Smoothing and the Environmental Manifold

Siddharth Unnithan Kumar[a,b,*], Philip K. Maini[a], Luca Chiaverini[b], Andrew J. Hearn[b], David W. Macdonald[b], Żaneta Kaszta[b], Samuel A. Cushman[b,c]

a: Mathematical Institute, University of Oxford, Oxford, United Kingdom

b: Wildlife Conservation Research Unit (WildCRU), Department of Zoology, University of Oxford, Oxford, United Kingdom

c: US Forest Service, Rocky Mountain Research Station, Flagstaff, AZ, USA

* Corresponding author. Email address: siddharth.unnithankumar@gmail.com

Abstract

How the observed occurrences of a species relate to environmental gradients is a fundamental question in community ecology. In this paper, we present a new approach to address this question, using the smoothing function, a method not previously recruited into this ecological context. Using simulation techniques, we explore its accuracy in recovering known species distributions from simulated noisy data, and we compare the smoothing function's predictive abilities to two widely used methods in this field, the generalised linear model (GLM) and random forest machine learning. In studying the smoothing function, we are led to consider a new analytical tool for ecology, which we call the environmental manifold. It is given by the shape of the data cloud of sampled predictor variables, and has deep relevance to ecological niche theory. Hitherto not considered in ecological analyses, it plays a fundamental role in understanding the species-environment relationship, and we utilise it to compare the performance and behaviour of these three methods.

The results of our analysis find both random forest and smoothing to be robust to the complexities of the species-environment relationship, and also, to a degree, the shape of the environmental manifold. In contrast, the GLM's accuracy depends heavily on the complexity of the species-environment relationship, and is also affected by the geometry of the environmental manifold. Furthermore, the smoothing function is seen to be more accurate than random forest

in every combination of species-environment relationship and environmental manifold shape, and also less affected by sampling bias. This suggests the promising role that such smoothing functions can have in ecological analyses. Our results also support the robustness of random forest machine learning to nonlinearity in both the species-environment relationship, and for the first time, the complexity of the shape of the environmental manifold. We conclude by discussing the implications and uses of the environmental manifold in ecological practice and theory, including its importance for niche theory, understanding species distributions, and conservation policy decisions.

Keywords: Kernel smoothing; Environmental manifold; Random forest machine learning; Species-environment relationship; Predictive modeling; Hutchinsonian niche

Tired of all who come with words, words
but no language
I went to the snow-covered island.
The wild does not have words.
The unwritten pages spread themselves
out in all directions!
I come across the marks of roe-deer's
hooves in the snow.
Language, but no words.

Tomas Tranströmer

1 Introduction

In community ecology, a central question is to understand the relationships between observed occurrences of a given species and the environmental factors (such as elevation, or forest cover) which influence these patterns of occurrence. The ecological niche of a species is an idea encompassing a variety of related concepts, such as the Grinnellian and Eltonian niches, which are used to describe and understand these relationships (Soberón 2007). The formulation of the niche concept which has perhaps most informed quantitative ecological analyses is the Hutchinsonian niche, proposed in the mid-20th century by G. Evelyn Hutchinson (Hutchinson 1957). The Hutchinsonian niche has two aspects: the fundamental niche, and the realised niche. The former is commonly considered to be the region of abstract multi-dimensional environmental space (or 'niche space') corresponding to the

combination of conditions theoretically required for a species to survive and reproduce; the latter is defined to be the subset of this region which the species is actually found to inhabit (Holt 2009). There has been a great deal of work in recent times focused on describing the properties and shape of these two aspects of the Hutchinsonian niche (Blonder et al. 2014; Broennimann et al. 2012), with deeper insights now made possible by the revolutionary advances in high-quality GIS and occurrence data and computational resources (Samuel A. Cushman and Huettman 2010).

In light of this, a range of statistical tools has been developed to study the species-environment relationship quantitatively, and to describe the shape and dimensions of the fundamental and realised niche (Hegel et al. 2010). Two of the most popular tools for this endeavour are the generalised linear model (GLM) and the random forest machine learning algorithm (Nelder and Wedderburn 1972; Breiman 2001; Liaw, Wiener, et al. 2002). The GLM proceeds by assuming a linear relationship between species abundance and environmental variables, and then estimating a linear combination of environmental variables that best predicts species responses. It is possible to incorporate non-linearities into the GLM (such as square or exponential relationships), but these must be specified precisely when setting up the model, which is difficult a priori (Whittingham et al. 2006; Ash et al. 2021). Indeed, by prescribing the functional shape prior to analysis, the GLM is limited in its ability to accurately account for the complex and unknown relationships among species responses and environmental variables; in this paper, we configure the GLM with linear predictors, as it is most often used in ecological analyses (McGarigal et al. 2016). Random forest is a machine learning algorithm, which uses bootstrapping of classification and regression trees to recover the relationships between species occurrences and environmental variables. In contrast with the GLM, it is nonparametric (Evans et al. 2011), and with its techniques of bagging, subsampling and cross-validation, random forest has exceptional predictive ability, and very good performance in accounting for nonlinear and interactive effects (Breiman 1996; Cutler et al. 2007; Mi et al. 2017).

In this paper, we compare the behaviour and performance of the GLM and random forest in predicting simulated species-environment relationships with a third method, called the smoothing function. The smoothing function, in essence, numerically smooths the species response along environmental gradients in multiple dimensions; it can be used to describe and predict species-environment relationships nonparametrically, and to address scale dependence of species-environment relationships in environmental space. This is important because, while there has been much attention to scale dependence of species-environment relationships in geographical space (Wiens 1989; Levin 1992), scale dependence in environmental space has been almost completely unexplored in the ecological liter-

ature. In geographical space, scale dependence concerns a species' selection of habitat features at multiple spatial scales, which is a fundamental ingredient in the inference of the relationship between a species and their habitat, and in which smoothing in geographical space plays a key role (Samuel A Cushman and McGarigal 2002; Chandler and Hepinstall-Cymerman 2016). Scale dependence and smoothing in environmental space, however, is an inherently different phenomenon. Smoothing in environmental space is tied to the scales at which niche dimensions influence species occurrence, and it relates to the sensitivity and degree to which changes in niche variables influence and limit species occurrence. Therefore, it is crucial to consider scale dependence in environmental space in order to accurately measure the effective niche structure of a species and to assess how issues such as sampling bias affect its estimation. Our smoothing method addresses this second aspect of scale, as illustrated in Section 2.1. A similar smoothing method, for measuring niche overlap in environmental space, appeared in Broennimann et al. 2012, which suggested the promising role such functions could have in ecological analyses. In Broennimann et al. 2012 it was said that 'the use of a kernel smoother makes the process of moving from geographical space to multivariate environmental space independent of both sampling effort and arbitrary choice of resolution in environmental space', a claim which we will revisit in our discussion.

