considerations (Alcocer and Stephens, 2012): The sizes of the statistical samples of the variables and their degree of correlation. Too fine a grid and there will be no co-occurrences, too rough and there will be little to no discrimination.

For a given grid size, chosen by consideration of the relative sample sizes of the biotic variables, one can look for a coarse grained value of each abiotic variable within a given cell. As mentioned, there are different ways to do this. Here, we will consider the presence of a given value of a given abiotic variable to be equivalent to the "presence" of that variable in the corresponding cell. Thus, if in a given cell, values 1 and 2 of Annual Mean Temperature appear, then for the purposes of calculating co-occurrences, it is as if there were "collections" of these variables. Thus, the 19 Worlclim variables, each with 10 possible values, can be thought of as 190 different "species" and compared directly with the collection data.

We used 3337 square cells of linear size 25 km, which corresponds to an average of 20-point collections per cell (once again, for details and a discussion of the dependence on the grid size, see Stephens et al., 2009; Alcocer and Stephens, 2012). We consider B_i (x_α) as a measure of the presence of the taxa i in the spatial cell x_α. The particular measures, B_i, we can use depend on the availability of data – presence only, presence/absence, abundance, etc. Our main object of interest is $P(B_i(x_\alpha)|I(x_\alpha))$, the probability that the distribution measure $B_i(x_\alpha)$ takes a certain value in the spatial cell x_α conditioned on, $I(x_\alpha)$, which is composed of all biotic and abiotic factors that affect species distributions corresponding with their niche (Soberón and Peterson, 2004).

2.2.1. Co-occurrences as proxies for ecological interactions

We adopt a non-parametric "data mining" approach, modeling a species' distribution directly using available biotic and abiotic data, the former being a direct result of the past and present interactions of all relevant causative factors - climatic, phylogenetic, co-evolutionary and ecological.

We will take the taxon distribution, B_i (*Lutzomyia* or *L. rufus*), and a subset of potential niche variables $I' \subseteq I$. We are interested in the probability $P(B_i|\mathbf{I}') = N_{B_i \text{ AND } I'}/N_{I'}$, where $N_{B_i \text{ AND } \mathbf{I}'}$ is the number of spatial cells where there is a co-occurrence of the taxon B_i and the niche variables $\mathbf{I}'$, and $N_{I'}$ is the number of cells where the niche variables take their stated values. The niche profile $\mathbf{I}'(x_\alpha)$ associated with a spatial cell x_α then determines the probability of the distribution variable, $B_i(x_\alpha)$, in that cell, and one now has a predictive model. The problem of calculating $P(B_i|\mathbf{I}')$ directly is that both $N_{B_i \text{ AND } I'}$ and $N_{I'}$ are likely to be zero when the number of taxa or niche variables considered simultaneously is large, as there will tend to be no co-occurrences of so many variables. This can be ameliorated by considering a reduced number of both class and feature variables ($\mathbf{I}_k$). For instance, $P(B_i|\mathbf{I}_k)$ is determined by the number of co-occurrences of the taxon B_i and the niche variable $\mathbf{I}_k$ and, in principle, allows us to find the most important statistical associations between the niche variables and the taxa distributions. However, $P(B_i|\mathbf{I}_k)$ being a probability does not account for sample size. For example, if $P(B_i|\mathbf{I}_k) = 1$, this may be as a result of there being a coincidence of B_i and $\mathbf{I}_k$ in one spatial cell or 1000. Obviously, the latter is more statistically significant. To remedy this we consider the following test statistic

$$\varepsilon(B_i|I_k) = \frac{N_{I_j}(P(B_i|\mathbf{I}_k) - P(B_i))}{(N_{I_j}P(B_i)(1 - P(B_i)))^{1/2}} \quad (1)$$

which measures the statistical dependence of B_i on $\mathbf{I}_k$ relative to the null hypothesis that the distribution of B_i is independent of $\mathbf{I}_k$ and randomly distributed over the grid, i.e., $P(B_i) = N_{B_i}/N$, where $N_{B_{i_j}}$ is the number of grid cells with point collections of species B_i and N is the total number of cells in the grid. The sampling distribution of the null hypothesis is a binomial distribution where, in this case, every cell is given a probability $P(B_i)$ of having a point collection of B_i. The numerator of Eq. (1) then, is the difference between the actual number of co-occurrences of B_i and $\mathbf{I}_k$ relative to the expected number if the distribution of point collections were obtained from a binomial with sampling probability $P(B_i)$. As we are talking about a stochastic sampling the numerator must be measured in appropriate "units". As the underlying null hypothesis is that of a binomial distribution, it is natural to measure the numerator in standard deviations of this distribution and that forms the denominator of Eq. (1). In general, the null hypothesis will always be associated with a binomial distribution as in each cell we are carrying out a Bernoulli trial ("coin flip"). However, the sampling probability can certainly change.

The quantitative values of $\varepsilon(B_i|\mathbf{I}_k)$ can be interpreted in the standard sense of hypothesis testing by considering the associated p-value as the probability that $|\varepsilon(B_i|\mathbf{I}_k)|$ is at least as large as the observed one and then comparing this p-value with a required significance level. In the case where $N_{B_j} \geq 5 - 10$ then a normal approximation for the binomial distribution should be adequate, in which case $\varepsilon(B_i|\mathbf{I}_k) = 2$ would represent the standard 95% confidence interval. When a normal approximation is not accurate then other approximations to the cumulative probability distribution of the binomial must be used.

In the case where $\mathbf{I}_k = B_k$, that is when we consider the effect of another taxon on the target species, then $P(B_i|B_k)$ and $\varepsilon(B_i|B_k)$ are measures of the statistical association between the two taxa, $\varepsilon(B_i|B_k)$ having the added advantage of having built into it the degree of statistical confidence that one may have about the association. Note that such a statistical association does not necessarily prove that there is a direct "causal" interaction between the two taxa. Rather, it allows for a statistical inference that may be validated subsequently. However, similarly, for abiotic variables it is just a statistical association and does not prove a causal link between the taxa distribution and the corresponding abiotic variable.

2.2.2. Constructing predictive models

Probabilities $P(B_i|\mathbf{I}')$, or proxies thereof, where $\mathbf{I}'$ is of high dimension, can be constructed using different classification models, such as neural networks, discriminant analysis, Maxent, etc. A particularly transparent, simple and effective approximation is the Naive Bayes approximation (Hand et al., 2001)

$$P(B_i|I) = \frac{P(I|B_i)P(B_i)}{P(I)} = \frac{\prod_{k=1}^{N} P(I_k|B_i)P(B_i)}{P(I)}$$

where in the first equality, Bayes rule has been used, and in the second it has been assumed that the niche variables $\mathbf{I}_k$ are independent. The product here is over the N niche variables under consideration as conditioning factors for B_i. In the case of the relationship between *Lutzomyias* and mammals, N represents the number of mammal species. Although the Naive Bayes approximation assumes independence of the variables $\mathbf{I}$, it is well known to be very often a robust approximation even when it is known that this assumption is not valid (Zhang, 2004).

A score function that can be used as a proxy for $P(B_i|\mathbf{I}')$ is

$$S(B_i|I') = \sum_{k=1}^{N} S(B_i|I_k) = \sum_{k=1}^{N} In\left(\frac{P(I_k|B_i)}{P(I_k|\bar{B}_i)}\right)$$

where $\bar{B}_i$ is the complement of the set B_i. For example, if B_i is the set of cells with presence of taxon B_i then $\bar{B}_i$ represents the set of cells without presence. $S(B_i|\mathbf{I}')$ is a measure of the probability to find the distribution variable B_i when the niche profile is $\mathbf{I}'$. It can be applied

to a spatial cell x_α by determining the niche profile of the cell, $\mathbf{I}'(x_\alpha)$. As an example, for two biotic niche variables, B_2 and B_3, that take values 1 (corresponding to the fact that there is a point collection associated with that cell) and 0 (there is no point collection associated with the cell), the four possible biotic niche profiles of any cell are $(B_2, B_3) = (0,0), (0,1), (1,0)$ and $(1,1)$. The score contributions of each biotic variable are $S(B_i|B_2)$ and $S(B_i|B_3)$, calculated using the above formula. Hence, $S(B_i|\mathrm{I}') = S(B_i|B_2,\ B_3) = S(B_i|B_2) + S(B_i|B_3)$. Thus, for any given spatial cell x_α one can assign a niche profile, i.e. values of B_2 and B_3, from whence it is possible to assign a corresponding score. If there is no statistical association between B_i and B_2 or B_3 then the corresponding score contributions are zero. An overall zero score then signifies that the probability to find B_i is the same as would be found if B_i were distributed randomly. If the score is positive then there is a higher than random probability to find B_i present and on the contrary if the score is negative. As each niche factor is treated separately in $\varepsilon(B_i|\mathbf{I}_k)$ or $S(B_i,\mathbf{I}_k)$ we can thus evaluate the relative contribution of any given niche factor and compare it to the contribution of any other.

