Estimating the fundamental niche: accounting for the uneven availability of existing climates

Jiménez, L.[1,*] and Soberón, J.[1]

[1]Biodiversity Institute, University of Kansas, 1345 Jayhawk Blvd, Lawrence, KS 66045, USA

[*]Corresponding author: laura_jimenez@ku.edu

January 26, 2021

Abstract

In the last years, studies that question important conceptual and methodological aspects in the field of ecological niche modeling (and species distribution modeling) have cast doubts on the validity of the existing methodologies. Particularly, it has been broadly discussed whether it is possible to estimate the fundamental niche of a species using presence data. Although it has being identified that the main limitation is that presence data come from the realized niche, which is a subset of the fundamental niche, most of the existing methods lack the ability to overcome it, and then, they fit objects that are more similar to the realized niche. To overcome this limitation, we propose to use the region that is accessible to the species (based on its dispersal abilities) to determine a sampling distribution in environmental space that allow us to quantify the likelihood of observing a particular environmental combination in a sample of presence points. We incorporate this sampling distribution into a multivariate normal model (Mahalanobis model) by creating a weight function that modifies the probabilities of observing an environmental combination in a sample of presences as a way to account for the uneven availability of environmental conditions. We show that the parameters of the modified, weighted-normal model can be approximated by a maximum likelihood estimation approach, and used to draw ellipsoids (confidence regions) that represent the shape of the fundamental niche of the species. We illustrate the application of our model with two worked examples: (i) using presence points of an invasive species and an accessible area that includes only its native range, to evaluate whether the fitted model

predicts confirmed establishments of the species outside its native range, and (ii) using presence data of closely related species with known accessible areas to exhibit how the different dispersal abilities of the species constraint a classic Mahalanobis model. Taking into account the distribution of environmental conditions that are accessible to the species indeed affected the estimation of the ellipsoids used to model their fundamental niches.

Keywords: *fundamental niche; realized niche; environmental space; presence data; weighted distribution; accessible area*

1 Introduction

In recent years, there has been substantial progress in the fields of ecological niche modeling (ENM) and species distribution modeling (SDM) (Guisan et al., 2013). However, there is still debate about what aspects of the niche are being estimated by these methods (Jiménez-valverde et al., 2008; Lobo, 2008; Warren, 2012). Specifically, conventional ENM/SDM approaches that are based on presence-only data estimate objects that are between the realized and the fundamental niches (Peterson, Soberón, Pearson, et al., 2011), in the case of ENM, or between the actual and potential distributions of the species (Jiménez-valverde et al., 2008), in the case of SDM.

The distinction between the fundamental and the realized niche, as proposed by Hutchinson (1957), is essential to understand what kind of objects are being estimated by the different correlative statistical models used in ENM/SDM. The fundamental niche of a species is the set of environmental conditions where, in the absence of biotic interactions, the population growth rate is positive (Peterson, Soberón, Pearson, et al., 2011). The realized niche is a subset of the fundamental niche that is determined by abiotic factors (environmental conditions), biotic factors, and dispersal limitations (Soberón, 2007). Estimating the fundamental niche of a species is of particular importance when using the estimated niche to model species distributions at other times or in different regions, such as when using ENM/SDM to predict the effects of climate change or the spread of invasive species (Tingley et al., 2014). However, estimating the fundamental niche of a species is also substantially more difficult thn estimating the realized niche, and it requires experimental data on the physiology of the species (Hoogenboom & Connolly, 2009; Jiménez et al., 2019).

The relationship between modeling niches and modeling geographic distributions is mediated by Hutchinson's duality (Colwell & Rangel, 2009), which is the relationship between geographic

and environmental spaces. With the right resolution, a discrete set of geographic coordinates can be made to have a one to one relationship with a discrete set of environmental vectors (Aspinall & Lees, 1994; Soberón & Nakamura, 2009). This fundamental correspondence allows us to move back and forth between modeling niches and modeling geographic distributions.

As a consequence of Hutchinson's duality (Colwell & Rangel, 2009), and because presence data come from the area currently occupied by a species, a sample of presence records may not reflect all the environmental potentiality of a species (Jiménez-valverde et al., 2008; Lobo, 2008) and estimating niches from geographic presence data probably approximates the realized niche (Soberón & Nakamura, 2009). This means that correlative models aiming to estimate the fundamental niche of a species from presence-only data will be constrained by the imposed limitations of the set of environments where the species can be observed (Owens, Ribeiro, et al., 2020). As a consequence, failing to acknowledge and (somehow) include this constraint into a model will lead to severe uncertainties and drawbacks when trying to predict the effect of climate change on the distribution of the species (Lobo, 2016), or possible invasion scenarios.

Presence data are often spatially biased and noisy. There are techniques to deal with some of the common types of problems (Chapman, 2005), such as lack of accuracy in the reported coordinates, nomenclatural and taxonomic errors, and presence of geographic or environmental outliers. We work under the assumption that a cleaning and preparation process precede the application of ENM/SDM methods. However, there are still other types of bias contained in presence data. Here, we will focus on the bias induced by defining the spatial region that is regarded as relevant for the study. Selecting different study regions produce different sampling universes, and part of the specification of a model is the definition of the sampling region. We shall use the idea of explicitly defining an "M" hypothesis to set the sampling universe (N. Barve et al., 2011). Under the BAM framework (Peterson, Soberón, Pearson, et al., 2011), the region M contains all the sites that the species is hypothesized to have been able to reach from some past time. Not all sites in M are adequate to sustain viable populations, and it is known that some sites inside of M could be sink populations (N. Barve et al., 2011).

In geography (G-space), M is usually a connected and continuous set (i.e. a single polygon) describing a region of space that a species can reach by movements (dispersal and migration). However, in order to perform practical computations, this set is first converted into a discrete grid of coordinates, and then their environmental values are used to build an environmental space (E-space, the space where the fundamental niche is defined and where it make sense to

estimate it). The Hutchinson's duality permits to establish a one-to-one relationship between the elements of these sets (coordinates in G-space and multidimensional vectors in E-space), but this relationship does not preserve distance: nearby elements in G-space (or E-space) not necessarily are also nearby in E-space (or G-space). This creates a serious, and mostly ignored problem: random sampling in G-space does not imply random sampling in E-space.

It is important to stress that most of the ENM/SDM applications implicitly assume that any (relevant) environmental combination could be observed in a sample of presences. In recent studies, however, it has being shown that acknowledging the effect of M on the sample can improve the performance of SDMs (Cooper & Soberón, 2018; Owens, Ribeiro, et al., 2020; Saupe, N. Barve, et al., 2017), but the emphasis is on the geographical part of the problem. The key lesson from these works is that M limits the presence of the species to a discrete set of multivariate environmental combinations with sampling probabilities that are not uniform. Therefore, if the goal is to estimate an object closer to the fundamental niche using presence-only data, the empirical distribution of M in E-space (i.e., the distribution of accessible environments for the species) should be used to inform the statistical model about the uneven distribution of available sampling points. In other words, the E-space is not sampled uniformly when using presence points because M imposes biases in E.

In a previous contribution, we proposed a Bayesian argument to combine correlative techniques with partial information from physiological experiments to obtain an approximation to the fundamental niche. However, we noticed the major problem of environmental combinations not being uniformly available in E-space. In this contribution, we consider the situation where the existing environmental combinations in E-space have different probabilities of being recorded as presences, and these probabilities are determined by M. Particularly, we consider the event of observing an environmental combination $\boldsymbol{x} \in \mathbb{R}^d$ (where the species was recorded as present), with a certain probability of being recorded or included in the sample. In a usual random sampling on the random variable $\boldsymbol{X}$ with probability density function (pdf) $f(\boldsymbol{x};\theta)$, the probability of selection of each environmental combination is the same, regardless of the value of $\boldsymbol{x}$, so that the pdf at the observation $\boldsymbol{x}$ is $f(\boldsymbol{x};\theta)$. However, in a biased sampling on $\boldsymbol{X}$, the probability of selection of an environmental combination is proportional to a predetermined weight function $w(\boldsymbol{x})$, implying that the pdf at the observation $\boldsymbol{x}$ is no longer $f(\boldsymbol{x};\theta)$. Here, we determine the form of the weight function, $w(\boldsymbol{x})$, and the resulting pdf under biased sampling.

We illustrate the application of the proposed statistical model with two worked examples.

In the first one, we estimate the fundamental niche of *Vespa mandarinia*, the Asian giant hornet whose alien presence has being recorded in Europe, and recently confirmed in the United States and Canada (Wilson et al., 2020). In this example, we use presence records and an M hypothesis that only includes the native range of this species to fit a convex shape (an ellipsoid) and get an estimate of the fundamental niche, then, we evaluate the fitted model using the presences recorded in the invaded regions. In the second worked example, we again use ellipsoids, presence records, in this case of different hummingbird species, and species-specific M hypothesis to exhibit how the different M scenarios constraint a classic Mahalanobis model. We identify scenarios under which we expect the Mahalanobis model to deviate from the fundamental niche and be closer to the realized niche of the species.

2 Materials and Methods

2.1 Modeling approach

Our aim is to take into account the structure of the environmental space when attempting to estimate a fundamental niche. In order to address this problem, we will follow Austin (2002) who suggested to include three major components when modeling in ecology: (1) an ecological model that describes the ecological assumptions to be incorporated into the analysis, and the ecological theory to be tested, (2) a statistical model that includes the statistical theory and methods used, and (3) a data model that takes into account how the data were collected or measured. The ecological model will be described in the following section and it includes a detailed definition of the fundamental niche as a function of fitness and its relationship with the environmental combinations in which a species has been observed. The statistical model and the data model will be addressed in subsequent sections.

Ecological model: relevant concepts in the study of the fundamental niche

The fundamental niche of a species, N_F, is the set of all environmental conditions that permit the species to exist (Hutchinson, 1957; Peterson, Soberón, Pearson, et al., 2011). Let E ($\subseteq \mathbb{R}^d$) be d-dimensional environmental space influencing fitness (measured, for example, as the finite rate of increase in a demographic response function). Furthermore, define a function, $\Lambda(\boldsymbol{x}) : E \longrightarrow \mathbb{R}$, that relates each environmental combination, $\boldsymbol{x} \in E$, to fitness (Jiménez et al., 2019; Pulliam, 2000). If the fitness function is of the right shape, there is a value of fitness,

λ_{min}, interpreted as minimum survivorship that defines the border of the fundamental niche (Etherington & Omondiagbe, 2019; Jiménez et al., 2019). This is, λ_{min} is the threshold above which the fitness is high enough as to support a population, and any environmental combination with a lower fitness level is outside the fundamental niche: $N_F = \{\boldsymbol{x} \in E | \Lambda(\boldsymbol{x}) \geq \lambda_{min}\}$. Notice that, using this notation, $\Lambda(\mu) = \lambda_{max}$.