In addition to our smoothing function, we introduce a new concept called the environmental manifold, which we describe in detail in Section 2.2. Briefly, the environmental manifold is the shape of the data cloud of sampled environmental variables in multi-dimensional environmental space. It turns out to be a crucial ingredient in understanding the observed response between a species and their environment, as discussed in Sections 2.1 and A.1, and its multivariate nature affords new insights into ecological data by allowing us to observe and analyse the nonlinear interactions of several environmental variables simultaneously. A deep exploration of its properties and relations to other ecological analyses lies beyond the scope of this paper, but we will touch on its foundational connection to the Hutchinsonian realised niche in Section 4. As will become clear from its construction in this paper, the environmental manifold is precisely the *realisable* niche - the geographically realised subset of environmental space on which the realised niche is constrained to occur - suggesting profound implications for both ecological practice and theory.

After describing the smoothing function and the environmental manifold, our analysis will involve simulating species-environment relationships, and investigating the ability of the GLM, random forest and smoothing methods to recover these relationships. Simulation modeling has particular advantages for this kind of exploration. Namely, simulation enables us to stipulate, a priori, a known relationship

between multiple environmental gradients and the species response (Samuel A. Cushman and Erin L. Landguth 2010; Shirk, S. Cushman, and E. Landguth 2012). This in turn enables us to have a 'known truth' to which we can compare the performance of the different methods in terms of their ability to correctly identify the variables and estimate the parameters of their relationship with species occurrence (e.g. Atzeni et al. 2020; Chiaverini et al. 2021). Using this simulation framework, we address the following question: Do all three methods recover the known relationship, and do they perform equally well when the relationship is complex? To do this, we evaluate their performance along two gradients of increasing complexity: (1) from a linear response to a nonlinear response, and (2) from a simple isotropic environmental manifold, across a range of more interesting geometries, to the actual complexity of a three-dimensional environmental manifold from a case-study landscape in Borneo, where our team has studied the patterns of biodiversity in relation to environmental gradients (Hearn et al. 2018; Kaszta et al. 2019).

2 Methods

2.1 The smoothing function

The method which we call the smoothing function is mathematically described as a Gaussian kernel smoothing function (similar to the Gaussian kernel density function seen in Silverman 1986). Explicitly, given a collection of occurrence data (which can be binary presence-absence data, or abundance data with multiple occurrences) distributed along some environmental gradient, the value $V(x)$ at each occurrence point x is replaced by a weighted average $S(x)$ of the values at the neighbouring occurrence points, where the weighting function K_σ is, in this case, a Gaussian function. This can be expressed mathematically as

$$S(x) = \frac{\sum K_\sigma(x, x') \cdot V(x')}{\sum K_\sigma(x, x')}, \quad \text{where} \quad K_\sigma(x, x') = e^{-\frac{(x-x')^2}{2\sigma^2}}$$

with the sum taken over all points x' in the data set. Sample code for the smoothing function is provided in Section A.2.

In contrast with the GLM, no underlying relationship is assumed, and thus this approach is nonparametric, like the random forest method. The smoothing method can be used for predicting species distributions, by using the smoothed occurrence points to obtain an estimated probability distribution of species occurrence, and in this paper we will compare its predictive ability with

that of the GLM and random forest. One key difference between the smoothing and the other two methods is that it may be expressed as a genuine mathematical function on the occurrence data points themselves, and is thus more straightforward to work with from a mathematical point of view, as will be discussed in the appendices. On the other hand, in comparison with the regression coefficients predicted by the GLM, and the bagging rules used by random forest, the smoothing approach, by its inherently local nature (of smoothing around the provided occurrence data points) need not give precise information on extrapolating predictions of occurrence probabilities for very different environmental conditions to those from which the occurrence data were gathered.

In Figure 1, we illustrate the effect of this Gaussian smoothing on the response of the Sunda clouded leopard to elevation, in Sabah, Borneo, using data from Hearn et al. 2018. We see the effect of smoothing as we increase σ, starting from the top-left diagram in which $\sigma = 0$ (which means that no smoothing has been applied). An analogous multivariate smoothing happens when we consider multiple environmental variables simultaneously.

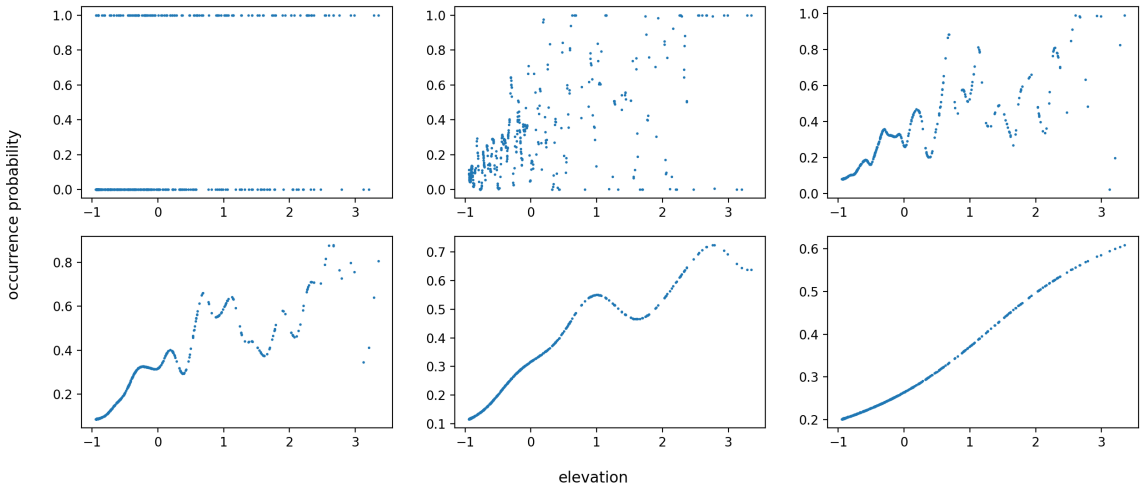

Figure 1: The effects of smoothing on binary presence-absence data. The top-left graph consists of the original occurrence data, with increasing smoothing as we move from left to right along the first and then second row ($\sigma = 0.01, 0.05, 0.1, 0.3, 1$ respectively). The data are occurrences of the Sunda clouded leopard, taken with respect to an increasing gradient of elevation, which has been normalised with standard scaling to have zero mean and unit variance.