2.3. Model analysis

The relation between the score function $S(B_i|\mathbf{I})$ and the niche variables $\mathbf{I}$ is determined as outlined above using a subset of randomly chosen grid cells – the training set, consisting of 70% of the data. $S(B_i|\mathbf{I})$ is a monotonic measure of the probability to find the taxa present and exhibits different characteristics as a function of geographic location, x_α as the niche profile $I(x_\alpha)$ is different for different locations. For instance, there may exist large/small sub-regions with very high/low scores. Model performance was tested on the remaining 30% of data, retained as an independent test set.

To examine model performance as a function of score and thereby, implicitly, as a function of location, we divided the grid cells into deciles. The 10th decile, as shown in tables below, corresponds to the 10% of grid cells with the highest score values, the 9th decile to the next 10% of grid cells with highest score values, etc. This allows us to establish predictability profiles across the different score deciles for the different niche models. A performance measure we will use here is to calculate for each score decile the percentage of associated target species collections. Summing over all the score deciles would yield 100% of the collections. The larger the percentage in the higher decile is, the better, more discriminating, the model. A random model would yield 10% of true positives in each decile whereas a perfect model would locate all point collections in the highest ranked cells. Thus, large changes in score passing from one decile to another correspond to the fact that the associated model discriminates well between one decile and another.

As the top/bottom score deciles correspond to best/worst niche conditions, by using ε we can determine for these deciles what are the most correlated niche factors, considering separate analyses for climatic, land cover and biotic factors. We can also examine to what extent these niche factors concur with previous knowledge about the ecological interactions of *Lutzomyia* and *L. rufus* or concentrate on those niche factors that have particular importance. For instance, for *Lutzomyia* there are several mammal species that are known reservoirs of the Leishmania parasite. We can determine to what extent these particular species are associated with the top/bottom deciles. Similarly, for *L. rufus* potentially interesting biotic factors are those that correspond to known preys or to potential competitors.

Having identified the most relevant factors associated with the most/least suitable niche variables for *Lutzomyia* and *L. rufus*, we can investigate how the contribution of these factors changes as we pass from one score decile to another. In order to do this a linear regression was performed, where the two variables involved are the log of the number of cells associated with a given score decile that are occupied by the taxa of interest and the log of the number of cells in the same decile that are occupied by the niche variables of interest.

3. Results

Before presenting the results it is important to state that their chief purpose is as case studies, to show what our modeling methodology is capable of. Thus, for instance, we may learn about the relative importance of abiotic versus biotic factors for two representatives of a pair of very different taxa – an insect and a mammal. The fact that we chose these particular species is, as far as the methodology is concerned, purely coincidental. Any taxa whatsoever could have been chosen with similar results – viz. – we may see how the methodology allows one to construct a more complete model for the realized niche of an organism, create more accurate distribution models, compare with what is already known about an organism and formulate new hypotheses and make new predictions.

3.1. Relative influence of biotic versus abiotic factors

We will first consider what can be gleaned about the relative effects of biotic versus abiotic factors in the niches of our representative taxa – an insect and a mammal. There are two different but related ways in which one can pose the analysis. One is to imagine that we know nothing about the biology of the different taxa being studied. We then use the model results to try and infer what their differences are by contrasting their relative niches. A goal of this might be to identify the taxon by studying its interaction with its niche. The other way is to make hypotheses about their comparative biology and ecology from existing knowledge and to see to what extent those hypotheses are validated. In this article, we will concentrate more on this second type of analysis. As a metric for the relative influence of biotic versus abiotic factors we consider three different score models – one using only abiotic variables (Worldclim) and two others, one using only mammals collections and the other land cover, as biotic factors. We rank our grid cells in terms of the three different scores and then divide the ranked cells into deciles. So, for example, decile 10 would be the 10% of cells with the highest scores for a given model while decile 1 would be the 10% of cells with the lowest scores.

Lutzomyia are a genus of small sandflies, various species of which are known vectors of the disease Leishmaniasis. The females need blood meals from mammals for the reproductive cycle, laying eggs in damp surroundings shortly after having fed. These mammals are then potential reservoirs for the Leishmania parasite. On the other hand, *Lynx rufus* is a vagile mammal that is a strict carnivore. What might one generally expect? First of all, *Lutzomyia*, as an insect, would be expected to be less robust to climatic factors than a mammal, such as *Lynx rufus*. Secondly, one might at first thought expect an insect to be less influenced by the presence of mammals than a carnivore that feeds principally on other mammals. However, in the present case the insect is a haematophage that depends on mammals for blood meals. We might therefore expect the relative contributions of the mammal variables to be somewhat similar.

We consider the average score as a function of score decile, the interpretation being that positive scores correspond to variable combinations associated with the niche, while negative scores correspond to an "anti-niche." We will use two representations of average score: one where the score is simply the sum of the scores of the corresponding variables in each cell, and a second, where we divide the score by the number of variables in the corresponding class – 190 for abiotic, 427 for mammal and 77 for land cover.

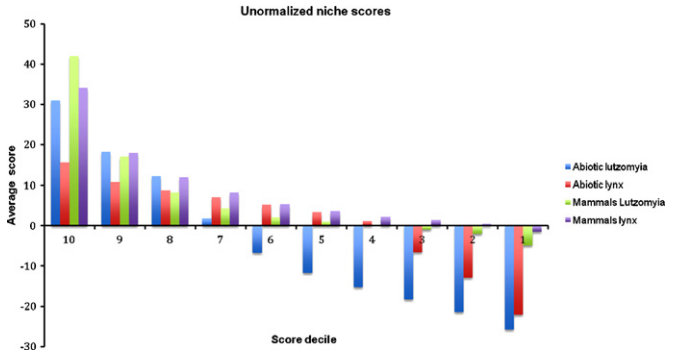

Fig. 1. Relative importance of different niche factor types for *Lutzomyia* and *Lynx rufus*.

In Fig. 1 we see the average score for each model as a function of each decile of grid cells. The relative variation in score across the deciles, comparing both abiotic and biotic variables for our two taxa, is very revealing. First of all, note the great deal of similarity for both *Lutzomyia* and *L. rufus* in the score distributions for the mammal variables. This might seem somewhat counterintuitive but is perfectly consistent with the hypothesis that biotic interactions with mammals should be similar for the two taxa given the nature of their feeding habits, i.e., both require mammals on which to feed; *Lutzomyia* are haematophages that are not known to be selective in which mammals they feed on. In this sense, the presence of mammal species is a necessary condition for the presence of *Lutzomyia*. Of course, one can argue that this is only an inference about potential biotic interactions. We will see that this inference can be made quite compelling shortly. Turning now to abiotic factors: We see that climatic factors, such as temperature and precipitation, are much more important for *Lutzomyia* than for *Lynx rufus*, as one might expect for an insect versus a mammal. The distribution of abiotic (climatic) scores shows that the role of climate is predominantly negative, indicating more where *Lutzomyia* cannot be, when compared to biotic variables, while the effect of land cover is quite similar to that of the climatic factors (Fig. 2).

In Fig. 2, we see the relative contributions per variable of the three variable types, where we have normalized the total score contribution by the number of corresponding variables. Once again we can see that climatic variables are much more discriminating for the insect than the mammal. In the case of mammals as biotic factors, we can see that their presence can be a good indicator of where to find either of our two taxa but do not influence very much where not to find them. Thus, in terms of anti-niche, abiotic factors are much more important than biotic ones.

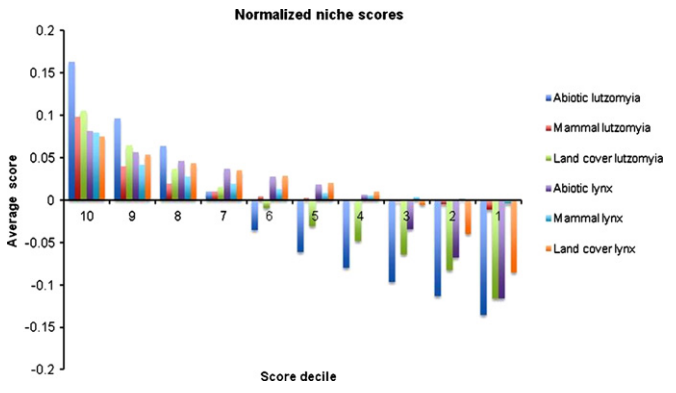

Fig. 2. Relative average score contribution of niche factors of different types for *Lutzomyia* and *Lynx rufus*.