Assumptions about the shape of the response of a species to an environmental variable (fitness function) are central to any predictive modeling effort (Austin, 2002; Soberón & Peterson, 2019). We assume that the level surfaces of the fitness function, Λ, and therefore, the frontier of the fundamental niche is convex (Drake, 2015). Specifically, we model these level surfaces through ellipsoids in multivariate space (Brown, 1984; Jiménez et al., 2019; Maguire, 1973) since these are simple and manageable convex sets that are defined through a vector μ that indicates the position of the optimal environmental conditions (the centre of the ellipsoid), and a covariance matrix Σ that defines the size and orientation of the ellipsoid in E. These parameters can also be interpreted as the parameters of a multivariate normal density function, $f(\boldsymbol{x}; \theta)$, where the corresponding random variable $\boldsymbol{X}$ represents an environmental combination in E where the species could be recorded as present (Jiménez et al., 2019). These ellipsoids are also know as Mahalanobis models (Farber & Kadmon, 2003) because the Mahalanobis distance defined by θ is equivalent to calculating the quadratic form that defines $f(\boldsymbol{x}; \theta)$. Ellipsoids have been used to test the niche-centre hypothesis (Osorio-Olvera et al., 2020). Under this framework, it is expected that the closer an environmental combination is to the mean μ, the higher the suitability value associated to that combination, and therefore, the higher the abundance of the species there.

We transform fitness values, $\Lambda(\boldsymbol{x})$, by simply calculating $f(\boldsymbol{x}; \theta)/f(\mu; \theta)$ which give us a value between 0 and 1 that can be interpreted as a "suitability" index for the environmental combination $\boldsymbol{x}$. The central assumption here is that there is a monotonic transformation between $\Lambda(\boldsymbol{x})$ and $f(\boldsymbol{x}; \theta)$ in the sense that high values of fitness (around λ_{max}) correspond to high suitability values (around μ). Therefore, we can work with the normal model to delimit the environmental requirements of a species and to test the niche-centre hypothesis.

Other central concepts in the study of the fundamental niche are the ones of the existing niche (Jackson & Overpeck, 2000) and the realized niche (Hutchinson, 1957), two subsets of the fundamental niche that result from considering the existing climatic conditions in the region of study, and, additionally, the biotic interactions with other species, respectively. We assume

that the following relationships between three niche concepts – the fundamental niche N_F, the existing niche $N^*(t; M)$ and the realized niche N_R – are fulfilled (Peterson & Soberón, 2012; Peterson, Soberón, Pearson, et al., 2011):

$$N_F \supseteq N_F \cap E(t) = N^*(t; M) \supseteq N_R(t; M). \tag{1}$$

These concepts are illustrated in Figure 1 and constitute our ecological model. In theory, every single environmental combination inside the ellipse is suitable for the species, and, if the species is able to reach one of these sites, it could persist there indefinitely (in the absence of biotic factors such as predators). However, at a given point in time, t, only a discrete subset of environmental combinations that can be mapped into E-space exist in the geographic space. This constitute the *existing niche* of the species. Notice that this set of points splits up into different regions when represented in geographic space; some of the points are in North America (purple points) and others are in South America (green points). Species have dispersal limitations that may prevent its individuals from colonizing all the environments in its existing niche. Thus, if a species is native to North America and is only able to reach the area covered in orange (henceforth called M, then, (i) its realized niche will be a subset of purple points (because they are suitable and the species has access to them), and (2) the set of green points constitutes its potential niche (suitable points that are not accessible to the species). Therefore, there is only a discrete set of environmental combinations from the fundamental niche that is available to the species and we expect to observe higher abundance of the species in those conditions that exist and are close to the center of the ellipse.

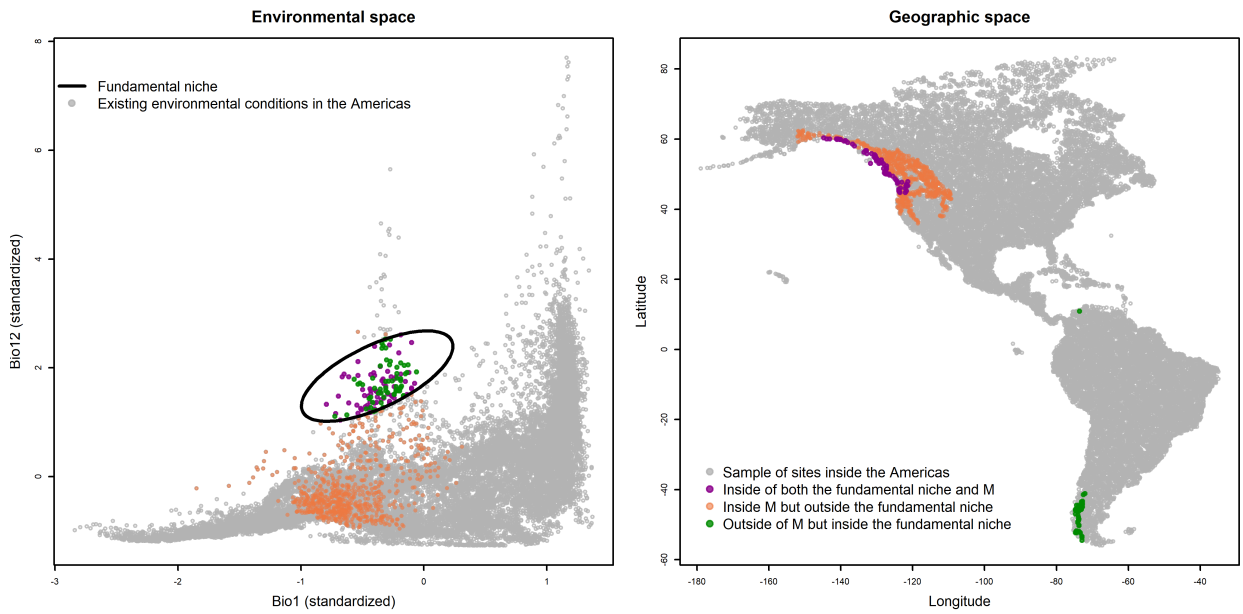


Figure 1: Different subsets of environmental combinations that are of interest for niche estimation. Each grey point in geographic space corresponds to a grey point in E-space, and vice versa. The ellipse represents the border of a fundamental niche where only the green and purple environmental combinations exits somewhere in geography (existing niche) and only the purple points are accessible to the species. The corresponding regions in geographic space are highlighted with the same colors.

Statistical and data models: two-stage sampling

Given a sample of environmental conditions where the species has been observed as present, $D = \{\boldsymbol{x}_1, ..., \boldsymbol{x}_n\}$, we propose a likelihood function for the parameters that describe the fundamental niche of the species ($\theta = (\mu, \Sigma)$). Suppose that the environmental space E is defined through d environmental variables (i.e., $D \subset E \subseteq \mathbb{R}^d$ and each point $\boldsymbol{x}_i$ has d coordinates). If E were a uniform grid of points embedded in $\mathbb{R}^d$, then the occurrence sample D could be considered as a random sample from a multivariate normal variable with density function

$$f(\boldsymbol{x}; \theta) = (2\pi)^{-d} |\Sigma|^{-1/2} \exp\left[-\frac{1}{2}(\boldsymbol{x} - \mu)^T \Sigma^{-1} (\boldsymbol{x} - \mu)\right], \tag{2}$$

which could be used to define a likelihood function and estimate θ.

Unfortunately, this is not the case. First, because the environmental combinations that actually exist in the planet do not represent a uniform sample from the whole multivariate space E because Huthinson's Duality does not preserve distances. If we take a uniform grid in geographic space and we map it into environmental space, the resulting cloud of points will be concentrated in some regions, leaving some other regions empty, as seen in Figure 1. The second

reason is that occurrences can only come from E_M – the set of environmental combinations associated to all the sites in the accessible region M. The irregular shape of E_M induces a sampling bias such that the probability of recording the species as present in an environmental combination is no longer given directly by $f(\boldsymbol{x};\theta)$, these probabilities will be affected by the availability of environmental conditions in E_M.

To account for the sampling bias induced by E_M, suppose that when the event $\{\boldsymbol{X} = \boldsymbol{x}\}$ occurs (meaning that the species was recorded as present at a site with this environmental conditions), the probability of recording it changes depending on the observed $\boldsymbol{x}$. We represent this probability by $w(\boldsymbol{x})$. Thus, the set of observed environmental combinations, D, can be considered as a random sample of the random variable $\boldsymbol{X}_w$ with probability density function

$$f_w(\boldsymbol{x};\theta) = \frac{w(\boldsymbol{x})f(\boldsymbol{x};\theta)}{E[w(\boldsymbol{X})]}, \tag{3}$$

where

$$E[w(\boldsymbol{X})] = \int w(\boldsymbol{x})f(\boldsymbol{x};\theta)dx.$$

Notice that $f_w(\cdot)$ is an example of a weighted density function where $E[w(\boldsymbol{X})]$ is the normalizing factor makes the total probability function equal to unity (Lele & Keim, 2006; Patil & Ord, 1976; Patil & Rao, 1978). Patil & Rao (1978) call this normalizing factor the *visibility factor* which captures the idea that samples from E are not uniform. The observed presences can only come from E_M (the visible set of environmental combinations), which may include regions in $\mathbb{R}^d$ where the environmental combinations are abundant (associated to high values of $w(\boldsymbol{x})$). But, even though the points inside these abundant regions might not be close to μ (i.e., where $f(\boldsymbol{x};\theta)$ is small), the probability of observing the species in these points could be higher than the one associated to other points closer to μ (i.e., where $f(\boldsymbol{x};\theta)$ is large) that do not exist in E_M (they are not visible), or whose weight $w(\boldsymbol{x})$ is too small.

On the other hand, equation 3 can also be described as the resulting model for a two-stage sampling design that accounts for the random process under which an environmental combination $\boldsymbol{x} \in D$ is observed (Patil & Rao, 1978). Suppose that nature produces a sample of size N of environmental conditions inside the fundamental niche with probabilities given by the density function $f(\boldsymbol{x};\theta)$ (this sample may contain any point in E), but, because the species not only requires the right abiotic conditions to maintain a population, – it also needs specific biotic conditions and it can only disperse to a finite set of sites – the recorded sample will not

include all the N observations. Instead, only a subsample of size $n < N$ is selected by drawing observations from the original sample (of size N) with a chance proportional to $w(\boldsymbol{x})$.

Let us illustrate the two-stage sampling design using the fundamental niche shown in Figure 1. Suppose that a species' fundamental niche is defined by the bivariate normal distribution with parameters:

$$\mu_1 = (-0.38, 1.85), \Sigma_1 = \begin{pmatrix} 0.0412 & 0.0366 \\ 0.0366 & 0.075 \end{pmatrix},$$

where the first dimension represent annual mean temperature (x-axis), and the second represents annual precipitation (y-axis). The level curves of the corresponding density function, $f(\boldsymbol{x}; \mu_1, \Sigma_1)$, are ellipses; they are plotted in the left panel of Figure 2 where the largest ellipse corresponds to a 99% confidence region. In the first stage of the sampling process, we generated a sample of size $N = 100$ from this distribution and plotted it on top of the ellipses. As expected, most of the environmental combinations in this sample are close to the center of the ellipses (green points inside the ellipses plotted in the left panel of Figure 2). The middle panel of Figure 2 shows all the environmental combinations that exist inside the region that is accessible to the species, E_M (purple and orange points, in both Fig. 1 and Fig. 2); we identified the subset of points that exist in E_M and are inside the 99% confidence region of $f(\boldsymbol{x}; \theta)$, and colored them in purple. Then, we estimated the density function of the accessible environments, $\hat{h}(\cdot; E_M)$, using a kernel method. The resulting level curves of $\hat{h}(\cdot; E_M)$ correspond to the orange regions in the middle panel of Fig. 2. We used $\hat{h}(\cdot; E_M)$ to define the weights, $w(\boldsymbol{x}) = \hat{h}(\boldsymbol{x}; E_M)$, which were used in the second stage of the sampling process to select a subsample of size $n = 25$ – assuming that not all the $N = 100$ points from the original sample exist in E_M – from the first sample of environments from the fundamental niche. The resulting sample, $D = \{\boldsymbol{x}_1, ..., \boldsymbol{x}_n\}$, is shown in the right panel of Figure 2 (purple triangles).