We see that a correlation emerges as we smooth the graph, from an initial collection of binary values to a smoothly increasing graph. In the first two cases, noise obscures any clear pattern, but in the last case, we have potentially smoothed away all information aside from the most basic pattern

of a positive correlation. The fourth and fifth graph contain information more coherent than the first, and more rich than the last, with the apparent emergence of two peaks as elevation increases. Figure 2 gives another example, using the same empirical occurrence data for the clouded leopard, but now looking with respect to human footprint:

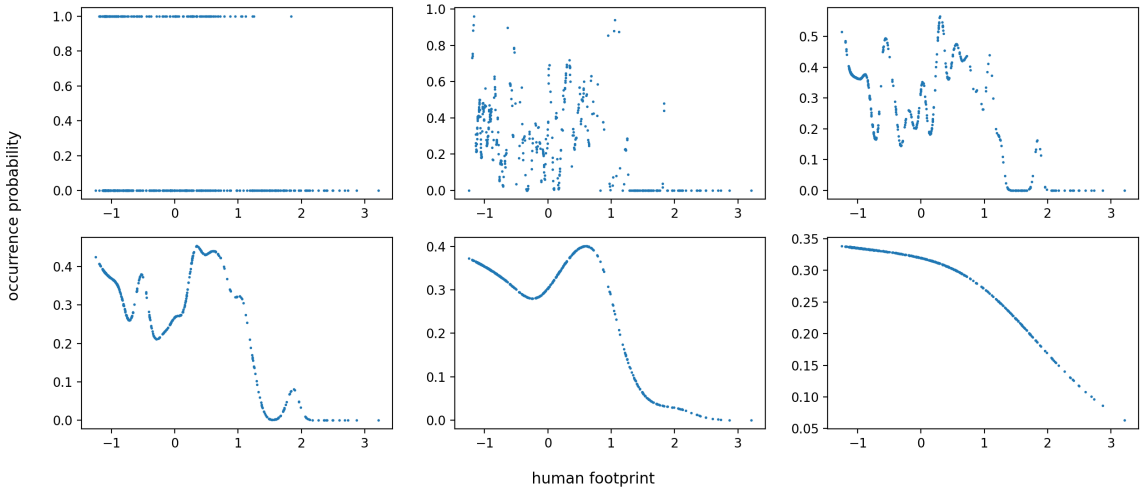

Figure 2: Another examples of the effects of smoothing. The data are again occurrences of the Sunda clouded leopard, now with respect to increasing gradient of human footprint. The smoothing function has been applied to the latter five graphs with the same values of σ, namely $\sigma = 0.01, 0.05, 0.1, 0.3, 1$ respectively.

As one could expect, with extensive smoothing we see a negative correlation between occurrences of the Sunda clouded leopard and increasing human footprint. But, beyond this basic inference, we can say little else - how could we explain the clear peak in the fifth graph, in which the their occurrence increases as human footprint increases along a portion of this environmental gradient? Similarly, how shall we understand the double peak seen in the fifth elevation graph?

Now comes the key point. What we are observing here is the correlation of the species occurrence with this particular environmental variable, across different sites in geographical space. But across these different sites, other environmental variables will be varying too; this means we are not observing their response to this variable independently of the other variables to which they may be also responding. Indeed, rather than the Sunda clouded leopards favouring intermediate levels of human footprint (as one may think Figure 2 suggests), we may suspect that the *covariance* of the environmental variables is the cause of this pattern; that, for example, in regions of Borneo for which human footprint increases from low to intermediate, another environmental variable (such as elevation, or

forest cover) is varying simultaneously, and their positive response to this environmental variable outweighs their negative response to increasing human footprint. Crucially, even if the underlying response between a species and *every* environmental gradient is linear, the observed correlations may exhibit the nonlinear and multimodal shapes seen in Figures 1 and 2, due to the complexity not of the species response but rather the nonlinear covariance of environmental variables; see Section A.1 for details and examples of this phenomenon. So, to better observe the underlying response between species and environment, we need to understand the shape of how the environmental variables themselves vary together, which leads us to consider the environmental manifold.

2.2 The environmental manifold

The environmental manifold, described in detail below, is the shape of the data cloud of sampled environmental variables. It encodes precisely how this set of environmental variables co-vary, and in practice is obtained from empirical GIS data. The observed correlations above provide one motivation for its importance, the implications of which for ecological theory and practice are revisited in Sections 4 and A.1. A second reason is its fundamental relevance to the Hutchinsonian niche, which was mentioned in Section 1 and will be revisited in Section 4. Finally, it is a concept which opens avenues for new mathematical insights into ecological data, affording a geometric viewpoint hitherto unavailable in ecology, which will be discussed in Section B. Let us now see how to construct the environmental manifold.

Constructing the environmental manifold

We choose to illustrate the environmental manifold using three environmental variables across the island of Borneo: elevation, forest cover and human footprint, given their demonstrated importance as a predominant influence on biodiversity (D. W. Macdonald et al. 2020) generally, and particularly for the wild felid species (Hearn et al. 2018) in this region (Figure 3). Elevation is derived from the SRTM data set (Jarvis et al. 2008). Forest cover is a continuous variable ranging from 0 (no tree cover) to 100 (full canopy closure) and is taken from the NASA global tree cover data set (Hansen et al. 2013). Human footprint is obtained from the global human footprint data set (WCS and CIESIN 2005). In this analysis we evaluate these environmental variables at different geographical scales, following Hearn et al. 2018 and Kaszta et al. 2019, who showed that a species response to these variables (illustrated by the Sunda clouded leopard) is scale dependent, and is strongest and

clearest when measured relative to the focal mean of these variables at a given radius. These variables were evaluated at 90m for elevation, 30m for forest cover and 1,000m for human footprint. For our analysis, all variables were re-sampled to a common resolution of 250m, following D. W. Macdonald et al. 2020.