3.2. Predictive power

We now move to a discussion of the predictive power of the different model types using Fig. 3, which shows the percentage of true positives in the different score deciles for the three different models – abiotic, biotic and land cover – and also for a model which contains all three classes of variables. From comparing the performance of the abiotic and biotic models in the top decile we see that the biotic model is more predictive than the abiotic or land cover models for both taxa. For *Lutzomyia* about 45% of point collections are found in the top decile for the biotic model compared to around 35% for the others. This means that biotic variables are more important than abiotic ones in determining the optimal niche regions associated with the highest probability to find *Lutzomyia*. This is even more the case for *Lynx rufus* where the biotic model leads to twice as high a percentage of true positives in the top score decile when compared to the abiotic one, Note that the abiotic model in this case leads to a substantial number of false negatives in the lower deciles. This is further evidence that climatic variables do not discriminate between niche and anti-niche as well as the biotic variables.

The most significant feature of Fig. 3 is the enhanced performance of a model that incorporates all three variable types as opposed to just one, leading not only to more true positives in the top decile but also to less false negatives in the lower deciles. The reason for this enhanced performance is that the different niche dimensions are complementary, biotic variables capturing predictability that is not present in the abiotic ones and vice versa. For *Lutzomyia* the model with all variables predicts more than 50% of the *Lutzomyia* point collections in the top decile, i.e., a 500% increase over random chance, while for *Lynx rufus* it is over 60%. To check the statistical significance of these results we randomly repeated the division of data into training and test sets 30 times (Fig. 4). We then considered the average performance of each model – abiotic, biotic, vegetation, all – in the top decile over the 30 trials and then considered the following statistical diagnostic for any model type

$$\varepsilon' = \frac{\bar{X}_b - \bar{X}_a}{\sqrt{(\sigma_b^2/n_b) + (\sigma_a^2/n_a)}}$$

For *Lutzomyia* the niche model with mammals is more predictive than abiotic and vegetation only. Once again, however, combining all variable types leads to a more predictive model, showing how a niche model built with abiotic and biotic variables together is more accurate and robust. In this case $\varepsilon' = 3.86$ with an associated p value of 0.007, thus showing that including in the full set of niche dimensions at our disposal leads to better prediction models. For *Lynx rufus* we find a value of 23.96 for the average performance of the biotic model over the abiotic model and 26.64 for the model with all niche variables relative to the abiotic model. Finally, ε' for the all variables model versus the biotic model is 4.78, corresponding to p value of 0.0019.

3.3. Constructing niches

An important advantage of our modeling framework is that it allows us to establish a transparent and unique ecological profile for each individual cell of our grid, thereby allowing us to quantify the relative influence of each niche variable at each geographic location and its impact on the presence of a species there. As the top decile of scores best characterizes the ideal niche, while the bottom decile is associated with anti-niche, we constructed lists of the most relevant abiotic and biotic variables for these deciles and ranked them from highest to lowest in terms of ε value. From these lists we selected the highest quartile of ε values for the top decile of scores and the lowest quartile (most negative) of epsilon

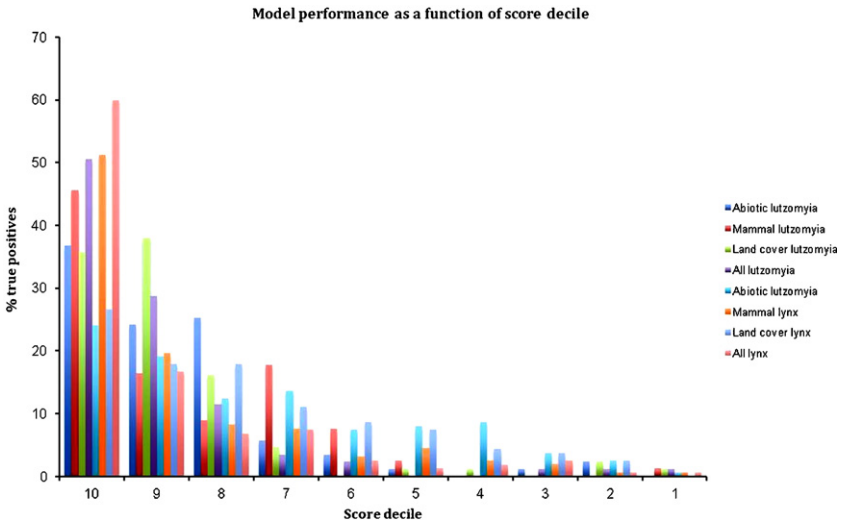

Fig. 3. Relative performances of different niche models for *Lutzomyia* and *Lynx rufus*.

values for the bottom decile (Tables 2 and 3). In this way, we can determine what the most important variables are in determining both the presence and absence of the target species. With these niche/anti-niche characterizing factors in hand we can then quantify how their presence changes as we pass from one score decile to another (Table 4). The results can be seen in Figs. 5 and 6. For instance, in the top left hand graph the horizontal axis is the log of the percentage of cells that are associated with the abiotic niche-defining factors given in Table 2 that are associated with the optimal *Lutzomyia* niche (top score decile), while the vertical axis corresponds to the log of the percentage of cells that are occupied by *Lutzomyia*.

For *Lutzomyia* in the top score decile, the optimum conditions are related to climate, emphasizing the role of precipitation as a key

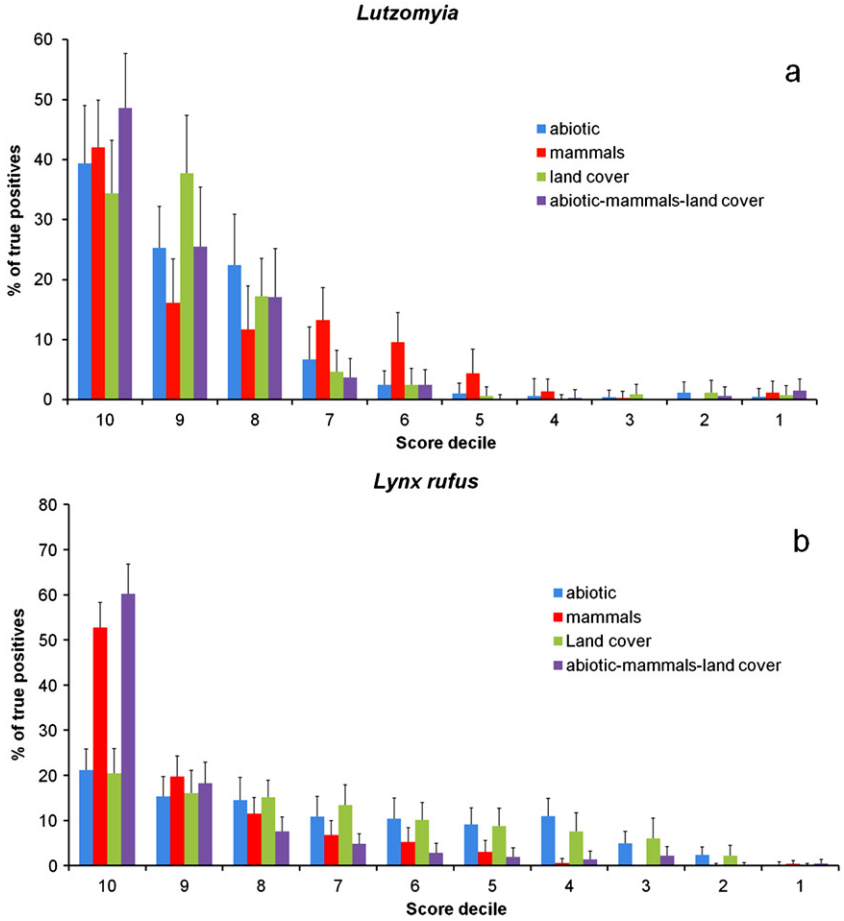

Fig. 4. Average number of collection points found for each model type in each score decile using 30 random samples of each set of variables for (a) *Lutzomyia* and (b) *Lynx rufus*.

Table 2
Abiotic and biotic variables in the top and bottom deciles for *Lutzomyia* with epsilon and score values.