Notice that if we use a simple likelihood approach based on the simulated sample to estimate the parameters μ_1 and Σ_1 (this is equivalent to fitting a Mahalanobis model), we recover a 99% confidence ellipse that is smaller (violet ellipse in the right panel of Fig. 2) than the theoretical fundamental niche of our species (largest, blue ellipse in the right panel of Fig. 2). More importantly, the estimated center of this ellipse does not coincide with the optimal environmental conditions, μ_1. Therefore, if we estimate the parameters that describe the fundamental niche of the species without taking into account the distribution of available environments, we can not claim that the model fully recover the fundamental niche of the species. Moreover, if we try to

test the niche-center hypothesis using the ellipse recovered from a Mahalanobis model, we will be expecting high abundances around an environmental combination that is not the optimum.

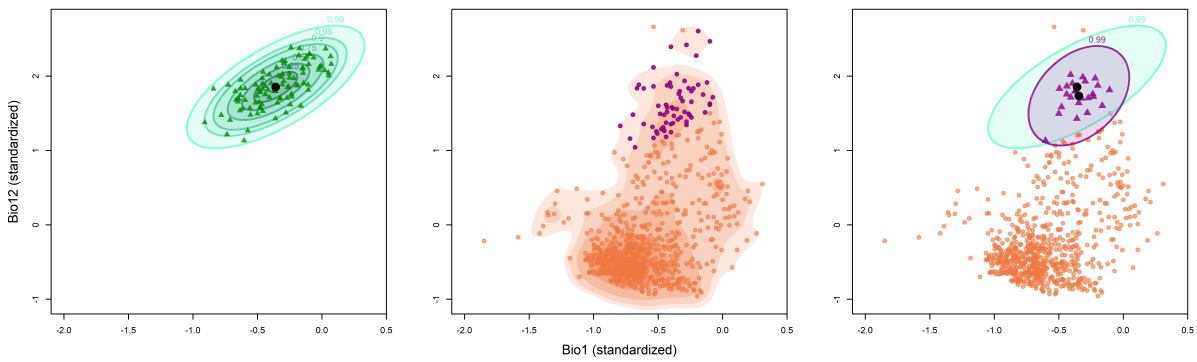


Figure 2: Left: theoretical fundamental niche represented as ellipsoids that correspond to the confidence regions of a normal distributions with parameters μ_1 and Σ_1 and a sample of size $N = 100$ (green triangles) simulated from this distribution (first stage of sampling process). Middle: environmental combinations accessible to the species, set E_M (orange and purple circles) and contour levels of the kernel density function estimated with these points, $\hat{h}(\boldsymbol{x}; E_M)$ (orange regions). Right: subsample of size $n = 25$ (purple triangles) selected in the second stage of the sampling process using the weights defined by $\hat{h}(\cdot; E_M)$, points in E_M (orange circles), and theoretical fundamental niche of the species (green ellipse).

Coming back to the application of the method, we will assume that the sample of environmental combinations where the species of interest was observed as present, D, is a random sample of the random variable $\boldsymbol{X}_w$ with probability density function given by Eq. 3. Thus, we can define the likelihood function of the parameters of interest $\theta = (\mu, \Sigma)$ as follows:

$$\mathcal{L}(\theta|D) \propto \prod_{i=1}^{n} f_w(\boldsymbol{x}; \theta) = \prod_{i=1}^{n} \frac{w(\boldsymbol{x}) f(\boldsymbol{x}; \theta)}{E[w(\boldsymbol{X})]}, \tag{4}$$

where the function $w(\boldsymbol{x})$ will be approximated using all the accessible environmental combinations and a kernel density procedure. As for the expected value $E[w(\boldsymbol{X})]$ (which was defined as integral; see Eq. 3), we will get a Monte Carlo estimate of this quantity. Notice that the function $w(\boldsymbol{x})$ does not depend of the parameters of interest, and it can be ignored when maximizing the log-likelihood function. Furthermore, the analytical form of $w(\boldsymbol{x})$ is unknown, hence the analytical form of $E[w(\boldsymbol{X})]$ is also unknown. However, we can sample environmental combinations from the availability distribution by randomly choosing points from M and extracting the values of the environmental variables at those sampled sites. Using this sample, we can get a Monte-Carlo estimate of $E[w(\boldsymbol{X})]$ for any fixed value of θ. This method has been used before in the statistical modeling literature and it is called method of simulated maximum likelihood

(Lele & Keim, 2006; Robert & Casella, 1999). Thus, we can obtain a Monte-Carlo estimate of the log-likelihood function as follows:

$$\ell(\theta|D) = \sum_{i=1}^{n} \log(f_w(\boldsymbol{x}_i;\theta)) = \sum_{i=1}^{n} \left\{ \log(f_w(\boldsymbol{x}_i;\theta)) - \log\left[\frac{1}{K}\sum_{j=1}^{K} f_w(\boldsymbol{x}_j^*;\theta)\right]\right\}, \tag{5}$$

where $\boldsymbol{x}_j^*, j = 1, 2, ..., K$ is a random sample with replacement from the distribution $w(\boldsymbol{x})$. The size of this sample, K, needs to be large enough as to ignore Monte-Carlo error. Once that sample is generated, we can apply standard optimization techniques to minimize the negative log-likelihood function in Eq. 5 and obtain the maximum likelihood estimators, $\hat{\mu}$ and $\hat{\Sigma}$, respectively.

In summary, we will use a sample of occurrences for the species of interest together with a polygon that represents its accessible area. Inside these polygons, points were generated at random and their environmental values extracted. This was done with two different purposes: (1) to estimate a kernel density to define the weights in the likelihood function, and (2) to get a Monte-Carlo estimate of the log-likelihood function. Once we have all the elements of the likelihood function, we will calculate the maximum likelihood estimates of the parameters, $\hat{\mu}$ and $\hat{\Sigma}$. These estimated parameters will allow us to plot ellipses in E-space to represent the border of the estimated fundamental niche. In all the examples, we will plot the ellipses that correspond to the 99% confidence regions of the fitted multivariate normal distribution. We will compare this ellipses with the ones that correspond to the 99% confidence region of a standard Mahalanobis model (Farber & Kadmon, 2003) with estimated parameters $\hat{\mu}_0 = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i$ and $\hat{\Sigma}_0 = \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{x}_i - \hat{\mu}_0)(\boldsymbol{x}_i - \hat{\mu}_0)^T$ (i.e., the maximum likelihood estimates under a multivariate normal model described in Eq. 2). We hypothesize that the two ellipses will be similar in cases where E_M covers most of the fundamental niche of the species (i.e., the overlap between these two sets is high) and the distribution of points in E_M is approximately uniform.

Finally, we will project the resulting models back to geographic space. In order to do this, once we have the maximum likelihood estimates of the parameters of interest, $\hat{\theta}$, we use the multivariate normal density function given in Eq. 2 to calculate a suitability index, which can be plotted in G space, either as a continuous value or a binary region, (using a threshold). For interpretation purposes, it is convenient to standardize this index to the interval $(0, 1)$ which easily done by dividing $f(\boldsymbol{x};\theta)$ by the its maximum value, $f(\boldsymbol{x};\hat{\theta})$.

2.2 Data

The application of our method to approximate fundamental niches requires three types of data: (*i*) occurrence data, which should go through a thorough cleaning process before being used, (*ii*) an M hypothesis or geographic polygon that encloses all the accessible sites, taking into consideration the dispersal abilities of the species and geographic barriers, (*iii*) environmental layers cropped by the area of study from which we can extract values for the occurrences and the sites inside M. Below, we describe the particular datasets that we used to create our worked examples.

Occurrence data

We selected seven species to illustrate the use of the statistical model that we are presenting here. The first species is *Vespa mandarinia*, the Asian giant hornet (Matsuura & Sakagami, 1973; Matsuura, 1988). Being an invasive species, *V. mandarinia* can provide insight about whether the estimated niche fitted with our model is a good approximation to a true fundamental niche. This is, we can use the known locations that this species has been able to invade and we expect that the estimated niche contains this sites. We also selected six species of Hummingbirds for our analyses: *Amazilia chionogaster*, *Threnetes ruckeri*, *Sephanoides sephanoides*, *Basilinna leucotis*, *Colibri thalassinus*, and *Calypte costae*. These are used because very detailed M hypotheses are availbale for them.

Originally, all the occurrence data comes from the Global Biodiversity Information Facility database (GBIF; `https://www.gbif.org/`). We downloaded 1944 occurrence records for *V. mandarinia* (GBIF, 2020), which, after undergoing a standard cleaning procedure (Cobos et al., 2018) were reduced to 170 presences in the native range and one presence record in Europe. The cleaned dataset was then spatially thinned by geographic distance (at least 50 km away) to avoid having an extra source of sampling bias (Anderson, 2012), ending up with a final sample of 46 presence records to fit the niche models (see red points in Fig. 3).

For the hummingbird species, we used the cleaned occurrence records that were used by Cooper & Soberón (2018), which are available at `https://github.com/jacobccooper/trochilidae`. Cooper & Soberón made a thorough revision and cleaning of the presence records, eliminating misidentified individuals and synonyms. Additionally, they defined the accessible regions for most of the extant hummingbird species and used them to fit species distribution models. Given the success that they had at using this M hypothesis, we consider that their data provide a great

opportunity to test our proposed model.

Figure 4 shows the presence samples of each hummingbird species. The sample sizes range from 148, presences for *Amazilia chionogaster*, to 926 presences, for *Calypte costae*. We selected this six species because they occupy different regions of the Americas, from Southern United States to the Patagonia. Therefore, it will be interesting to compare the estimated niches of all the species in E-space and see how similar they are; particularly, how distant are the optimal environmental conditions among different species. The different species are likely to occupy different regions of E-space but their fundamental niches might share some environmental combinations.

M hypothesis

In the case of *V. mandarinia*, we decided to define the accessible area as a combination of buffers that represent the species' dispersal ability and the elevation range where the species in known to occur (850 - 1900 m). First, we delimited a region contained within a buffer area of 500 km around all the occurrence records in the sample, which accounts for the dispersal abilities of the hornets (Matsuura & Sakagami, 1973). Second, we clipped this region with an elevation layer to get rid of regions at elevations higher than 1900 m. The resulting polygon is shown in Figure 3 (outlined in blue). Notice that the polygon falls outside the continents, however, we will not consider these regions in the study. When we extract the environmental values of the sites inside this M, we only do it for the inland sites.

For the hummingbird species, we used the polygons generated by Cooper & Soberón (2018). These areas were hypothesized taking into account the known occurrences, topography, ecoregions, and estimated dispersal distances, as well as bounding by significant geographical barriers such as large rivers and mountains. This is, they took into account all the criteria that are know to yield more accurate models (N. Barve et al., 2011; Owens, Campbell, et al., 2013; Owens, Ribeiro, et al., 2020; Saupe, V. Barve, et al., 2012). Figure 4 shows the six polygons for the different species. Notice that there are species like *T. ruckeri* that occupy most of its accessible area (Fig. 4b), while some others, like *B. leucotis* and *C. costae*, occupy only a fraction of the accessible area. These species also show diversity in their range sizes, *S. sephanoides* seems to have a more restricted range, compared to *C. thalassinus*. Therefore, it will be interesting to make similar comparisons in E-space.