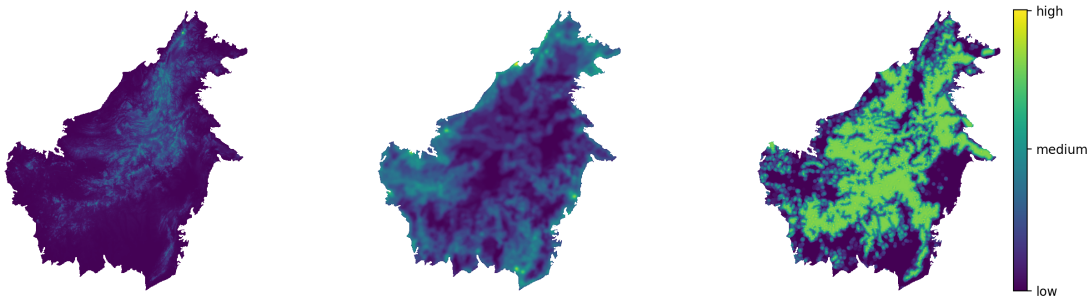

Figure 3: Three GIS layers of Borneo, which are elevation, human footprint and forest cover, respectively. Yellow corresponds to higher values and blue to lower values.

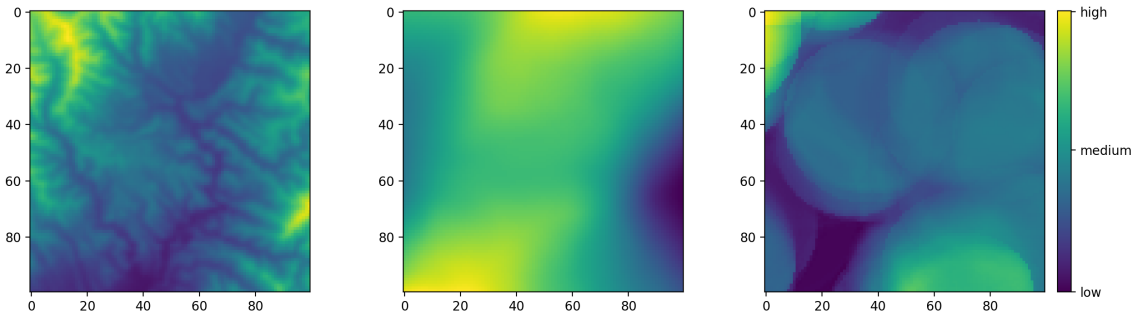

Figure 4: A 100×100 pixel portion of the three rasters in Figure 3, showing part of Sabah, Borneo.

Figure 4 shows a 100×100 pixel grid of obtained from the above three rasters, all from the same region in Borneo. Since we are viewing each environmental variable in a different plot, it is difficult to see how they co-vary: namely, how do their values change simultaneously as we move across this portion of Borneo? Recall from Section 2.1 that the knowledge of this simultaneous variation of environmental variables is crucial to our understanding of a species-environment relationship. Now, the exact nature of this covariance becomes apparent if we move from geographical space to environmental space (referred to by G. Evelyn Hutchinson in Hutchinson 1957 as 'niche space', and similar to the 'gridded environmental space' in Broennimann et al. 2012). This is simply the space where the axes are given by the different environmental variables, and we label this space as $\mathbb{E}$.

Moving from geographical space to environmental space means that each pixel in our GIS layer of Borneo is mapped into three-dimensional space (since we are considering here the three environmental variables mentioned above), to the point whose coordinates are given by the environmental variable values at that pixel. For example, if a pixel has elevation α_1, with human footprint value α_2 and forest cover value α_3, then this pixel of the Borneo layer gets sent to the point $(\alpha_1, \alpha_2, \alpha_3)$ in our three-dimensional environmental space $\mathbb{E}$. We do this for all pixels in any chosen portion of Borneo, and we call the resulting shape the environmental manifold, labeling this object by M. In Figure 5, we see how our 100×100 portion of Borneo twists and bends in a highly nonlinear manner, demonstrating the geometrically complex covariance of these environmental variables. In Figure 6, we see the graph of this same portion of Borneo in a two-dimensional environmental space, where in each case the two axes are given by two of the three environmental variables considered above.

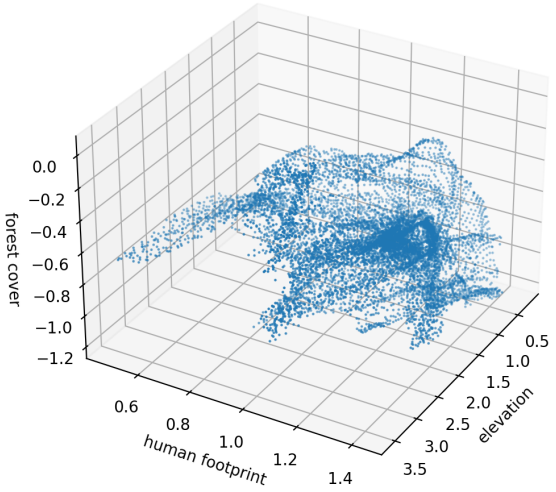

Figure 5: Visualising the environmental manifold. This is the result of mapping the pixels from the 100×100 pixel portions of the GIS layers into three-dimensional environmental space, with axes given by the three environmental variables (elevation, human footprint and forest cover). The three environmental variables have been normalised using standard scaling to have zero mean and unit variance. Intuitively, we can think of this mapping from geographical space into environmental space as folding and contorting a piece of paper, where the paper is the flat GIS layer of Borneo in geographical space (parameterised by latitude and longitude).

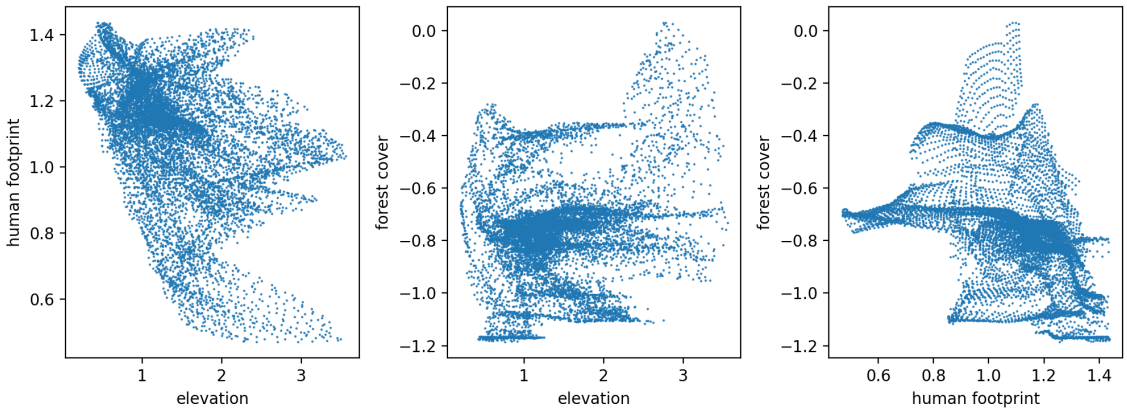

Figure 6: The same 10,000 points have now been plotted in two-dimensional environmental space, with axes given by the three different possible pairings of the three environmental variables above. Visually, this is equivalent to flattening the shape in Figure 5 onto the floor or the two walls of the axes seen in Figure 5.