Top decile				Bottom decile			
Optimal niche conditions for *Lutzomyia*				Suboptimal niche conditions for *Lutzomyia*			
Abiotic variables	Range	Epsilon	Score contribution	Abiotic variables	Range	Epsilon	Score contribution
BIO17	88–219	*8.960*	*5.013*	BIO12	42–507	*−5.604*	*−2.279*
BIO1	23.3–26.4	*8.938*	*1.006*	BIO16	18–218	*−5.001*	*−2.328*
BIO11	22.2–25.3	*8.873*	*2.322*	BIO18	1–249	*−3.839*	*−3.799*
BIO14	26–63	*8.782*	*4.916*	BIO6	3.1–3.4	*−3.761*	*−2.931*
BIO4	25.35–33.09	*7.543*	*2.152*	BIO7	26.3–28.4	*−3.544*	*−8.853*
BIO6	13.4–16.6	*7.524*	*3.293*	BIO2	16.5–18.4	*−3.535*	*−2.997*
BIO13	392–774	*7.107*	*12.913*	BIO11	2.9–12.5	*−3.271*	*−4.482*
BIO7	28.5–30.6	*7.012*	*3.803*	BIO4	3310–7184	*−2.971*	*−9.551*
BIO16	1019–2019	*6.925*	*12.175*	BIO19	192–383	*−2.940*	*−0.448*
BIO19	192–383	*6.618*	*4.157*	BIO10	28.9–32.3	*−2.669*	*−0.916*
BIO12	1906–3302	*6.314*	*8.701*	BIO1	10.3–19.9	*−2.189*	*−1.033*
BIO2	9.8–10.8	*6.130*	*4.458*	BIO3	3.7–5.5	*−2.130*	*−3.576*
BIO18	623–746	*5.748*	*1.260*	BIO8	28.4–31.7	*−1.964*	*−0.731*

Reservoirs			Reservoirs		
Reithrodontomys gracilis	8.892	2.640	*Sigmodon hispidus*	*6.946*	*1.244*
Heteromys gaumeri	8.800	2.234			
Heteromys desmarestianus	8.716	2.381			
Ototylomys phyllotis	7.559	2.028			
Peromyscus yucatanicus	7.249	2.116			
Sigmodon hispidus	6.946	1.244			
Didelphis marsupialis	5.774	1.662			
Oryzomys melanotis	3.494	1.387			
Marmosa mexicana	2.773	1.541			

Land cover			Land cover		
Cloud forest	6.642	1.408	Subtropical scrub	*−1.675*	*−1.527*
Tropical evergreen forest	6.603	4.476	Subtropical scrub with secondary vegetation	*−1.849*	*−1.658*
Cloud forest with secondary vegetation	*6.028*	*1.459*	Xeric scrub with secondary vegetation	*−2.092*	*−3.640*
Tropical evergreen forest with secondary vegetation	*6.007*	*4.344*	Xeric scrub	*−2.924*	*−4.044*
Agriculture areas	*5.966*	*1.736*	Mesquite	*−3.337*	*−1.714*
Human settlement	*4.947*	*0.577*	Grassland	*−3.734*	*−1.874*
Deciduous tropical forest with secondary vegetation	*4.081*	*1.013*	Mangroves	*−4.063*	*−2.000*

variable for its distribution. These results show annual and seasonal precipitation are between 900 and 3000 mm, and with temperature from 20 to 30 °C, indicating that the probability to find this species is higher from humid tropical to sub-humid climates ($R^2 = 0.554$, $p = 0.014$). Additionally, the most favorable vegetation types for *Lutzomyia* are deciduous forests and tropical evergreen, followed by agricultural areas ($R^2 = 0.809$, $p < 0.001$).

Nine mammal species have been confirmed as reservoirs of this blood-sucking insect vector (Canto-Lara et al., 1999; Van Wynsberghe et al., 2000; Sosa, 2004), there being a commensalism interaction between *Lutzomyia* and the mammals described. All these nine reservoir species are found in the top decile (Table 2), indicating the potential importance of their presence as a food resource for the vectors ($R^2 = 0.736$, $p = 0.002$). On the contrary, the variables that would indicate suboptimal conditions in the bottom decile for the presence of *Lutzomyia*, where the probability of finding *Lutzomyia* is smaller, are: low rainfall, hot and dry climates ($R^2 = 0.763$, $p < 0.001$), temperate forests and semi-desert scrub ($R^2 = 0.902$, $p < 0.001$). Only one known reservoir (*Sigmodon hispidus*) appears in the bottom decile and this is a mammal with a particularly wide distribution in Mexico.

For *L. rufus* we determined its niche landscape and projected it into our geographic region of interest. In the range of probabilities we can observe suitable areas in southern México where, in fact, this species has not been registered (Fig. 8). To identify the most relevant variables for *L. rufus*, we characterized its ecological profile and analyzed it with a simple log-linear regression (Fig. 6) as was the case with *Lutzomyia*. For the top score decile the most relevant niche variables were taken to be those associated with the highest quartile of ε values (Table 3). Temperature is the most important variable in places with annual and seasonal averages between −2 °C and 20 °C and its rainfall threshold is from 80 to 160 mm average. These results indicate a high probability to find this species in temperate to warm–dry zones ($R^2 = 0.505$, $p = 0.021$). Moreover, the most optimal vegetation types are temperate forests of pine, oak and fir, as well as semi-desert scrub, grasslands and crop fields ($R^2 = 0.417$, $p = 0.044$). The presence of *L. rufus* in agricultural areas could be associated with the presence of domestic animals (Lariviere and Walton, 1997).

Considering the potential importance of predator-prey interactions between *L. rufus* and other mammals we listed the 27 previously reported species that have been confirmed as prey to *L. rufus*, lagomorphs being the most important component in its diet (Lariviere and Walton, 1997; Aranda et al., 2002). Of these 27, 25 are in the top decile (Table 3), thus indicating that the wide availability of food resources and the change in their availability is an important factor in determining the range of *L. rufus* ($R^2 = 0.735$, $p = 0.002$). Projecting these ecological requirements onto a map we identified the most important areas in central and north Mexico for presence of *L. rufus* (Fig. 8).

Table 3
Abiotic and biotic variables in the top and bottom deciles for *L. rufus* with epsilon and score values.