Environmental variables

The climatic layers used to create the models for each species came from the WorldClim database (Hijmans et al., 2005). We use only two out of the 19 variables: annual mean temperature (Bio1), and annual precipitation (Bio12); both variables were recorded at a 10 arcmin resolution. We clipped each layer using the polygons that correspond to the M hypotheses for each species.

For all the species that we selected for the analyses, both of the variables that we selected are biologically meaningful (see Matsuura & Sakagami (1973) for the Asian hornet, and Root (1988) for Hummingbirds). Furthermore, in the case of the hummingbirds, we wanted to compare their estimated niches. Comparisons of estimated niches in E-spaces with different axes and/or dimensions are not possible, therefore, we only looked at these two dimensions of their niches.

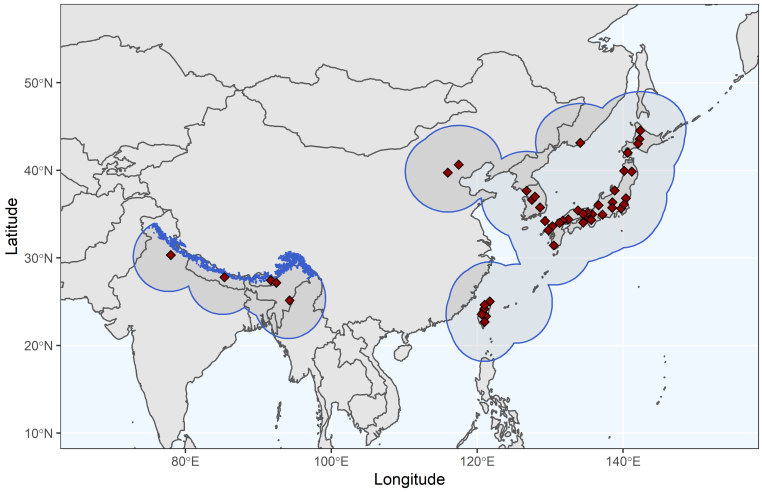


Figure 3: Occurrence records (red squares) of *V. mandarinia* in its native range, and the M hypothesis created from a combination of buffers and the known elevation range of the species (regions delineated with blue lines). The sample of occurrences went down to 46 records after the cleaning and thinning processes.

All the analysis were done in R version 3.6.3 (R Core Team, 2020). We used existing packages for different steps in the data preparation, analysis and visualization: ggplot2 (Wickham, 2016), ks (Duong, 2020), raster (Hijmans, 2020), rgdal (Bivand, Keitt, et al., 2020),rgeos (Bivand & Rundel, 2019), scales (Wickham & Seidel, 2020), sf (Pebesma, 2018). Additionally, we created functions that can be used to reproduce our examples, as well as to apply our methodology to other species. These functions can be consulted at `https://github.com/LauraJim`.

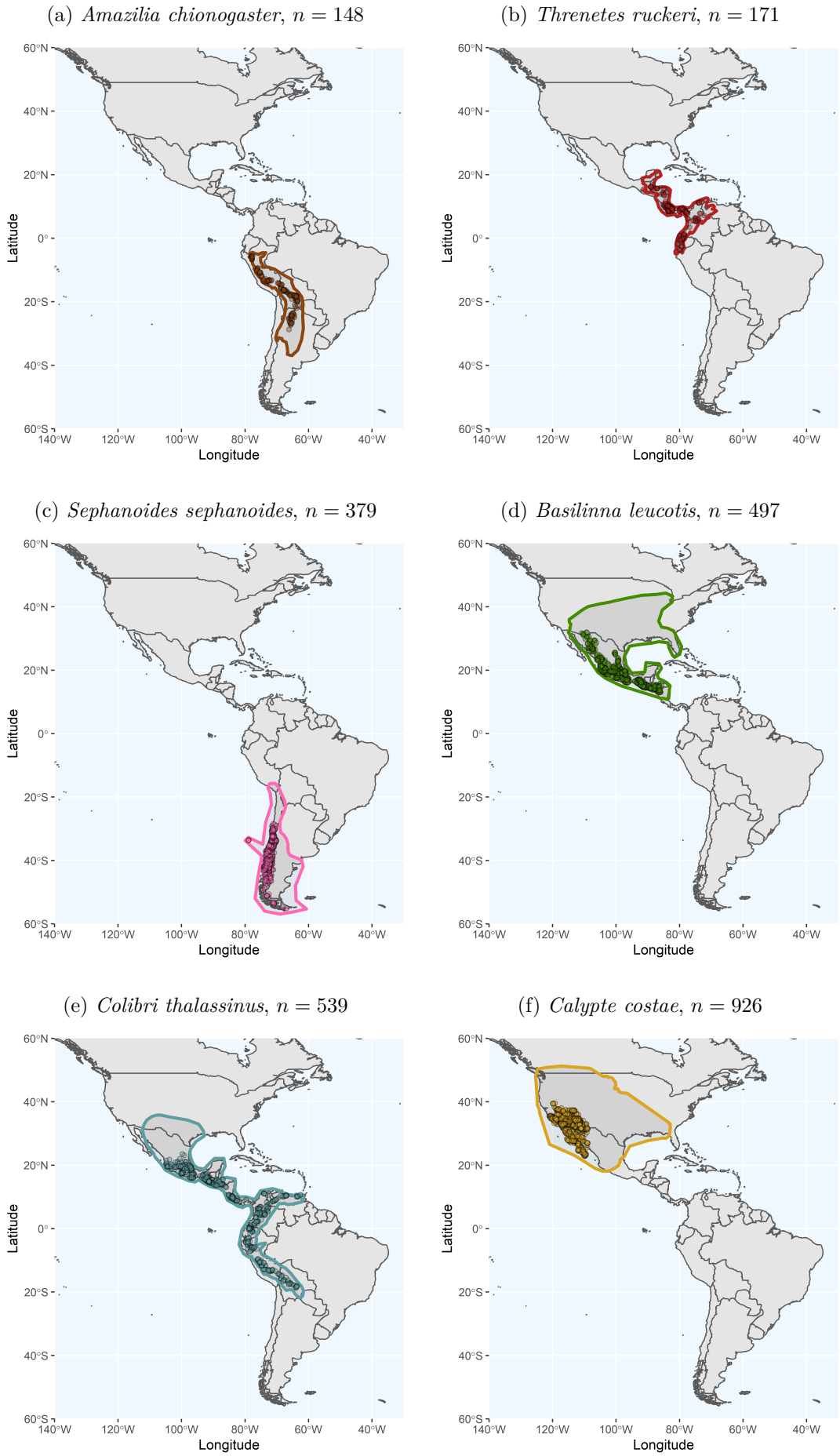


Figure 4: Occurrence samples and M polygons for the six species of hummingbirds selected for the study. (a) *A. chionogaster* will be represented with brown occurrences along this study, (b) *T. ruckeri* in red, (c) *S. sephanoides* in pink, (d) *B. leucotis* in green, (e) *C. thalassinus* in pale blue, and (g) *C. costae* in yellow. The sample sizes (n) are given in each panel.

3 Results

3.1 Estimated fundamental niche of *Vespa mandarinia*

We took a random sample of 10,000 sites inside the accessible (native) region of the Asian giant hornet (shown in Fig. 3), and extracted their environmental values (E_M) so we could plot them in E-space. The resulting set E_M is shown in Figure 5 as a cloud of points in the background (grey open circles). We used this E_M and the 46 presence records of *V. mandarinia* (red points in Fig. 5) to determine the log-likelihood function (Eq. 5) of the parameters that describes its fundamental niche under the Weighted-Normal model, $\theta = (\mu, \Sigma)$. We maximized this log-likelihood function to get the maximum likelihood estimates (MLEs) $\hat{\mu}$ and $\hat{\Sigma}$. Additionally, we used the 46 presence records only to get MLEs under the Mahalanobis model, $\hat{\mu}_0$ and $\hat{\Sigma}_0$. The resulting MLEs for both models are given in Table 1 and the corresponding 99% confidence regions are plotted in Figure 5.

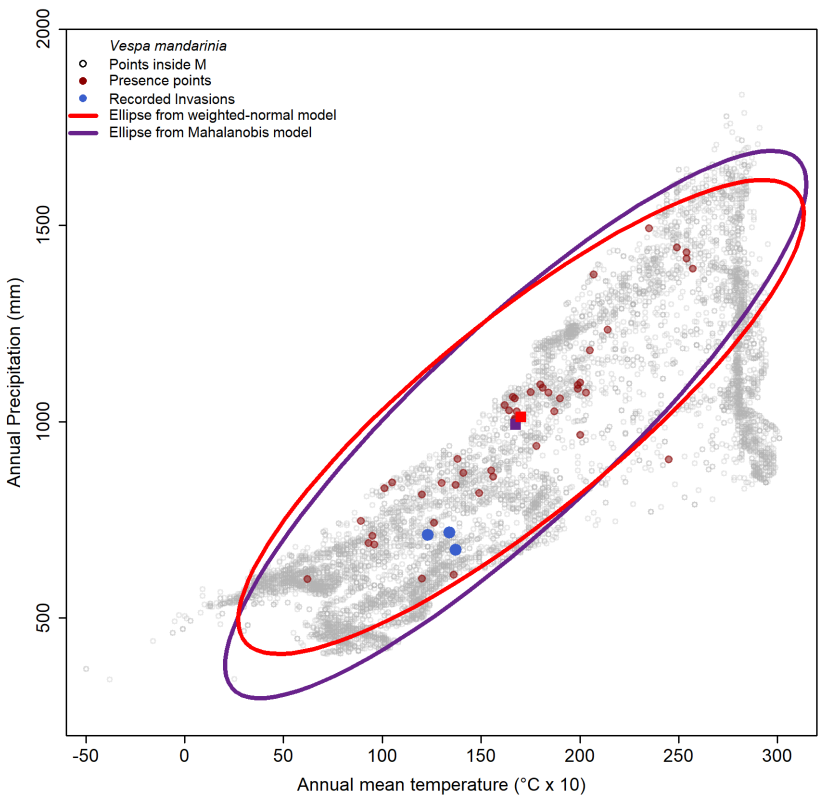


Figure 5: Estimated ellipses (99% confidence regions) from the Weighted-Normal model(red) and the Mahalanobis model (purple) defined by $(\hat{\mu}, \hat{\Sigma})$ and $(\hat{\mu}_0, \hat{\Sigma}_0)$, respectively, as given in Table 1. The red points are the presences from the native range of *V. mandarinia* used to fit the models, while the blue points are presences recorded outside the native range. The centers of the ellipses are indicated with a square of the same color as the corresponding model. The grey points represent E_M.

The MLEs of μ (the optimal environmental conditions) estimated under each model, are very similar (purple and red squares in Figure 5). However, the models predict different limits for the fundamental niche of the Asian giant hornet. Under the Mahalanobis model, the estimated N_F is larger (purple ellipse in Fig. 5). As we can see in Table 1, the variance predicted by the Weighted-Normal modelfor annual precipitation is smaller. Nevertheless, both estimated ellipses contain 45 out of the 46 presence points and all of the presences from the invaded regions are inside the ellipses (blue points in Fig. 5). The presences that come from the invaded regions are not very close to the center of the ellipses, but they are placed in a region of E_M where environmental combinations are well represented in G-space.