Despite its seeming complexity, we are actually working with a shape which in some sense we understand very well, since we know it is in fact sampled from a portion of two-dimensional geographical space, which means that we know which points on the environmental manifold neighbour each other in geographical space. Thus, we can create a linear approximation of this surface from these points, and hence obtain a clearer representation of its shape. In Figure 7, we see this for three different 20×20-pixel portions of the above environmental manifold, where we have 'joined up the dots' by using the knowledge from geographical space of which points are contiguous with each other. In this paper, we stick to using three environmental variables for the sake of visualisation. In the same manner, however, we may construct the environmental manifold using as many variables as desired, as discussed in Appendix B.

2.3 Comparison of smoothing with GLM and random forest

Analysis methodology

With the smoothing function and environmental manifold now developed, we study the accuracy of the GLM, random forest and smoothing methods in recovering simulated species-environment relationships, as we vary both the relationship z and also the geometry of a simulated environmental manifold M. These simulated environmental manifolds will be constituted of 10,000 points in environmental space, just like the empirical environmental manifold obtained from the GIS layers in

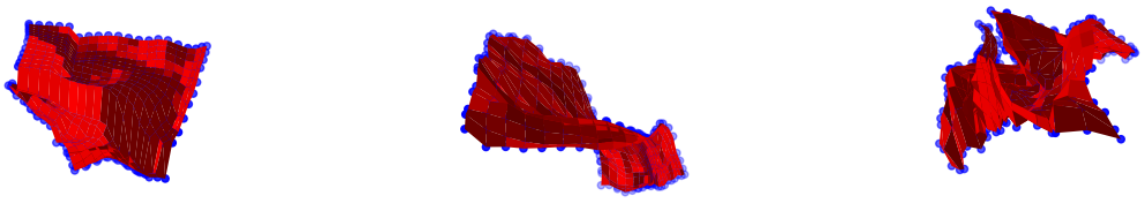

Figure 7: Using the knowledge from the raster layers of which points are adjacent to each other, we obtain a linear approximation to any chosen portion of Borneo's surface when mapped into environmental space.

Section 2.2. We will then describe how a simulated data set is created from a chosen relationship z and environmental manifold M. The metric of accuracy will be calculated from the difference between the true simulated values and values predicted by each of the three models: it will be given by the Pearson correlation coefficient and mean squared error between the simulated and predicted values at each of the 10,000 points on M for the GLM and smoothing function; for random forest, at each of the 1,000 points on M obtained from a 9 to 1 split between train and test data.

We will consider the linear relationship $z_1 = X + Y - Z - 1$ and the nonlinear relationship $z_2 = (0.3X)^3 + YZ + (0.2Y)^4 - 1$ for our simulated species-environment relationships, where X, Y, Z are variables in our simulation which play the role of environmental variables, and which are scaled to have zero mean and unit variance. We chose these two relationships since they gave a distribution of binary values on the environmental manifolds which did not lead to an extreme of either very few presences or absences, and we used a nonlinear relationship of this form so as to include a variety of possible nonlinearities (such as cross terms and different powers); other than these two aspects, the particular coefficients of z_1 and z_2, and the form of z_2, could just as well have been chosen differently. For the shape of the environmental manifold, we consider four cases: a collection of random points (sampled from a uniform distribution), a plane, a sphere, and the portion of the empirical environmental manifold seen in Figure 5. These are illustrated in Figure 8. We chose the uniformly random sample since it represents a null model for the environmental manifold, and the plane and sphere in order to have a variety of different possible geometries for our study, noting that aspects of all three shapes may be realised in empirical environmental data. In this simulation, we

use the term 'empirical environmental manifold' for the shape arising from the empirical GIS layers of geographical space in Section 2.2, to distinguish it from the simulated shapes in our analysis. See the end of this section for a summary of this methodology.

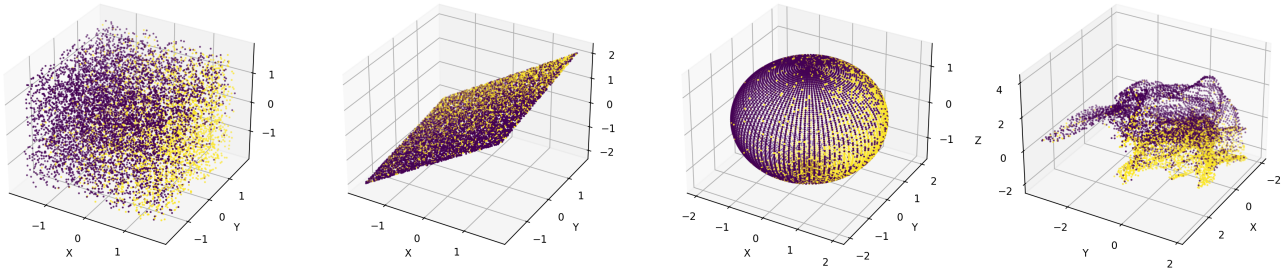

Figure 8: The four different shapes for the simulated environmental manifold, constituted of 10,000 points: respectively, a uniform random selection, a plane, a sphere, and the empirical environmental manifold itself. The colouration is given by presence (yellow) and absence (blue), given by the linear simulated relationship z_1. See Section 2.3 for how these binary values are obtained.

After this, we again look at the accuracy of the GLM, random forest and smoothing methods with the same linear and nonlinear species response (given by z_1 and z_2 respectively), but now on two different subsets of points from the empirical environmental manifold, rather than using the fully sampled empirical environmental manifold as before. This corresponds more closely to the situation which arises in reality from camera trap data. Our first subset is a sample of 5000 points selected randomly across Borneo. In the second, we use the actual 484 locations where camera traps were placed in Sabah to collect species occurrence data in Hearn et al. 2018.