Top decile				Bottom decile			
Optimal niche conditions for *L. rufus*				Suboptimal niche conditions for *L. rufus*			
Abiotic variables	Range	Epsilon	Score contribution	Abiotic variables	Range	Epsilon	Score contribution
BIO1	−2.7–16.7	*5.488*	*6.109*	BIO9	19.8–29.7	*−4.177*	*−0.821*
BIO6	−9.4–3.4	*5.327*	*3.005*	BIO11	19–28.6	*−3.930*	*−5.379*
BIO8	2.2–14.7	*4.797*	*1.096*	BIO6	6.8–19.9	*−3.578*	*−1.902*
BIO4	25.35–48.95	*4.704*	*1.393*	BIO1	23.3–29.7	*−3.452*	*−3.128*
BIO9	−3.5–16.4	*4.687*	*5.758*	BIO16	619–1618	*−3.060*	*−3.268*
BIO11	−3.6–16.5	*4.632*	*7.050*	BIO7	11.5–21.4	*−2.853*	*−1.656*
BIO16	219–418	*4.602*	*0.524*	BIO17	88–219	*−2.782*	*−1.091*
BIO5	7.7–30.5	*4.330*	*1.777*	BIO2	7.3–11.9	*−2.594*	*−0.954*
BIO10	−2.7–22	*4.266*	*2.33*	BIO13	238–620	*−2.59*	*−3.996*
				BIO12	974–3302	*−2.512*	*−1.413*
				BIO14	26–63	*−2.253*	*−4.666*
				BIO18	374–870	*−2.219*	*−1.068*
Preys				Preys			
Spermophilus variegatus		*13.824*	*1.883*	*Sylvilagus floridanus*		11.004	1.439
Sylvilagus floridanus		*11.004*	*1.439*	*Neotoma mexicana*		8.034	1.378
Neotoma albigula		*9.143*	*1.604*	*Didelphis virginiana*		5.553	1.054
Microtus mexicanus		*8.846*	*1.776*	*Nasua narica*		5.270	1.147
Dipodomys ordii		*8.636*	*1.565*	*Odocoileus virginianus*		4.457	1.589
Dipodomys merriami		*8.618*	*1.306*				
Neotoma mexicana		*8.034*	*1.378*				
Sigmodon leucotis		*6.275*	*1.982*				
Sylvilagus audubonii		*5.972*	*1.556*				
Didelphis virginiana		*5.553*	*1.054*				
Cratogeomys merriami		*5.385*	*2.031*				
Nasua narica		*5.270*	*1.147*				
Dipodomys deserti		*5.057*	*2.059*				
Dipodomys nelsoni		*4.972*	*1.453*				
Odocoileus virginianus		*4.457*	*1.589*				
Romerolagus diazi		*4.427*	*4.362*				
Dipodomys gravipes		*4.296*	*2.465*				
Dipodomys spectabilis		*4.039*	*1.366*				
Neotomodon alstoni		*3.860*	*1.589*				
Ammospermophilus harrisii		*3.700*	*2.128*				
Dipodomys agilis		*3.469*	*1.248*				
Spermophilus tereticaudus		*2.332*	*1.366*				
Dipodomys simulans		*1.875*	*1.877*				
Mustela frenata		*1.810*	*0.928*				
Sylvilagus cunicularius		*1.743*	*1.030*				
Potential competitors				Potential competitors			
Leopardus pardalis		*3.373*	*1.147*	*Leopardus pardalis*		3.373	1.147
Panthera onca		*2.559*	*0.928*	*Panthera onca*		2.559	0.928
Leopardus wiedii		*1.597*	*0.735*	*Leopardus wiedii*		1.597	0.735
Herpailurus yagouaroundi		*1.138*	*0.524*	*Herpailurus yagouaroundi*		1.138	0.524
Land cover				Land cover			
Grassland		*4.883*	*0.629*	Low forest evergreen with secondary vegetation		−2.088	−0.430
Plantation forest		*4.738*	*1.934*	Savannah		−2.202	−1.907
Xeric scrub with secondary vegetation		*4.283*	*1.094*	Cloud forest with secondary vegetation		−2.439	−2.061
Oyamel forest		*4.274*	*1.256*	Mangrove		−2.506	−1.191
High mountain meadow		*4.042*	*1.812*	Tropical evergreen forest with secondary vegetation		−2.540	−3.532
Agriculture areas		*3.903*	*0.745*	Tropical evergreen forest		−2.566	−3.575
Xeric scrub		*3.955*	*0.678*	Deciduous tropical forest		−2.924	−1.816
Coniferous forest		*3.878*	*0.565*	Deciduous tropical forest with secondary vegetation		−3.143	−2.471
Quercus forest		*3.858*	*0.475*				
Human settlement		*3.661*	*0.356*				
Coniferous forest with secondary vegetation		*3.631*	*0.591*				
Quercus forest with secondary vegetation		*3.457*	*0.468*				

Turning now to the bottom decile, we once again ranked the list of variables and took the lowest quartile of epsilon values corresponding to the most suboptimal conditions for *L. rufus* (Table 3). Prominent among these variables are annual and seasonal precipitation ranges between 900 and 3000 mm and temperature average from 23 °C to 30 °C. Thus, hot and humid climates are unfavorable for it ($R^2 = 0.508$, $P = 0.021$). Regarding vegetation, the probability of occurrence of this species in tropical forests is small ($R^2 = 0.711$, $P = 0.002$).

An important change in the optimal conditions for this species in the lower scoring deciles is the absence of a large number of typical prey species. In the bottom decile for example, only five confirmed prey species are present, only one of which is a lagomorph. It is important to highlight that we also found the presence

Table 4
Percentage of occupied cells across one score decile to another for *Lutzomyia* and *Lynx rufus*, and their optimal and suboptimal niche conditions described in Tables 2 and 3.

Decile	*Lutzomyia*	Optimal niche conditions			*L. rufus*	Optimal niche conditions		
		Reservoirs	Abiotic	Land Cover		Preys	Abiotic	Land Cover
10	14.2	66.0	100.0	98.4	31.4	93.9	99.4	98.4
9	8.1	29.8	100.0	97.4	8.7	67.6	99.4	99.0
8	3.2	11.4	99.7	90.6	3.5	45.1	99.0	98.1
7	1.0	8.4	94.2	71.8	3.9	31.3	97.7	95.1
6	0.6	6.2	71.8	68.5	1.3	21.2	97.7	91.9
5	0.0	3.6	39.9	55.8	0.7	14.7	96.7	83.4
4	0.0	5.2	32.0	58.9	1.0	11.4	84.1	65.3
3	0.3	5.2	19.4	51.8	1.3	30.5	77.9	83.4
2	0.3	6.5	9.4	37.9	0.3	22.1	49.7	80.5
1	0.3	7.7	3.5	38.9	0.3	20.6	54.8	84.8

Decile	Suboptimal niche conditions		Suboptimal niche conditions		
	Abiotic	Land Cover	Felines	Abiotic	Land Cover
10	76.1	6.8	5.8	30.1	18.4
9	61.8	7.4	4.9	26.9	23.6
8	80.8	16.6	1.0	25.1	25.1
7	95.1	35.1	2.0	29.6	25.1
6	99.4	65.9	0.7	29.0	28.3
5	100.0	77.3	2.0	36.8	41.4
4	100.0	80.3	3.2	73.7	52.3
3	100.0	76.7	14.9	94.2	89.0
2	100.0	85.8	10.4	100.0	99.7
1	100.0	89.7	11.9	100.0	100.0

of other species of felines (*Leopardus pardalis*, *L. wiedii*, *Herpailurus yagouaroundi* and *Panthera onca*) as potential competitors. However, when we analyzed the presence of other felines alone we noted these are not a variable that particularly restricts the presence of *L. rufus* ($R^2 = 0.0343$, $P = 0.608$) as they are present in all deciles. From this analysis, we consider that it is a combination of low availability of prey combined with presence of other felines as potential competitors that is limiting the distribution of *L. rufus* in some regions.

4. Discussion

Two of the most important objectives in ecology are to understand why a species can be in one place and not in another and which types of variable, either biotic or abiotic, are the most important in determining its presence or absence. However, a fundamental barrier that must be overcome in order that we can evaluate the relative importance of different niche variables is to establish a framework within which a fair statistical comparison can be made between them. The subtleties of doing this have meant that conventional ecological niche models have been based only on species–climate relationships (Guisan and Zimmermann, 2000; Guisan and Thuiller, 2005; Guisan et al., 2006). These can show the potential distribution for species, but they cannot say why this species is absent when the model says otherwise. This limitation is principally due to the difficulty of integrating different variable types (e.g., point collections and continuous environmental layers). However, it is generally accepted that integrating biotic variables into niche models would generate more robust and precise models for explaining species' distributions (Martin, 2001; Araujo and Luoto, 2007; Heikkinen et al., 2007; Wiens et al., 2009; Araújo et al., in press).

In this paper we proposed a novel methodology, based on data mining techniques, that integrates and analyzes both biotic and abiotic variables to thereby more fully characterize the ecological niche of a species and its associated geographical range. Our results represent an important next step in building distribution models. Integrating different types of variables in the construction of the niche allowed us not only to determine the probability of presence of the species but also to precisely determine the relative contributions of the different types of niche variables that contribute to these probabilities.

Comparing the predictability of the different niche models generated in our study, we noted that for both case studies, the predictability in the top decile is higher using mammals only than models with abiotic variables or vegetation only, but when we generate a model that integrates all variables the predictability increases even further. Therefore, a model that includes different variable classes is more predictive than one that does not and further, we can identify those variables that are most relevant in describing the suitable (niche) and unsuitable (anti-niche) areas. Hence, optimum conditions are given by those variables whose spatial expression is associated better with the presence of species, whereas suboptimal conditions are due to a loss of a relevant niche variable or the presence of an anti-niche variable.

Using two such contrasting examples as target species permitted us to identify the large differences in relative importance between biotic and abiotic factors that can occur from one species to another. For an insect like *Lutzomyia*, we saw clearly that abiotic variables are more negatively correlated, indicating that macroclimate plays an important role in determining the limits of its distribution. We found that precipitation is the most important variable and that when this variable changes it leads to a reduction in the probability to find this species. However, when we projected these ecological requirements on the geographic region of interest, we observed that there are point collections beyond these limits, principally in the north of México (Fig. 7).