We calculated a standardized suitability index for the Asian giant hornet using the MLEs from the Weighted-Normal model and Eq. 2. We created a worldwide suitability map (Fig. 6) to visually assess if there are regions around the world where the species could establish based on the species' environmental requirements alone. In Figure 6, we can see that the West Cost of northern part of the United States and the southern part of Canada, where the species was already confirmed to be established, have a low to moderate suitability index. Similarly, most of Europe's territory has a low to moderate suitability, particularly around the site where the species was already recorded. We provide additional maps focused closely on the West Coast of North America, Europe, and the native range of the species in the Supplementary material. Additionally, there are other regions that are highly suitable for the species, such as western United States, the mountains ranges of Mexico and the Andes, and the Brazilian and Ethiopian highlands.

Table 1: MLEs of the Weighted-Normal model ($\hat{\mu}$ and $\hat{\Sigma}$) and the Mahalanobis model ($\hat{\mu}_0$ and $\hat{\Sigma}_0$) obtained from the 46 presences of *V. mandarinia* inside its native range.

Model	Weighted-Normal		Mahalanobis	
Parameters	$\hat{\mu}$	$\hat{\Sigma}$	$\hat{\mu}_0$	$\hat{\Sigma}_0$
V. mandarinia	(170.21, 1012.19)	$\begin{pmatrix} 2223.45 & 8031.54 \\ 8031.54 & 39554.33 \end{pmatrix}$	(167.52, 992.97)	$\begin{pmatrix} 2348.47 & 9813.12 \\ 9813.12 & 52770.37 \end{pmatrix}$

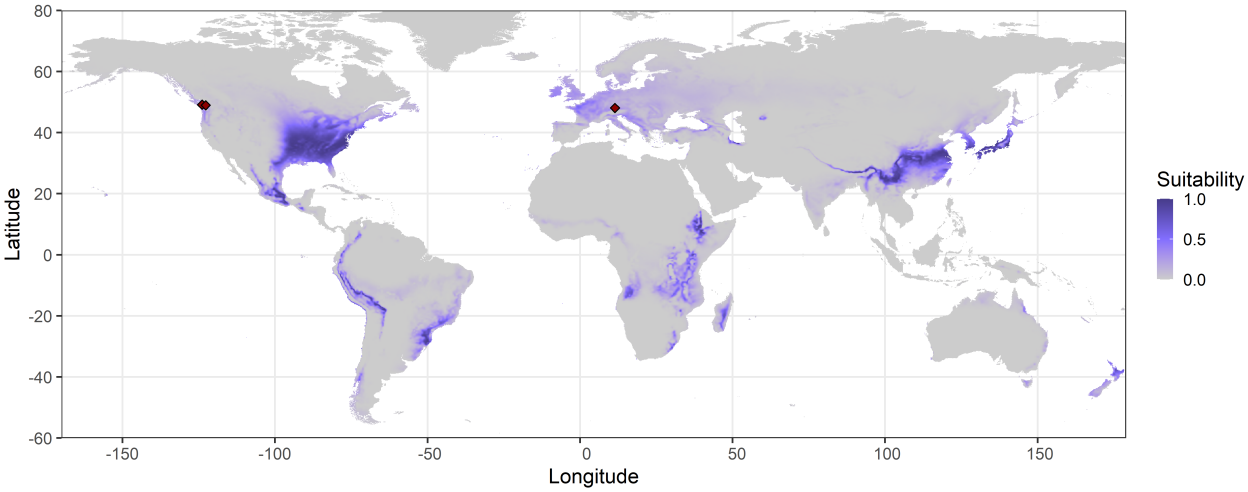


Figure 6: Worldwide suitability map of *V. mandarinia*. Dark purple shades correspond to highly suitable sites whose environmental combinations in E-space are close to the center of the estimated fundamental niche, and light purple shades correspond to sites with low suitability whose environmental combinations in E-space are near the border of the fundamental niche.

3.2 Estimated fundamental niches of Hummingbird species

In the case of the six hummingbird species that we chose for our analysis, not only the accessible areas (Ms) and presence records are located in different regions of G-space but also, in E-space, the corresponding E_M and presence points occupy different regions (see grey clouds of points in Fig. 7). Using these two sources of information (the M and the presences), we estimated the parameters μ and Σ that describe the fundamental niche of each species under the two models: the Mahalanobis model and the Weighted-Normal model. The resulting MLEs for each species are given in Table 2.

Notice that, except for *C. thalassinus*, the ellipses estimated under the Mahalanobis model are smaller than the ones recovered with the Weighted-Normal model. This is, the Weighted-Normal model predicted broader fundamental niches for most of these species. In the case of *A. chionogaster* and *B. leucotis*, the estimated centers of the ellipses ($\hat{\mu}_0$ and $\hat{\mu}$) and their orientation with respect to the axes are very similar. However, there is a clear difference between the centers of the estimated ellipses for the species *T. ruckeri*, *S. sephanoides*, *C. thalassinus*, and *C. costae*.

For all the hummingbird species, both models agreed on the sign of the covariance between the two environmental variables selected to describe and compare the fundamental niches. The

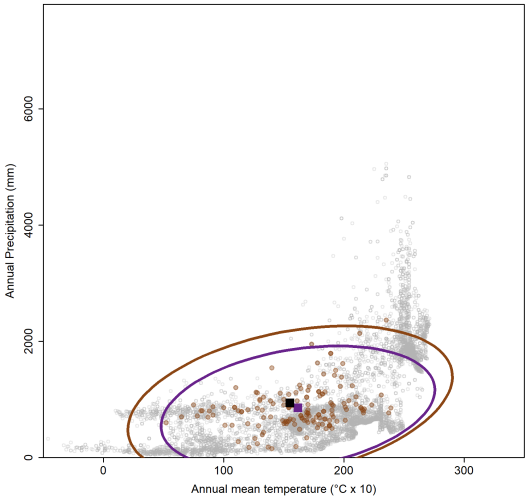
estimated ellipses of *S. sephanoides* are the most different regarding the magnitude of the estimated covariance between the two environmental variables. Additionally, the ellipses that come from the Weighted-Normal model contain more presence points than the ones that come from a simple Mahalanobis model (except for *C. thalassinus*).

It is also worth noticing that, under the Mahalanobis model, the estimated optimum temperature value for four out of the six species lies between the 16 and 18 degrees Celsius (see Table 2). This is, the Mahalanobis model predicts that the species temperature optima does not differ between these species, although they differ in their precipitation optima. On the other hand, under the Weighted-Normal model, the estimated optimum temperature value is clearly different for all the species (see Figure 8).

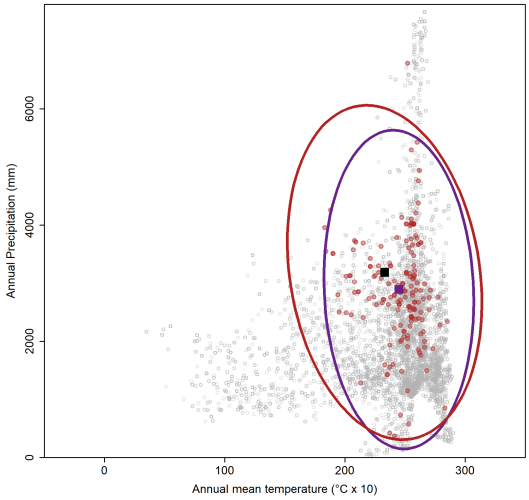
Table 2: Maximum likelihood estimates of the parameters that determine the fundamental niche of the six hummingbird species obtained with the Weighted-Normal model (second column) and the Mahalanobis model (third column).

Model	Weighted-Normal		Mahalanobis	
Species	$\hat{\mu}$	$\hat{\Sigma}$	$\hat{\mu}_0$	$\hat{\Sigma}_0$
A. chionogaster	(155.20, 939.80)	$\begin{pmatrix} 1975.96 & 6702.73 \\ 6702.73 & 191251.29 \end{pmatrix}$	(161.72, 854.68)	$\begin{pmatrix} 1405.95 & 3874.59 \\ 3874.59 & 123768.21 \end{pmatrix}$
T. ruckeri	(232.92, 3187.12)	$\begin{pmatrix} 710.18 & -4647.29 \\ -4647.29 & 897405.47 \end{pmatrix}$	(244.71, 2894.13)	$\begin{pmatrix} 420.79 & -1367.22 \\ -1367.22 & 816368.46 \end{pmatrix}$
S. sephanoides	(125.88, 1341.95)	$\begin{pmatrix} 1877.22 & -11075.44 \\ -11075.44 & 296005.42 \end{pmatrix}$	(108.98, 1208.83)	$\begin{pmatrix} 1024.16 & -10577.69 \\ -10577.69 & 542467.87 \end{pmatrix}$
B. leucotis	(180.99, 1310.40)	$\begin{pmatrix} 1556.64 & 7922.37 \\ 7922.37 & 565818.85 \end{pmatrix}$	(177.08, 1199.03)	$\begin{pmatrix} 1163.15 & 6300.40 \\ 6300.40 & 281258.818 \end{pmatrix}$
C. thalassinus	(146.58, 1826.31)	$\begin{pmatrix} 2103.17 & 11601.23 \\ 11601.23 & 466300.43 \end{pmatrix}$	(177.43, 1640.71)	$\begin{pmatrix} 1821.22 & 13570.96 \\ 13570.96 & 646108.77 \end{pmatrix}$
C. costae	(197.17, 266.32)	$\begin{pmatrix} 2287.66 & -5866.60 \\ -5866.60 & 37195.04 \end{pmatrix}$	(177.87, 277.69)	$\begin{pmatrix} 1563.16 & -3495.43 \\ -3495.43 & 23727.06 \end{pmatrix}$