Creating a simulated data set

To create the simulated data sets for our analyses, we consider three environmental variables X, Y, Z (which means that we are working in three-dimensional environmental space). We first create out simulated environmental manifold M by sampling 10,000 points from the three-dimensional environmental space, to obtain one of the shapes mentioned above. Then, we choose our species-environment relationship z, which is given by a function of these three variables. In our analysis, z is a polynomial in X, Y and Z, being one of the two functions z_1 or z_2 above.

Now, to obtain a resulting probability of occurrence from z, taking values between 0 and 1 (where values closer to 1 indicate a higher likelihood of our simulated species occurring, and values closer to

0 a lower likelihood) we use the logistic transform $\phi = e^z/(1+e^z)$. This transform produces a value $\phi(p)$ between 0 and 1 for each point p of the 10,000 points on our environmental manifold M, which corresponds to the probability of the simulated species occurring at each location, as a function of the environmental variables at that location and the specified species-environment relationship z.

With this probability of species occurrence, we now generate a simulated presence-absence data set which reflects those probabilities by, at each point p on M, sampling a random number $r(p)$ uniformly between 0 and 1, and calculating the difference $\phi(p) - r(p)$, giving a value between -1 and 1. If the resulting value is negative, we relabel it as 0, and else relabel it as 1. This is a stochastic simulation of presences and absences that reflect the stipulated probability of occurrence at each location as a function of z. Our simulated data set is the resulting collection of binary values, 0 and 1, at each point p on the environmental manifold. Figure 9 gives an illustration of the environmental manifold, first coloured by the value of ϕ on and then by the resulting binary value.

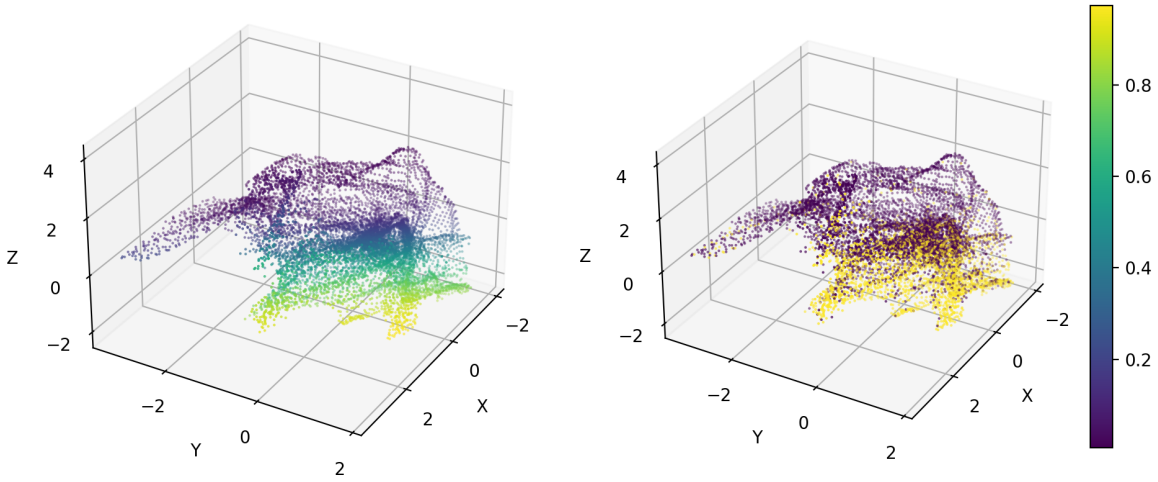

Figure 9: Colouration of the environmental manifold: on the left, by the probability surface ϕ, and on the right, by the resulting stochastic presence-absence value. On the left, yellow and blue correspond to a higher chance of presence and absence, respectively; on the right, to the resulting binary presence and absence values.

Model setup

We use a logit link function for the GLM, due to the binary nature of the simulated data. As mentioned in the introduction, in configuring the GLM we are also required to choose a priori the functional shape (e.g. linear, squared, exponential) used to predict the species-environment relationship. We choose the linear predictor for our analysis due to its use being by far the most prevalent in ecological studies (McGarigal et al. 2016). Since the GLM is a parametric model, whose parameters are the slope coefficients defining the predicted linear regression plane, we will also investigate its ability to recover these coefficients. The GLM and random forest were implemented in Python 3.7, the former with the `statsmodels` package and the latter with `sklearn`, versions 0.12.1 and 0.24.1 respectively (Seabold and Perktold 2010; Pedregosa et al. 2011).

For configuring the random forest and smoothing methods, there is the question of hyperparameters (also known as model parameters), which are the parameters chosen in the setup of a model and which determine how the model will run. With our study, we found that the only hyperparameter which, when changed from its default value, was found to increase the accuracy of random forest was the maximum depth of a tree (which is the number of splits in each decision tree). This means that we also used a 9 to 1 split between the train and test data, which is the default value in the `sklearn` package. For the smoothing function there is only one hyperparameter involved, namely the variance (or, bandwidth) σ of the Gaussian function K_σ defined in Section 2.1. Thus the question of tuning hyperparameters reduces to one in each case: for random forest, a maximum depth of 7 and 8 was most accurate for z_1 and z_2, respectively; for smoothing, the optimal value for σ depended on both relationship and environmental manifold, and varied between 0.05 and 1.

All analyses were performed in Python 3.7 using Jupyter notebook. Since the simulated occurrence data sets have a random component, we ran the simulation 100 times for each case of relationship, and environmental manifold or sample type, and have taken the mean average of the resulting correlation and error.

Summary of methodology

In summary, we first create our simulated environmental manifold M by sampling 10,000 points from environmental space, with axes X, Y, Z. We then define our species response z as a function of X, Y, Z, and apply the logistic transform $\phi = e^z/(1 + e^z)$ to obtain a probability of occurrence for each of the 10,000 points on M. From this we stochastically obtain the binary presence-absence data

on M as described in 2.3, which comprises our simulated presence-absence species occurrence data set. We apply the GLM, random forest and smoothing methods to this binary data, and determine their accuracy by calculating the resulting Pearson correlation coefficient and mean squared error from the difference between the true simulated values and predicted values at each point. We further evaluate the accuracy of all three methods on two smaller samples of environmental space, and lastly we determine the predictive ability of the GLM to recover the slope coefficients of the regression plane defining the linear species relationship $z_1 = X + Y - Z - 1$.