If we did not use other variables types in order to build a more complete niche model, these points could be taken as being errors in the point collection data. However, including biotic factors allowed us to determine why *Lutzomyia* is predicted to be present beyond these climatic limits. Considering that *Lutzomyia* is a blood-sucking insect which has an interaction with mammals, we might surmise that presence of *Sigmodon hispidus* in that region could explain its occurrence. An important question to address is: How could it arrive at those sites? To answer this question it is necessary to

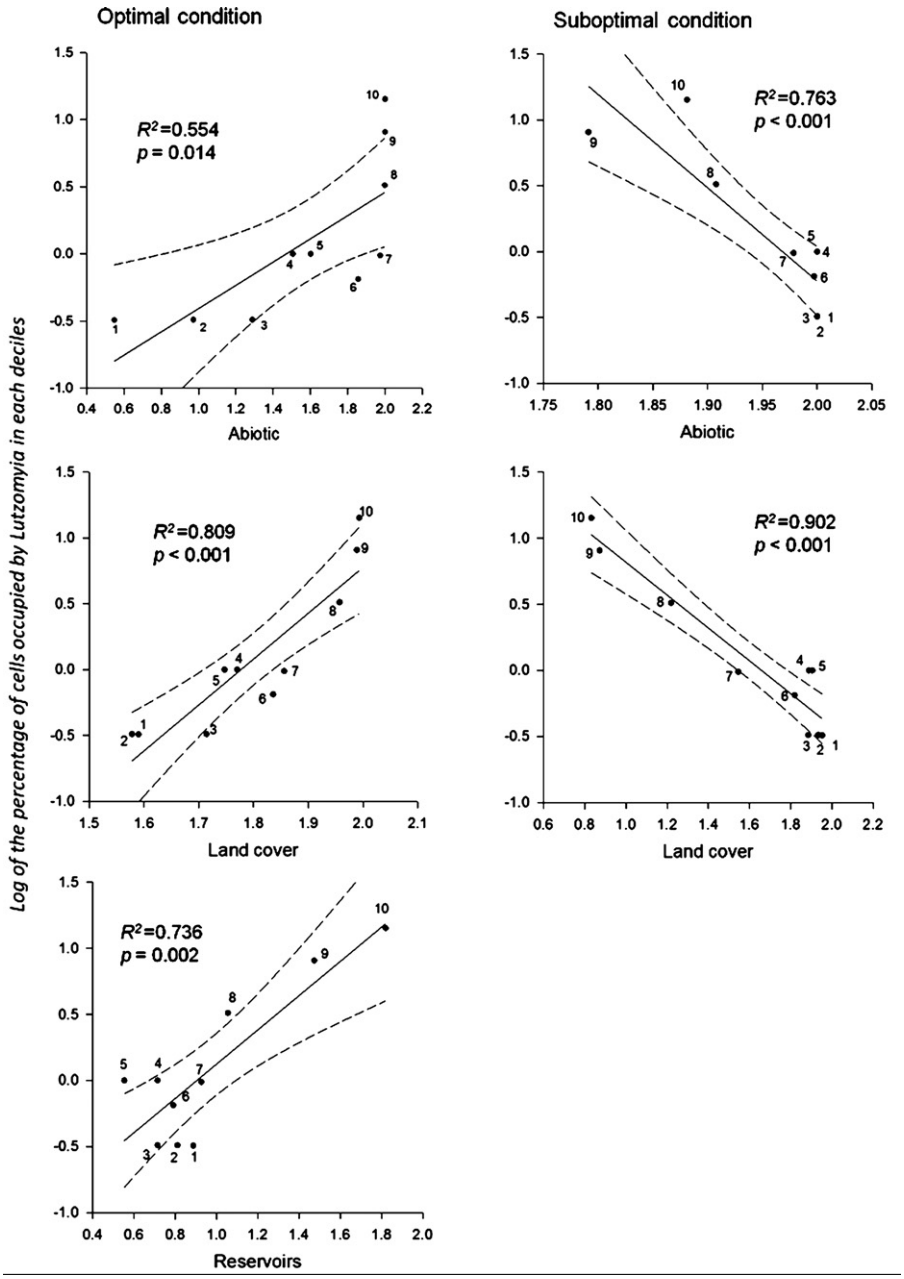

Fig. 5. Linear regression of the log of the percentage of cells occupied in each decile for optimal and suboptimal conditions versus log of the percentages of cells occupied for *Lutzomyia* as response variable. Within the chart we include the number of each of the deciles corresponding to Fig. 1.

analyze the inclusion of vegetation in the niche model. For this variable we found that tropical forests are an important factor for the presence of *Lutzomyia*. However, agricultural areas also play an important role in its distribution. If we consider the changes in primary vegetation to agricultural areas, these could be functioning as corridors for species associated with crops (González-Salazar and Stephens, 2012). This hypothesis may be sustained if we take into account the results of the log-linear regressions, where we saw that vegetation showed the highest correlation with the presence of *Lutzomyia*. If we consider the outcome of the model with all variables, we can establish a predictive scenario for the presence of this genus and identify areas of risk for leishmaniasis.

For *L. rufus* the relationship with abiotic variables was less negative than *Lutzomyia*. When we characterize its climatic requirements we saw that its distribution occurs in a heterogeneous climatic gradient from temperate to arid areas. We can then hypothesize that weather alone could not establish a limit to its distribution. An important question is then, why does it not occupy all the areas suitable for it? As climate is only one axis of its ecological niche, we need to analyze the other axes corresponding to biotic variables to explain this situation.

This case is, in fact, a good example to show the scope of our method because, as this species is a strict carnivore, there are really only two potential biotic interactions that could influence its distribution: predator–prey relationships and potential competition with other species of felines. Previously, Sánchez-Cordero et al. (2008) mentioned that in the south of the Isthmus of Tehuantepec, although exists areas with suitable fundamental niche conditions for this species, the presence of other felines may limit the distribution of *L. rufus* due to potential competition. However, when we characterize in our model the deciles in terms of mammals, we found that potential feline competitors are present in all deciles,

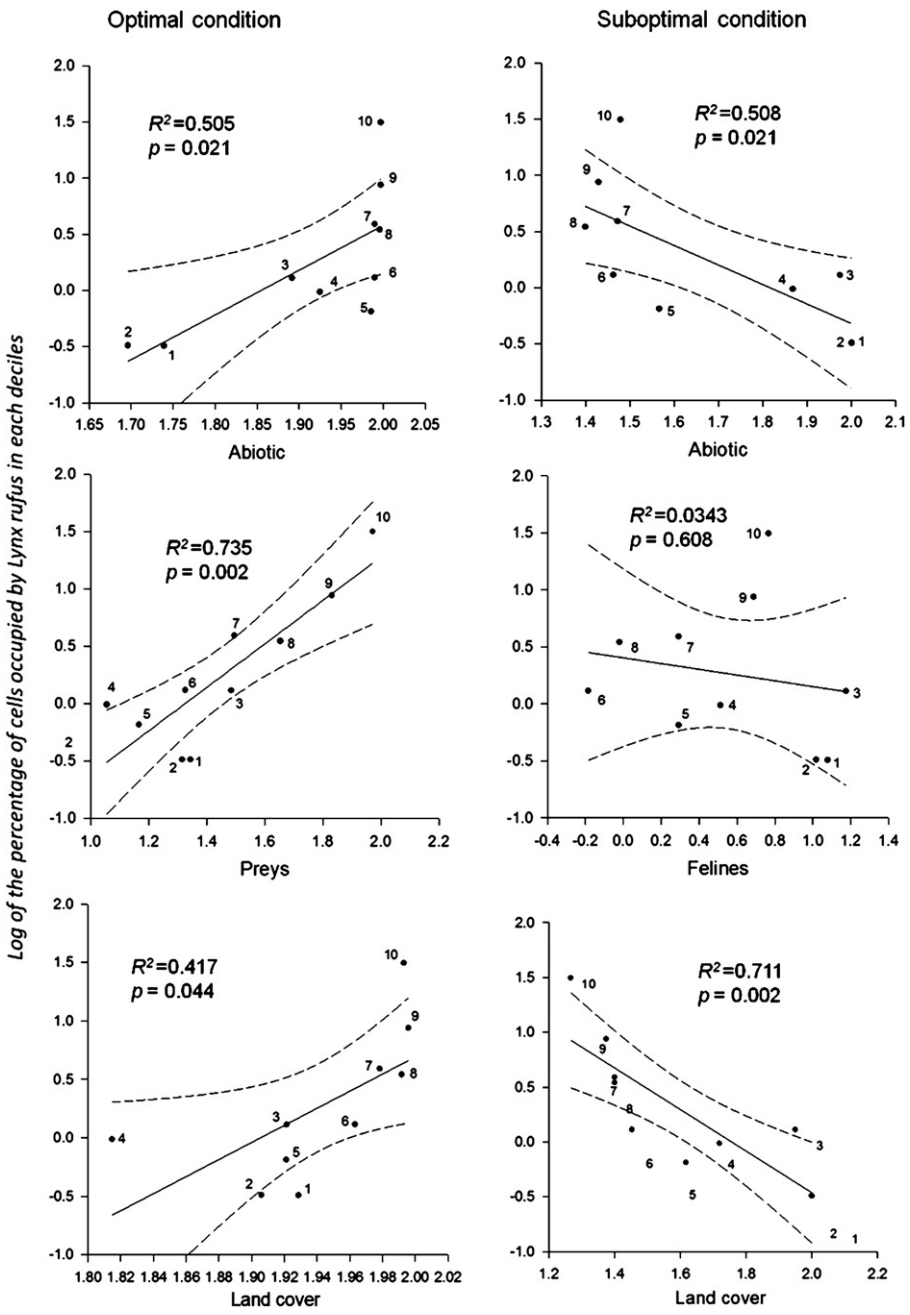

Fig. 6. Linear regression for log of percentage of cells occupied in each decile for optimal and suboptimal conditions versus log of percentages of cells occupied for *L. rufus* as response variable. Within the charts we include the number of each of the deciles corresponding to Fig. 1.