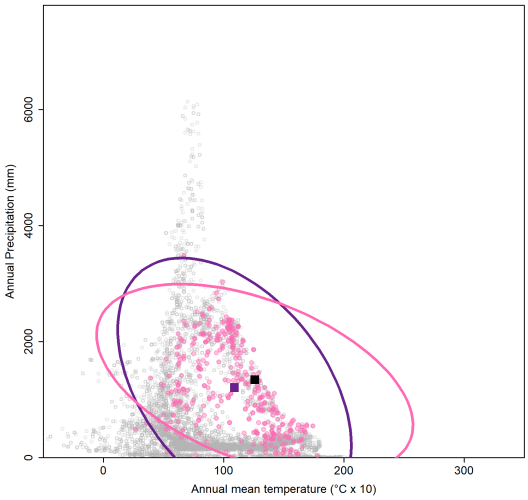
(a) *Amazilia chionogaster*



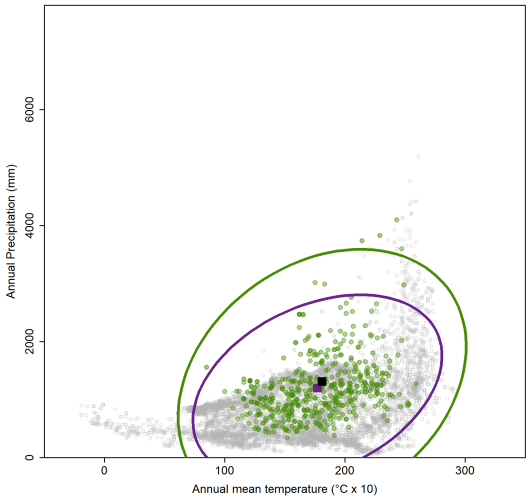
(b) *Threnetes ruckeri*



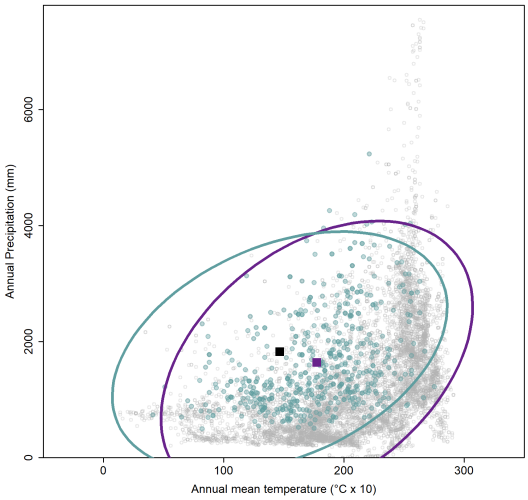
(c) *Sephanoides sephanoides*



(d) *Basilinna leucotis*



(e) *Colibri thalassinus*



(f) *Calypte costae*

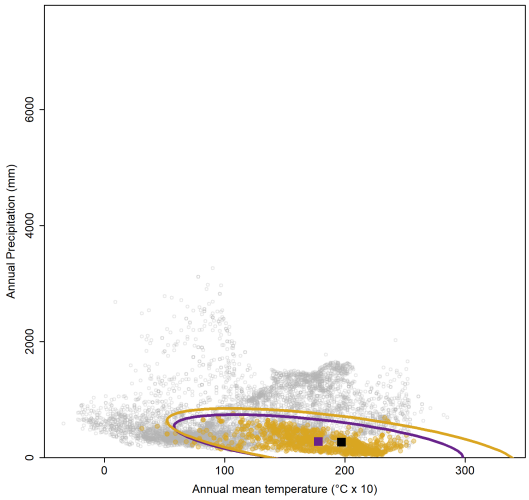


Figure 7: Estimated fundamental niches for the six species of hummingbirds. In all the panels, the purple ellipse represents the estimated niche from a Mahalanobis model, and the second ellipse represent the estimated niche from our proposed weighted model. The centers of both ellipses are marked with a purple square and a black square, respectively.

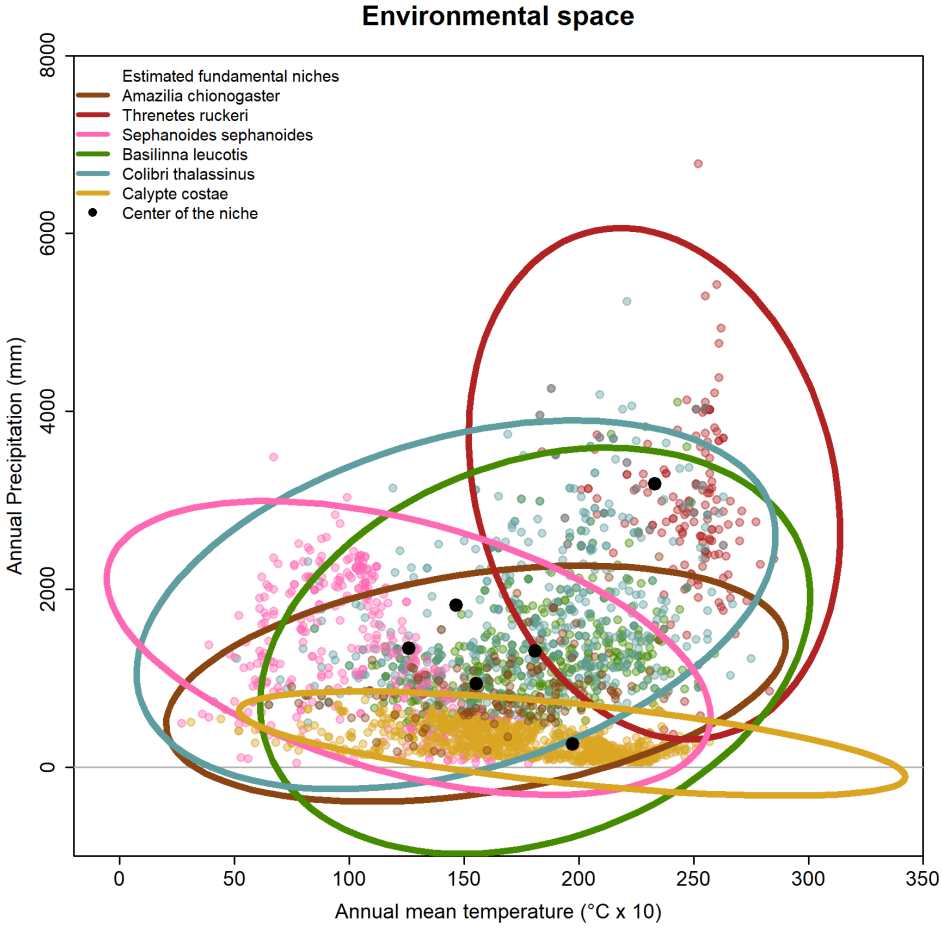


Figure 8: Comparison of the estimated fundamental niches (ellipses) for the six species of hummingbirds and the presence data used to fit the models. The centers of the ellipses are marked with a black circle

4 Discussion

In the last decades, we have seen a significant increase of the availability of presence data and ecological niche modeling (and species distribution modeling) software. However, at the same time, we have seen an increase in the number of recently published studies that question important conceptual and methodological aspects of ENMs (Austin, 2002; Godsoe, 2010; Jiménez-valverde et al., 2008; Lobo, 2008). For that reason, we agree with Jiménez-valverde et al. (2008) in concluding that the lack of a solid conceptual background endangers the advancement of the field. The development and application of ENM/SDM should be rooted in a good understanding of the conceptual background. We intend to lead by example in this work by explicitly relating the ecological theory to the statistical method developed to estimate the fundamental niche of a species.

It is essential that we acknowledge that the presence points of a species come from the realized niche. If our objective is to obtain an approximation to a fundamental niche, from presence data, then we need to account, among other things, for the bias induced in the sample by the available E-space in the region of interest. For this reasons, we proposed to define the distribution of accessible environments to the species based on the dispersal abilities of the species and geographic barriers. This is a definition of M, the area accessible to the species via dispersal (*sensu* Peterson, Soberón, Pearson, et al., 2011), in geographic space, but we highlight the implications to sampling in E, something which is seldom discussed. By projecting the accessible region into E-space, we determined its empirical probability density function and used it as a weight function in the multivariate normal distribution that we use to estimate the fundamental niche of the species of interest.

The main result we have is that taking into account the shape of the E-space defined by an M indeed affects the estimation of the ellipsoids we use to model fundamental niches.

We illustrated the application of the proposed method with two examples. In the first one, we showed how presence data of invasive species can be used to evaluate the estimated shape of the fundamental niche and how the fitted model can be projected back into geography to get a suitability map beyond the region M. The estimated ellipse of *V. mandarinia* estimated with our model contains the known invaded sites and the suitability map shows that the known invaded regions are indeed suitable to the species.

In the second example, we showed how to use existing M hypothesis and presence data to approximate the fundamental niches of closely related species and how to compare the fitted models in environmental space are hummingbirds and, according to phylogenetic taxonomies, they belong to different major clades: *T. ruckeri* is in the Hermits (Phaethornithinae), *C. thalassinus* is in the Mangoes (Polytmini), *S. sephanoides* is in the Coquettes (Lophornithini), *C. costae* is in the Bees (Mellisugini), *A. chianogaster* and *H. leucotis* are in the Emeralds (Trochilini). The clades were listed starting with the one having the oldest split and following the order in which the continued splitting (Hernández-Baños et al., 2014; McGuire et al., 2009). It is interesting to compare the estimated fundamental niches (ellipses) of these six species keeping in mind the phylogenetic relationships among them. For instance, *T. ruckeri* has the oldest split and its estimated fundamental niche (red ellipse in Fig. 8) is the most different among the six species. On the other hand, the rest of the species share some regions in environmental space that belong to their fundamental niches. This can also be concluded by comparing the

centers of the ellipses. These observations support the theory of conservatism of ecological niches, which predicts low niche differentiation between species over evolutionary time scales (Peterson, Soberón & Sánchez-Cordero, 1999).

The central idea in our model is that the set of accessible sites, represented in environmental space (E_M), restricts the sample space from which we can get occurrence records of the species. However, it is universal practice in ENM and SDM to use the environments in the occurrence data to fit niche models. Given that we use the distribution of E_M to inform the model about the uneven availability of environmental combinations, and therefore, about the bias of the sample of presence points induced by M, the delimitation of M is crucial. There is no unique way to delimit the accessible area of a species (N. Barve et al., 2011). When outlining M, we need to take into account several factors simultaneously: the natural history of the species, its dispersal characteristics, the geography of the landscape, the time span relevant to the species's presence and any environmental changes that occurred in that period, among others. Although we do not favor any approach as the bet way to estimate M, we acknowledge that all those factors have an important role in determining its shape. As a future work of this research project, a thorough analysis comparing different approaches to outline M and their effects in recovering the fundamental niche of a species under the proposed model is needed. This can be done using virtual species, where we know the true parameters that shape the fundamental niche, or invasive species, where we used the invaded locations as evaluation points to test the fitted model for the fundamental niche (as showed in one of our examples).

The presence of geographic sampling biases in primary biodiversity data and their implications (e.g., decreasing model performance) are broadly recognized (Kadmon et al., 2004; Meyer et al., 2016). Here, we focused on a particular type of bias which affects the estimation of a fundamental niche. It is important to notice that the other types of sampling biases may also affect the estimation of the fundamental niche (for instance, accessibility of a site to observers). In such cases, it would be relatively easy to incorporate the effect of another source of bias into the proposed model. We would keep using a multivariate normal distribution to represent the hypothesis that the fundamental niche has a simple, convex shape, but we would need to modify the way we define the weights in the likelihood function (Eq. 4). For example, suppose we want to include the effect of two sources of bias that affect the presence of the species, such as the bias induced by M and the differences in sampling intensity across the landscape due to differences in human accessibility (accessibility bias). Each source of bias will define a different sampling

probability distribution across the environmental combinations from the region of interest. If the types of biases are independent, we can calculate a single, general sampling probability distribution as the product of the individual sampling probability distributions, and substitute it in Eq. 4. This, the function $w(\boldsymbol{x})$ can be defined as the product of two probabilities, the probability of observing $\boldsymbol{x}$ considering the distribution of points in E_M and the probability of observing $\boldsymbol{x}$ taking into account the accessibility bias. We proposed to use a kernel to estimate the former, and Zizka et al. (2020) recently developed a method to quantify accessibility biases that could be used to determine the later.