3 Results

Simulation on the fully sampled environmental manifold

We present in Table 1 the results of our study on the fully sampled environmental manifold. For each shape and method, we see its accuracy in recovering the simulated species occurrences, first for the linear and then for the nonlinear relationship. In each case, the first of the two numbers gives the Pearson correlation coefficient between the predicted values and the true simulated values, and the second gives the mean squared error between these values, to four decimal places. A high correlation means that the predicted values are, on average, roughly proportional to the distribution of the true values; a low error means that there is little variance and bias between the individual predicted and true values. Thus, by these two metrics, the overall accuracy is higher when the first number is closer to 1, and when the second is closer to 0.

Simulation on sparser samples of environmental gradients

In empirical research, samples are taken at a subset of locations, rather than exhaustively (as was the case with the simulated data sets on the environmental manifold in Section 2.3). In this section, therefore, we present the results for the simulated relationships in the two sampling examples described above. This gives an illustration of the difference in inference and effectiveness of the three methods between 'ideal' sampling, consisting of a large random sample, and 'actual' sampling, which is typically smaller and nonrandom. These locations are plotted in environmental space in Figure 10, and the results of this analysis are shown in Table 2.

Relationship	Manifold	GLM	Random forest	Smoothing
	Random	(0.9997,0.0001)	(0.9837,0.0031)	(0.9948,0.0014)
Linear	Plane	(0.9995,0.0001)	(0.9678,0.0014)	(0.9936,0.0006)
$z_1 = X + Y - Z - 1$	Sphere	(0.9998,0.0001)	(0.9877,0.0024)	(0.9958,0.0008)
	Empirical	(0.9997,0.0001)	(0.9774,0.0030)	(0.9881,0.0017)
	Random	(0.0399,0.0332)	(0.9557,0.0038)	(0.9899,0.0025)
Nonlinear	Plane	(0.0730,0.0099)	(0.9552,0.0009)	(0.9870,0.0005)
$z_2 = (0.3X)^3 + YZ + (0.2Y)^4 - 1$	Sphere	(0.0776,0.0184)	(0.9241,0.0042)	(0.9812,0.0020)
	Empirical	(0.1747,0.0124)	(0.8984,0.0026)	(0.9642,0.0011)

Table 1: Pearson correlation coefficient and mean squared error for each relationship, environmental manifold geometry and method, shown to four decimal places.

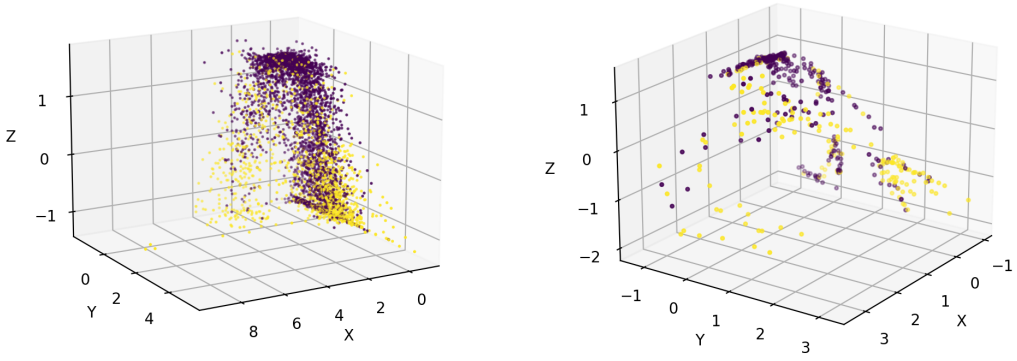

Figure 10: Two sets of locations, plotted separately in environmental space, with axes given by elevation, human footprint and forest cover. On the left, a 5000-point random sample of locations from across Borneo; on the right, the actual 484 locations of the camera traps in Sabah, as detailed in Hearn et al. 2018. As before, the axes have been normalised with standard scaling.

Relationship	Sample	GLM	Random forest	Smoothing
Linear	5000	(0.9995,0.0001)	(0.9760,0.0037)	(0.9875,0.0020)
$z_1 = X + Y - Z - 1$	484	(0.9946,0.0015)	(0.9513,0.0097)	(0.9797,0.0081)
Nonlinear	5000	(0.5292,0.0143)	(0.9174,0.0034)	(0.9610,0.0019)
$z_2 = (0.3X)^3 + YZ + (0.2Y)^4 - 1$	484	(0.5839,0.0112)	(0.7921,0.0067)	(0.8887,0.0040)

Table 2: Pearson correlation coefficient and mean squared error for each relationship, sample type and method, taken to four decimal places.

Sensitivity of the GLM to environmental covariance

Suppose the GLM has accurately recovered the true simulated values on the environmental manifold, as was the case with the linear relationship $z_1 = X + Y - Z - 1$ above. Since it is a parametric model, we may further study its accuracy in recovering the parameters $(\beta_0, \beta_1, \beta_2, \beta_3)$ defining the species-environment relationship $z = \beta_1 X + \beta_2 Y + \beta_3 Z + \beta_0$. So, in our case, can the GLM recover the coefficients $(1, 1, -1, -1)$ of the terms in the relationships $z_1 = X + Y - Z - 1$? Surprisingly, the accuracy in recovering the simulated values of the linear relationship (seen in Table 1) need not reflect its accuracy in recovering these coefficients.

When the environmental manifold had a nonlinear shape (as was the case with the random points, sphere, and empirical environmental manifold), the GLM recovered the parameters β_i with an average error of approximately 0.02. However, when the environmental manifold had the shape of a plane (which, in this case, was given by $4X+Y-2Z-5 = 0$), the GLM predicted $(-0.063, 0.705, 0.113, -1.013)$ for the coefficients, which, for β_1, β_2 and β_3 is extremely inaccurate (Table 3). This error still occurred, sometimes to a greater magnitude, when varying the coefficients β_i of the linear relationship z_1, rotating and translating the environmental manifold plane, or adding small degrees of noise or nonlinearities orthogonal to the plane. In the language of statistical linear algebra, this problem arises because the design matrix (which is the matrix of values of the predictor variable X, Y, Z) does not have sufficient rank to provide a unique value for the slope coefficients of the regression plane. This is equivalent to a geometric argument involving the environmental manifold, which is left to Section B.1.

Manifold	Predicted coefficients
Random	(0.975,0.998,-1.005,-1.037)
Plane	(-0.063,0.705,0.113,-1.013)
Sphere	(0.994,1.040,-1.040,-1.040)
Empirical	(0.994,0.983,-0.965,-0.988)

Table 3: The GLM's prediction for the slope coefficients $(1, 1, -1, -1)$ of the linear relationship $z_1 = X + Y - Z - 1$, with each of the four shapes from Section 2.3.