coexisting principally in the Pacific and the slopes of the Gulf of Mexico. Why then is there competition in the south and not in the north?

The answer to this question is probably availability of food resources for *L. rufus* so as to allow it to avoid inter-specific competition in the north. When we analyzed the bottom deciles corresponding to suboptimal conditions we noted an absence of prey species, principally lagomorphs. This would highlight the importance of preys as a determining factor for the distribution of a strict carnivore such as *L. rufus*. Thus, we attribute absence of *L. rufus* below the Isthmus of Tehuantepec primarily to the absence of prey species and secondarily to the presence of others felines, and changes in land cover.

Our results in both examples show the relative importance of biotic and abiotic components of the ecological niche to explain the distribution of species. Thus, although climatic factors might shape the boundaries of the species distribution, vegetation and biotic interactions allow us to identify the areas within these boundaries that are more suitable to occupy (Soberón, 2007). Nevertheless, constraints and the internal structure of the geographic ranges may vary through different parts of the distribution areas. An important step was characterizing the ecological profiles for our taxa. This permitted us to identify the most relevant niche variables, and which individual factors are the most important in characterizing both presence and absence areas in a gradient of probabilities. With this gradient we determine not just a limit to the range of a species but a continuous gradient of suitability/unsuitability. To visualize this in three dimensions we can think of niche (high presence probability regions) being associated with highlands and anti-niche (low presence probability regions) as lowlands. We call this view the

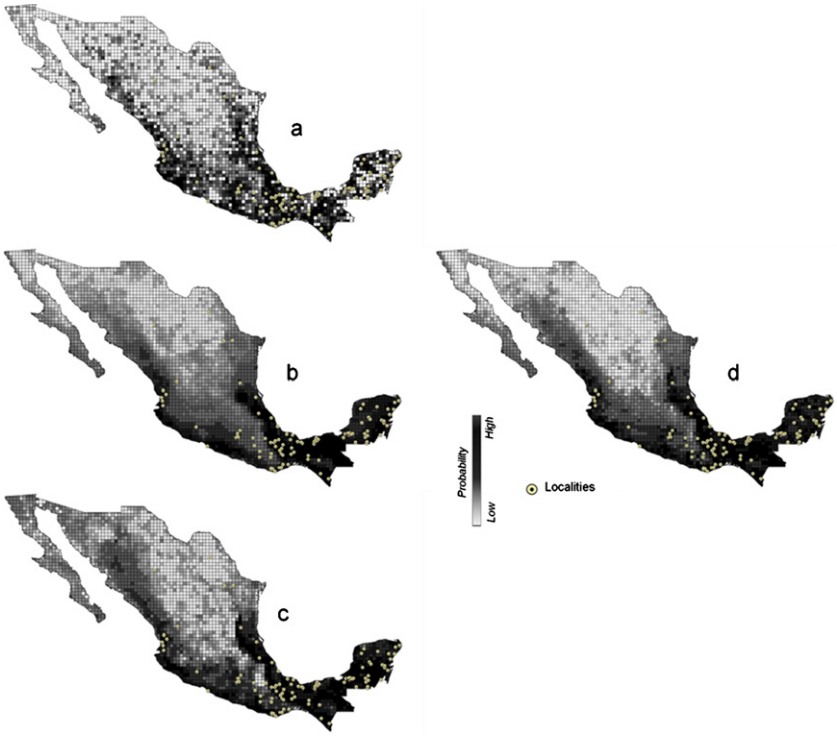

Fig. 7. Geographic projection of different niche models for *Lutzomyia*: (a) abiotic variables only; (b) mammals species only, (c) land cover only, and (d) abiotic variables, mammals and land cover combined.

"niche landscape" which allows us to infer what ecological processes occurred both inside and outside distribution areas.

Of course, one may argue that with respect to biotic interactions one can only infer their potential existence. This is true. All such modeling has this property. It is equally true of abiotic factors. One cannot more state with certainty that average annual temperature affects the distribution of, say, *Lynx rufus* than one can state that its distribution is affected by the presence of *Spermophilus variegates*. Both are inferences independent of which modeling algorithm or paradigm one chooses. In fact, we would argue that in many cases biotic interactions are much more directly responsible for the distribution of a species than abiotic factors. Is not it more sensible

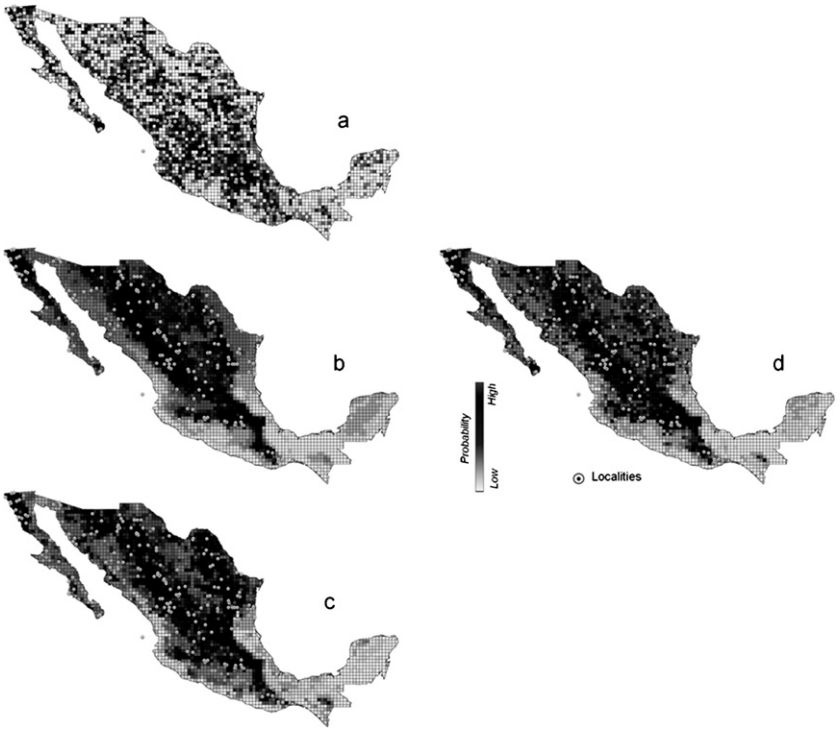

Fig. 8. Geographic projection of different niche models for *Lynx rufus*: (a) abiotic variables only; (b) mammals species only, (c) land cover only, and (d) abiotic variables, mammals and land cover combined.

biologically to infer that the presence or absence of a particular prey species of *Lynx rufus* is closer in the causal chain of factors affecting its distribution than a small change in average annual temperature? At any rate, it is preferable to establish a methodology within which such hypotheses may be formulated and later tested

5. Perspectives

A niche modeling methodology that allows us to include different types of variables, such as climate, vegetation, or biotic interactions, offers a fruitful framework within which to explain the ecological processes that occur from local to regional scales (Guisan et al., 2006; Heikkinen et al., 2007). Under the niche landscape viewpoint, we proposed a novel method with which to model the ecological niche. This representation can be usefully applied to diverse areas, such as patterns of biodiversity, emerging diseases, conservation, and global change as well as climatic change projections, as it allows us to determine how a particular variable can change and its relevance for species' distributions be understood. Additionally, this method could in principle be used to predict abundances (e.g. VanDerWal et al., 2009). We believe that optimal conditions in the top decile display where populations would be present in higher densities. This has potentially important implication in determining species priorities.