The proposed model to estimate fundamental niches can be generalized in different ways to represent more realistic scenarios. Above, we mentioned a way to include more than one type of biases but we can list some other possible modifications. It is possible to include information from physiological experiments if we consider the Bayesian approach proposed by Jiménez et al. (2019). For this, we need to transform the likelihood function (Eq. 4) into a posterior distribution that uses the tolerance ranges obtained from physiological experiments to define the *a priori* distributions of the parameters μ and Σ ($g_1(\mu)$ and $g_2(\Sigma)$, respectively) as follows:

$$f(\mu, \Sigma; D) = \mathcal{L}(\mu, \Sigma|D) g_1(\mu) g_2(\Sigma) \tag{6}$$

where

$$\mathcal{L}(\mu, \Sigma|D) \propto \prod_{i=1}^{n} \frac{w(\boldsymbol{x}_i) f(\boldsymbol{x}_i; \mu, \Sigma)}{\sum_{y \in E} w(y) f(y; \mu, \Sigma)}. \tag{7}$$

Notice that the function that determines the shape and size of the fundamental niche, $f(\cdot; \mu, \Sigma)$, can be modified to have asymmetrical confidence regions for species that are known to have asymmetrical response curves for the environmental variables under consideration (Jiménez et al., 2019). Furthermore, both the likelihood and the Bayesian approaches can be applied in cases where it makes sense for the species to include to more than two environmental variables in the delimitation of its fundamental niche. The downside of making these modifications in the proposed model is that they increase the complexity of the algebraic expression the model and the number of parameters to be estimated, which will require more computational power to either maximize the likelihood function (if we do not include physiological tolerances and use the likelihood approach), or simulate from the posterior distribution (when considering physiological tolerances, under the Bayesian approach).

Acknowledgements

We thank the KU-ENM group at the Biodiversity Institute of the University of Kansas for providing valuable comments. L.J. was supported by the Mexican Council of Science and Technology, CONACyT, grant 409052.

References

Anderson, R. P. (2012). Harnessing the world's biodiversity data: promise and peril in ecological niche modeling of species distributions. *Annals of the New York Academy of Sciences* 1260.1, pp. 66–80. DOI: 10.1111/j.1749-6632.2011.06440.x.

Aspinall, R. & B. Lees (1994). Sampling and analysis of spatial environmental data. *Advances in GIS Research. Taylor and Francis, Southampton*, pp. 1086–1098.

Austin, M. (2002). Spatial prediction of species distribution: an interface between ecological theory and statistical modelling. *Ecological Modelling* 157.2-3, pp. 101–118. DOI: 10.1016/S0304-3800(02)00205-3.

Barve, N., V. Barve, A. Jiménez-Valverde, A. Lira-Noriega, S. P. Maher, A. T. Peterson, J. Soberón & F. Villalobos (2011). The crucial role of the accessible area in ecological niche modeling and species distribution modeling. *Ecological Modelling* 222.11, pp. 1810–1819. DOI: 10.1016/j.ecolmodel.2011.02.011.

Bivand, R., T. Keitt & B. Rowlingson (2020). *rgdal: Bindings for the 'Geospatial' Data Abstraction Library*. R package version 1.5-10. URL: https://CRAN.R-project.org/package=rgdal.

Bivand, R. & C. Rundel (2019). *rgeos: Interface to Geometry Engine - Open Source ('GEOS')*. R package version 0.5-2. URL: https://CRAN.R-project.org/package=rgeos.

Brown, J. H. (1984). On the Relationship between Abundance and Distribution of Species. *The American Naturalist* 124.2, pp. 255–279. DOI: 10.1093/ehr/cepl85.

Chapman, A. D. (2005). Principles and methods of data cleaning–primary species and species-occurrence data, version 1.0. *Global Biodiversity Information Facility, Copenhagen* 75.

Cobos, M. E., L. Jiménez, C. Nuñez-Penichet, D. Romero-Alvarez & M. Simoes (2018). Sample data and training modules for cleaning biodiversity information. *Biodiversity Informatics* 13, pp. 49–50. DOI: 10.17161/bi.v13i0.7600.

Colwell, R. K. & T. F. Rangel (2009). Hutchinson's duality: The once and future niche. *Proceedings of the National Academy of Sciences of the United States of America* 106.SUPPL. 2, pp. 19651–19658. DOI: 10.1073/pnas.0901650106.

Cooper, J. C. & J. Soberón (2018). Creating individual accessible area hypotheses improves stacked species distribution model performance. *Global Ecology and Biogeography* 27.1, pp. 156–165. DOI: 10.1111/geb.12678.

Drake, J. M. (2015). Range bagging: A new method for ecological niche modelling from presence-only data. *Journal of the Royal Society Interface* 12.107. DOI: 10.1098/rsif.2015.0086.

Duong, T. (2020). *ks: Kernel Smoothing*. R package version 1.11.7. URL: https://CRAN.R-project.org/package=ks.

Etherington, T. & O. Omondiagbe (2019). virtualNicheR: generating virtual fundamental and realised niches for use in virtual ecology experiments. *Journal of Open Source Software* 4.41, p. 1661. DOI: 10.21105/joss.01661.

Farber, O. & R. Kadmon (2003). Assessment of alternative approaches for bioclimatic modeling with special emphasis on the Mahalanobis distance. *Ecological Modelling* 160.1-2, pp. 115–130. DOI: 10.1016/S0304-3800(02)00327-7.

GBIF (2020). *Occurrence download*. Accessed: 07 May 2020. DOI: 10.15468/dl.kzcgc2. URL: https://www.gbif.org/.

Godsoe, W. (2010). I can't define the niche but i know it when i see it: A formal link between statistical theory and the ecological niche. *Oikos* 119.1, pp. 53–60. DOI: 10.1111/j.1600-0706.2009.17630.x.

Guisan, A. et al. (2013). Predicting species distributions for conservation decisions. *Ecology Letters* 16.12, pp. 1424–1435. DOI: 10.1111/ele.12189.

Hernández-Baños, B. E., L. E. Zamudio-Beltrán, L. E. Eguiarte-Fruns, J. Klicka & J. García-Moreno (2014). The Basilinna genus (Aves: Trochilidae): An evaluation based on molecular evidence and implications for the genus Hylocharis. *Revista Mexicana de Biodiversidad* 85.3, pp. 797–807. DOI: 10.7550/rmb.35769.

Hijmans, R. J. (2020). *raster: Geographic Data Analysis and Modeling*. R package version 3.1-5. URL: https://CRAN.R-project.org/package=raster.

Hijmans, R. J., S. E. Cameron, J. L. Parra, P. G. Jones & A. Jarvis (2005). Very high resolution interpolated climate surfaces for global land areas. *International Journal of Climatology* 25.15, pp. 1965–1978. DOI: 10.1002/joc.1276.

Hoogenboom, M. O. & S. R. Connolly (2009). Defining fundamental niche dimensions of corals: synergistic effects of colony size, light, and flow. *Ecology* 90.3, pp. 767–780. DOI: 10.1890/07-2010.1.

Hutchinson, G. E. (1957). Concluding remarks. *Cold Sprig Harbor Symposia on Quantitative Biology*. Chap. 22, pp. 415–427.

Jackson, S. T. & J. T. Overpeck (2000). Responses of plant populations and communities to environmental changes of the late Quaternary. *Paleobiology* 26.4 SUPPL. Pp. 194–220. DOI: 10.1017/s0094837300026932.

Jiménez, L., J. Soberón, J. A. Christen & D. Soto (2019). On the problem of modeling a fundamental niche from occurrence data. *Ecological Modelling* 397.February, pp. 74–83. DOI: 10.1016/j.ecolmodel.2019.01.020.

Jiménez-valverde, A., J. M. Lobo & J. Hortal (2008). Not as good as they seem : the importance of concepts in species distribution modelling, pp. 885–890. DOI: 10.1111/j.1472-4642.2008.00496.x.

Kadmon, R., O. Farber & A. Danin (2004). Effect of roadside bias on the accuracy of predictive maps produced by bioclimatic models. *Ecological Applications* 14.2, pp. 401–413. DOI: 10.1890/02-5364.

Lele, S. R. & J. L. Keim (2006). Weighted distributions and estimation of resource selection probability functions. *Ecology* 87.12, pp. 3021–3028. DOI: 10.1890/0012-9658(2006)87[3021:WDAEOR]2.0.CO;2.

Lobo, J. M. (2008). More complex distribution models or more representative data. 82, pp. 14–19.

Lobo, J. M. (2016). The use of occurrence data to predict the effects of climate change on insects. *Current Opinion in Insect Science* 17, pp. 62–68. DOI: 10.1016/j.cois.2016.07.003.

Maguire, B. (1973). Niche Response Structure and the Analytical Potentials of Its Relationship to the Habitat Author (s): Bassett Maguire , Jr . Source : The American Naturalist , Vol . 107 , No . 954 (Mar . - Apr ., 1973), pp . 213-246 Published by : The University of C. *The American naturalist* 107.954, pp. 213–246.

Matsuura, M. & S. Sakagami (1973). A bionomic sketch of the giant hornet. *Vespa mandarinia*.

Matsuura, M. (1988). Ecological study on vespine wasps (Hymenoptera: Vespidae) attacking honeybee colonies: I. seasonal changes in the frequency of visits to apiaries by vespine wasps

and damage inflicted, especially in the absence of artificial protection. *Applied Entomology and Zoology* 23.4, pp. 428–440.

McGuire, J. A., C. C. Witt, J. V. Remsen, R. Dudley & D. L. Altshuler (2009). A higher-level taxonomy for hummingbirds. *Journal of Ornithology* 150.1, pp. 155–165. DOI: 10.1007/s10336-008-0330-x.

Meyer, C., P. Weigelt & H. Kreft (2016). Multidimensional biases, gaps and uncertainties in global plant occurrence information. *Ecology letters* 19.8, pp. 992–1006. DOI: 10.1111/ele.12624.

Osorio-Olvera, L., C. Yañez-Arenas, E. Martínez-Meyer & A. T. Peterson (2020). *Relationships between population densities and niche-centroid distances in North American birds*. DOI: 10.1111/ele.13453.

Owens, H. L., L. P. Campbell, L. L. Dornak, E. E. Saupe, N. Barve, J. Soberón, K. Ingenloff, A. Lira-Noriega, C. M. Hensz, C. E. Myers & A. T. Peterson (2013). Constraints on interpretation of ecological niche models by limited environmental ranges on calibration areas. *Ecological Modelling* 263, pp. 10–18. DOI: 10.1016/j.ecolmodel.2013.04.011.

Owens, H. L., V. Ribeiro, E. E. Saupe, M. E. Cobos, P. A. Hosner, J. C. Cooper, A. M. Samy, V. Barve, N. Barve, C. J. Muñoz-R. & A. T. Peterson (2020). Acknowledging uncertainty in evolutionary reconstructions of ecological niches. *Ecology and Evolution* 10.14, pp. 6967–6977. DOI: 10.1002/ece3.6359.

Patil, G. & J. Ord (1976). On Size-Biased Sampling and Related Form-Invariant Weighted Distributions. *Sankhyā: The Indian Journal of StatisticsSeries B* 38.1, pp. 48–61.

Patil, G. & C. Rao (1978). Weighted Distributions and Size-Biased Sampling with Applications to Wildlife Populations and Human Families. *Biometrics* 34.2, pp. 179–189.

Pebesma, E. (2018). Simple Features for R: Standardized Support for Spatial Vector Data. *The R Journal* 10.1, pp. 439–446. DOI: 10.32614/RJ-2018-009.

Peterson, A. T. & J. Soberón (2012). Integrating fundamental concepts of ecology, biogeography, and sampling into effective ecological niche modeling and species distribution modeling. *Plant Biosystems* 146.4, pp. 789–796. DOI: 10.1080/11263504.2012.740083.

Peterson, A. T., J. Soberón, R. G. Pearson, R. P. Anderson, E. Martínez-Meyer, M. Nakamura & M. B. Araújo (2011). Ecological niches and geographic distributions. *Princeton University Press* 49.11, pp. 1–314. DOI: 10.5860/choice.49-6266.

Peterson, A. T., J. Soberón & V. Sánchez-Cordero (1999). Conservatism of ecological niches in evolutionary time. *Science* 285.5431, pp. 1265–1267. DOI: 10.1126/science.285.5431.1265.

Pulliam, H. R. (2000). On the relationship between niche and distribution. *Ecology Letters* 3.4, pp. 349–361. DOI: 10.1046/j.1461-0248.2000.00143.x.

R Core Team (2020). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria. URL: https://www.R-project.org/.

Robert, C. & G. Casella (1999). *Monte Carlo statistical methods*. Springer Science & Business Media.

Root, T. (1988). Environmental Factors Associated with Avian Distributional Boundaries. *Journal of Biogeography* 15.3, p. 489. DOI: 10.2307/2845278.

Saupe, E. E., V. Barve, C. E. Myers, J. Soberón, N. Barve, C. M. Hensz, A. T. Peterson, H. L. Owens & A. Lira-Noriega (2012). Variation in niche and distribution model performance: The need for a priori assessment of key causal factors. *Ecological Modelling* 237-238, pp. 11–22. DOI: 10.1016/j.ecolmodel.2012.04.001.

Saupe, E. E., N. Barve, H. L. Owens, J. C. Cooper, P. A. Hosner & A. T. Peterson (2017). Reconstructing ecological niche evolution when niches are incompletely characterized. *Systematic Biology* 67.3, pp. 428–438. DOI: 10.1093/sysbio/syx084.

Soberón, J. (2007). Grinnellian and Eltonian niches and geographic distributions of species. *Ecology Letters* 10.12, pp. 1115–1123. DOI: 10.1111/j.1461-0248.2007.01107.x.

Soberón, J. & M. Nakamura (2009). Niches and distributional areas: Concepts, methods, and assumptions. *Proceedings of the National Academy of Sciences of the United States of America* 106.SUPPL. 2, pp. 19644–19650. DOI: 10.1073/pnas.0901637106.

Soberón, J. & A. T. Peterson (2019). What is the shape of the fundamental Grinnellian niche? *Theoretical Ecology* May. DOI: 10.1007/s12080-019-0432-5.

Tingley, R., M. Vallinoto, F. Sequeira & M. R. Kearney (2014). Realized niche shift during a global biological invasion. *Proceedings of the National Academy of Sciences* 111.28, pp. 10233–10238. DOI: 10.1073/pnas.1405766111.

Warren, D. L. (2012). In defense of 'niche modeling'. *Trends in Ecology and Evolution* 27.9, pp. 497–500. DOI: 10.1016/j.tree.2012.03.010.

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. URL: https://ggplot2.tidyverse.org.

Wickham, H. & D. Seidel (2020). *scales: Scale Functions for Visualization*. R package version 1.1.1. URL: https://CRAN.R-project.org/package=scales.

Wilson, T. M., J. Takahashi, S.-E. Spichiger, I. Kim & P. van Westendorp (2020). First Reports of Vespa mandarinia (Hymenoptera: Vespidae) in North America Represent Two Separate Maternal Lineages in Washington State, United States, and British Columbia, Canada. *Annals of the Entomological Society of America* 113.6, pp. 468–472. DOI: 10.1093/aesa/saaa024.

Zizka, A., A. Antonelli & D. Silvestro (2020). Sampbias, a Method for Quantifying Geographic Sampling Biases in Species Distribution Data. *Ecography*, pp. 1–8. DOI: 10.1111/ecog.05102.

Supplementary materials

On the main text, we provided a worldwide suitability map for *V. mandarinia* obtained with the Weighted-Normal model proposed to estimate fundamental niches. Here, we provide additional maps closely focused on some areas of concern for the invasive species: (1) the native range of the species in Figure S1, (2) Europe in Figure S2, and (4) the West Coast of North America in Figure S3. The same color scale used in Figure 6 was used in this figures. Dark purple shades correspond to highly suitable sites whose environmental combinations in E-space are close to the center of the estimated fundamental niche, and light purple shades correspond to sites with low suitability whose environmental combinations in E-space are near the border of the fundamental niche.

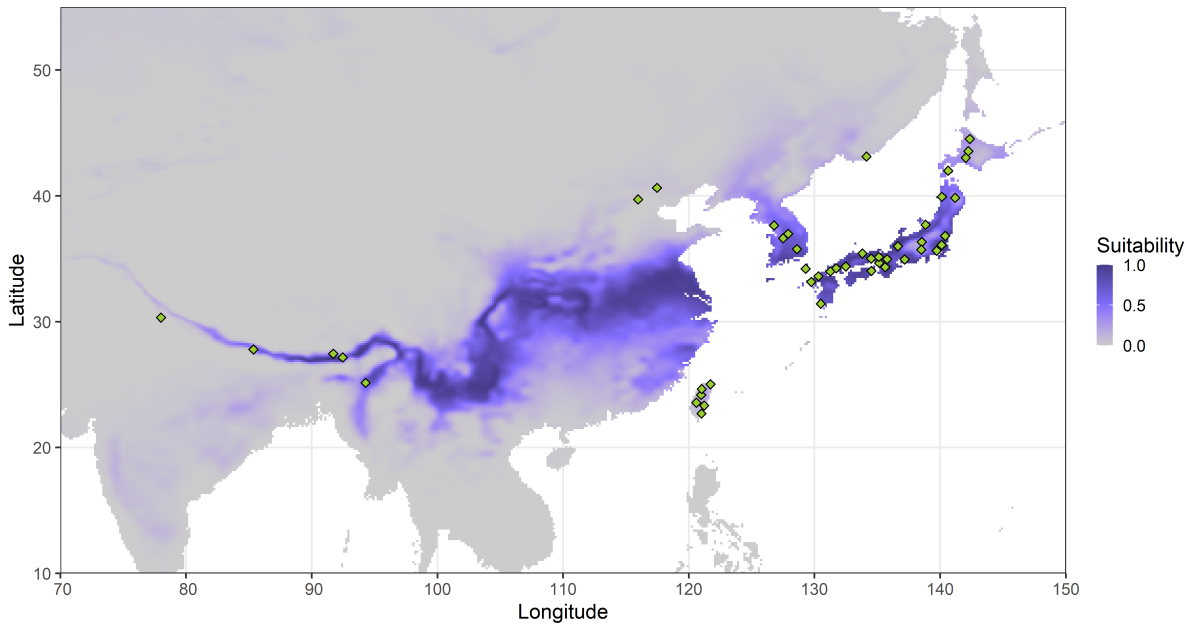


Figure S1: Suitability index obtained with the Weighted-Normal model plotted in the native range of *V. mandarinia*. The green squares are the occurrence points used to estimate the fundamental niche of the species.

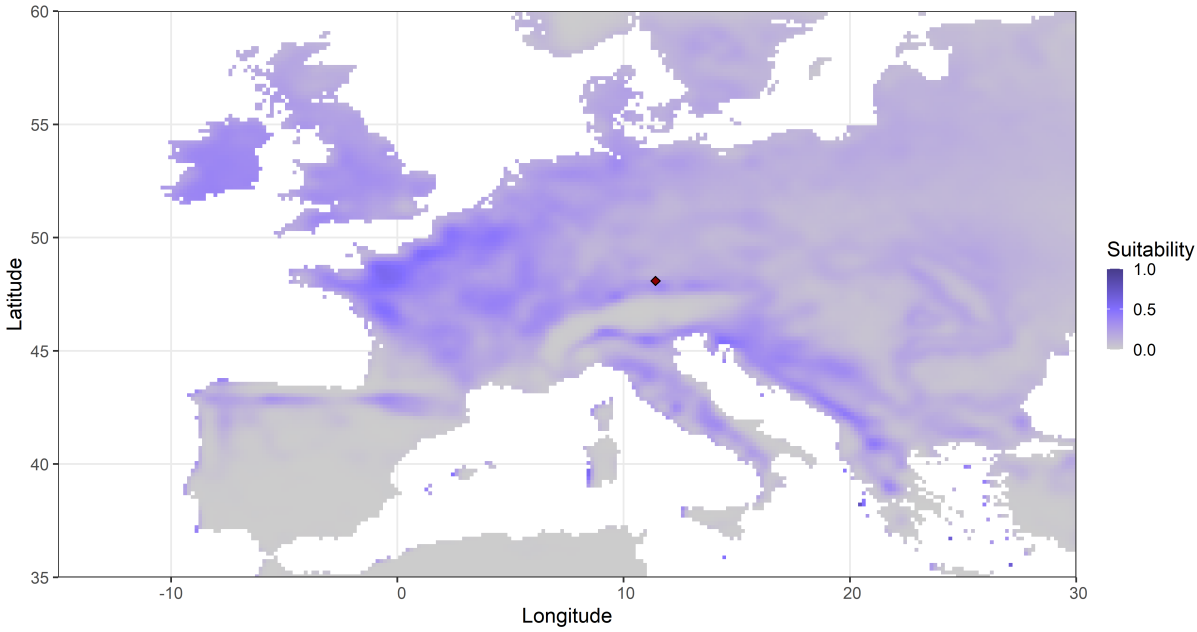


Figure S2: Suitability index obtained with the Weighted-Normal model for *V. mandarinia* plotted in the Europe. The red square is an occurrence point confirmed in Germany.

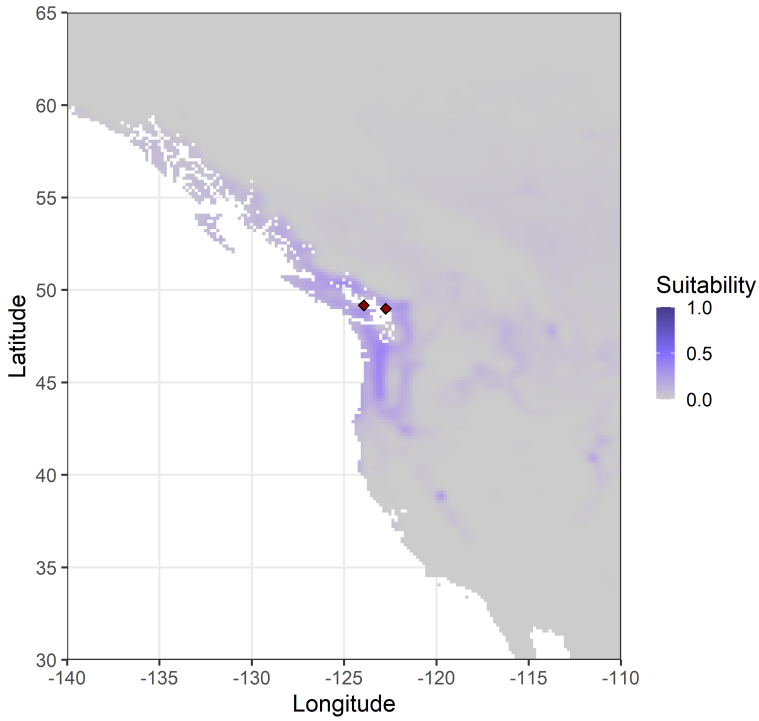


Figure S3: Suitability index obtained with the Weighted-Normal model for *V. mandarinia* plotted in the West Coast of Canada and the United States. The red squares are occurrence points confirmed in each country.