Acknowledgements

We thank an anonymous referee for valuable comments in regard to modeling niches beyond the species level, and for additional comments to improve this manuscript. CRS and CGS are grateful to CONACyT Grant 80156, support from CONACyT to the C3 – Centro de Ciencias de la Complejidad and to the Redes tematicas project Ciencia, Complejidad y Sociedad.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.ecolmodel.2012.10.007.

References

Alcocer, R., Stephens, C.R., 2012. Exploratory analysis of interrelations between co-located spatial Boolean features using network graphs. International Journal of Geographical Information Science 26, 441–468.
Aranda, M., Rosas, O., Ríos, J., García, N., 2002. Análisis comparartivo del la alimentación del gato montés (*Lynx rufus*) en dos diferentes ambientes de México. Acta Zoologica Mexicana 87, 99–109.
Araujo, M.B., Guisan, A., 2006. Five (or so) challenges for species distribution modelling. Journal of Biogeography 33, 1677–1688.
Araujo, M.B., Luoto, M., 2007. The importance of biotic interactions for modelling species distributions under climate change. Global Ecology and Biogeography 16, 743–753.
Araújo, M.B., Rozenfeld, A., Rahbek, C., Marquet, P.A. Using species coexistence networks to assess the impacts of climate change. Ecography, in press
Arntzen, J.W., Themudo, E., 2008. Enviromental parameters that determine species geographical range limits as a matter of time and space. Journal of Biogeography 35, 1177–1186.
Brown, J.H., 1995. Macroecology. University of Chicago Press, Chicago, IL, USA.
Brown, J.H., Stevens, G., Kaufman, D., 1996. The geographic range: size, shape, boundaries, and internal structure. Annual Review of Ecology, Evolution, and Systematics 27, 597–623.
Canto-Lara, S.B., Van Wynsberghe, N.R., Vargas-González, A., Ojeda-Farfán, F.F., Andrade-Narváez, F.J., 1999. Use of monoclonal antibodies for the identification of *Leishmania* spp. isolated from humans and wild rodents in the State of Campeche, Mexico. Memórias do Instituto Oswaldo Cruz 94, 305–309.
Guevara-Chumacero, L., López-Wilchis, R., Sánchez-Cordero, V., 2001. 105 años de investigación mastozoológica en México (1890–1995): una revisión de sus enfoques y tendencias. Acta Zoologica Mexicana 83, 35–72.
González-Salazar, C., Stephens, C., 2012. Constructing ecological networks: a tool to infer risk of transmission and dispersal of Leishmaniasis. Zoonoses and Public Health 59, 179–193.
Graham, C.H., Ferrier, S., Huettman, F., Moritz, C., Townsend Peterson, A., 2004. New developments in museum-based informatics and applications in biodiversity analysis. Trends in Ecology & Evolution 19, 497–503.
Guisan, A., Lehmann, A., Ferrier, S., Austin, M., Overton, J.M., Aspinall, R., Hastie, T., 2006. Making better biogeographically predictions of species' distributions. Journal of Applied Ecology 43, 386–392.
Guisan, A., Thuiller, W., 2005. Predicting species distributions: offering more than simple habitat models. Ecology Letters 8, 993–1009.
Guisan, A., Zimmermann, N.E., 2000. Predictive habitat distribution models in ecology. Ecological Modelling 135, 147–186.
Guralnick, R., Pearman, P., 2009. Using species occurrence databases to determine niche dynamics of montane and lowland species since the last glacial maximum. In: Körner, C., Spehn, E. (Eds.), Data Mining for Global Trends in Mountain Biodiversity. Taylor & Francis Group, Boca Raton, FL, pp. 125–134.
Hall, E.R., 1981. The Mammals of North America, vols. 1 and 2. Ronald Press, NY.
Hand, D., Mannila, H., Smyth, P., 2001. Principles of Data Mining. MIT Press, MA, USA.
Heikkinen, R.K., Luoto, M., Virkkala, R., Pearson, R.G., Körber, J.H., 2007. Biotic interactions improve prediction of boreal bird distributions at macro-scales. Global Ecology and Biogeography 16, 754–763.
Hijmans, R.J., Cameron, S.E., Parra, J.L., Jones, P.G., Jarvis, A., 2005. Very high resolution interpolated climate surfaces for global land areas. International Journal of Climatology 25, 1965–1978.
Lariviere, S., Walton, L.R., 1997. *Lynx rufus*. Mammalian Species No. 563. American Society of Mammalogists 563, 1–8.
Lomolino, M., Brown, J.H., Riddle, B., 2005. Biogeography. Signaler Sunderland, MA, USA.
Martin, T.E., 2001. Abiotic vs. biotic influences on habitat selection of coexisting species: climate change impacts? Ecology 82, 175–188.
Openshaw, S., 1983. The Modifiable Areal Unit Problem. Concepts and Techniques in Modern Geography. Geo Books, Norfolk, UK.
Palacio, J.L., Bocco, G., Velásquez, A., Mas, J.F., Takaki, F., Victoria, A., Luna, L., Gómez, G., López, J., Palma, M., Trejo, I., Peralta, A., Prado, J., Rodríguez, A., Mayorga, R., González, F., 2000. La condición actual de los recursos forestales en México: resultados del inventario forestal nacional 2000. Boletín del Instituto de Geografía 43, 183–203.
Pearson, R.G., Dawson, T.P., 2003. Predicting the impacts of climate change on the distribution of species: are bioclimate envelope models useful? Global Ecology and Biogeography 12, 361–371.
Ponder, W.F., Carter, G.A., Flemons, P., Chapman, R.R., 2001. Evaluation of museum collection data in biodiversity assessment. Conservation Biology 15, 648–657.
Sánchez-Cordero, V., Stockwell, D., Sarkar, S., Liu, H., Stephens, C., Gimenez, J., 2008. Competitive interactions between felid species may limit the southern distribution of bobcats *Lynx rufus*. Ecography 31, 757–764.
Soberón, J., 2007. Grinnellian and Eltonian niches and geographic distributions of species. Ecology Letters 10, 1115–1123.
Soberón, J., Peterson, A.T., 2004. Biodiversity informatics: managing and applying primary biodiversity data. Philosophical Transactions of the Royal Society of London Series B Biological Sciences 359, 689–698.
Sosa, E.I., 2004. Diagnóstico e identificación del subgénero de Leishmania en mamíferos silvestres mediante la técnica de reacción en cadena de la polimerasa (RCP). Universidad Autónoma de Yucatán, Facultad de Medicina, Unidad de Posgrado e Investigación, Mérida Yucatán, México.
Stephens, C., Giménez, J., González-Rosas, C., Ibarra-Cerdeña, C., Sánchez-Cordero, V., González-Salazar, C., 2009. Using biotic interaction networks for prediction in biodiversity and emerging diseases. PLoS One 4, e5725, http://dx.doi.org/10.1371/journal.pone.0005725.
Stockwell, D.R.B., Peters, D., 1999. The GARP modeling system: problems and solutions to automated spatial prediction. International Journal of Geographical Information Science 13, 143–158.
VanDerWal, J., Shoo, L.P., Johnson, C.N., Williams, S.E., 2009. Abundance and the environmental niche: environmental suitability estimated from niche models predicts the upper limit of local abundance. American Naturalist 174, 282–291.
Van Wynsberghe, N.R., Canto-Lara, S.B., Damián-Centeno, A.G., Itzá-Ortiz, M.F., Andrade-Narváez, F.J., 2000. Retention of *Leishmania* (Leishmania) *mexicana* in Naturally Infected Rodents from the State of Campeche, Mexico. Memorias do Instituto Oswaldo Cruz 95, 595–600.
Waltari, E., Guralnick, R.P., 2009. Ecological niche modelling of montane mammals in the Great Basin North America: examining past and present connectivity of species across basins and ranges. Journal of Biogeography 36, 148–161.
Wiens, J.A., Stralberg, D., Jongsomjit, D., Howell, C.A., Snyder, M.A., 2009. Niches, models, and climate change: assessing the assumptions and uncertainties. Proceedings of the National Academy of Sciences of the United States 106, 19729–19736.
Wolfe, N., Panosian, D.C., Diamond, J., 2007. Origins of major human infectious diseases. Nature 447, 79–283.
Zhang, H., 2004. Exploring conditions for the optimality of naive Bayes. International Journal of Pattern Recognition and Artificial Intelligence 19, 1830.