4 Discussion

Describing the Hutchinsonian realised niche and understanding the scale dependence in the response of a species to their environment are two central questions in modern ecology, both of which have received substantial attention in recent years (Holt 2020; Blonder et al. 2014; Chandler and Hepinstall-Cymerman 2016). In this work, we have illustrated how the tools of smoothing in environmental space, together with the hitherto unexplored environmental manifold, are powerful tools for studying these two realms. As explained throughout this paper, and treated in particular detail in Section A.1, the environmental manifold gives us a description of the realisable niche (that is, the subset of niche space in which the realised niche is constrained to occur); moreover, due to its multi-dimensional nature, the environmental manifold is a concept which can lead to surprising insights into ecological data by allowing us to consider and observe the nonlinear interactions of several environmental variables simultaneously. Furthermore, we have demonstrated that smoothing in environmental space performs very well in comparison with widely-used models such as the GLM, and with state-of-the-art machine learning tools such as random forest, with an ability to handle scale-dependence and nonlinearities in both the species response and the environmental manifold.

With the ability to stipulate known species-environment relationships, the tools of simulation give us the ability to quantify and compare the performance of the different statistical methods widely used in modeling species-environment relationships. Recent work has used simulation techniques similar to this study, also for the purpose of evaluating the efficacy of various methods to infer predictors of species occurrence (Atzeni et al. 2020; Chiaverini et al. 2021). This paper develops the simulation framework further, by investigating both linear and nonlinear relationships, and also different geometries of the environmental manifold. In our analysis, we used these techniques to

compare the predictive ability of the smoothing function with that of the GLM and random forest. Following Samuel A Cushman and Wasserman 2018 and Samuel A Cushman, E. A. Macdonald, et al. 2017, we found that the random forest model outperformed GLM when there were nonlinear and interactive terms in the species-environment relationship, and also found, for the first time, that random forest machine learning is robust to the shape of the environmental manifold. Furthermore, we found the smoothing function to be more accurate than the random forest in every setting, particularly in situations with greater sampling bias. Finally, we saw the GLM to be unable to recover the slope coefficients of the linear relationship with certain shapes of the environmental manifold.

Effects of functional complexity

To our knowledge, this is the first exploration of how functional complexity (which is determined by, for example, the number of variables and nonlinearities in the predicted response) interacts with environmental covariance in affecting the performance of different models to infer species-environment relationships. We found that when the relationships were linear, all three methods performed well in correctly predicting species occurrence. Furthermore, when the environmental manifold was of a highly nonlinear shape in our study, the GLM accurately recovered the slope coefficients β_i of the regression plane.

Nonlinearity of the relationship strongly affected the performance of the GLM, greatly reducing its ability to predict species occurrence and estimate slope coefficients. This sensitivity of the GLM to complex responses stands in contrast with recent studies such as Ash et al. 2021. On the other hand, random forest and smoothing performed very well with the nonlinear relationship - almost as well as they did on the linear relationship - showing these two methods to be capable of recovering complex nonlinear species responses. Moreover, the smoothing function proved more accurate than random forest, regardless of functional complexity.

Effects of geometric complexity

With the metrics of accuracy used in this study, the geometric complexity of the environmental manifold generally did not have a great influence on the performance of any of the three methods. The exception is the case of a planar environmental manifold affecting the ability of the GLM to recover the coefficients β_i. Upon further investigation, we found this error with the GLM to still

occur in other similar instances, such as when degrees of noise or nonlinearities were added orthogonal to the planar environmental manifold. Although a perfect planar geometry is not to be expected in empirical environmental data, this means that if there are portions of approximate constancy among any linear combination of the environmental variables, the GLM may be unable to accurately recover the parameters defining even a linear species response; this may very well arise if any of the environmental gradients are not well-sampled. From this, we conclude that the random forest and smoothing function are highly robust to the geometric complexity of the environmental manifold, but the ability of the GLM to predict regression coefficients can be strongly affected by certain types of environmental covariance.

Effects of sampling bias

With a linear species-environment relationship, we found the GLM to be the most accurate method for determining the species response, for all types of sampling; however, the strong sensitivity of the GLM to functional complexity remains, regardless of sampling effort. As with the fully sampled environmental manifold, in the case of the sparser empirical data (which reflects the sampling bias arising in practice) we found the smoothing function to be more robust than random forest, particularly with the more complex species response. We also see that the larger 5000-point random sample gives rise to greater model performance than with the smaller, opportunistic sample; all models, however, were most accurate with the exhaustive sampling represented by the environmental manifold. This shows that, while the smoothing function is not strongly affected by sampling bias, its accuracy cannot be considered to be independent of sampling, in contrast with what is claimed in regards to kernel smoothing methods in Broennimann et al. 2012.

Further developments for the environmental manifold

As motivated in Section 2.1, having a precise understanding of the way in which environmental gradients vary together is of central importance when studying species-environment relationships. Furthermore, as demonstrated in detail in Section A.1, observed nonlinearities in a species response may arise solely because of the nonlinear shape of the environmental manifold, even when all components of the underlying relationship are linear. This is of major importance for practitioners involved in conservation and management decisions, since predictions about species occurrences may be highly inaccurate when complex environmental covariance is not accounted for. With the increasing avail-

ability of highly accurate and fine-scale GIS data, we are now able to actually visualise, quantify and work with the environmental manifold as a tool in ecological analyses, which gives us access to a better understanding of the species-environment relationships we may observe from empirical data. In this paper we used only three environmental variables when working with the environmental manifold, for ease of visualisation and exposition, but the concept and insights of the environmental manifold are equally applicable for any number of environmental variables, suggesting the environmental manifold to be a powerful and practical new tool for ecological analyses with large data sets. The development of the environmental manifold as such a tool for policy and conservation should form the basis for future work, and should be configured to handle cloud computing and large volumes of data, in line with the latest developments in ecological informatics (Hoogen et al. 2019; Rey and Huettmann 2020).

Another salient aspect of the environmental manifold is its relevance to the Hutchinsonian realised niche, the modeling of which is of fundamental importance in community ecology. With the tools developed in this paper, we are now afforded the precise description of the subset of niche space which is geographically realised; namely, the environmental manifold is the *realisable* Hutchinsonian niche. This means that, even if the fundamental niche is isotropic and simply described, traditional niche descriptions will fail if the realisable niche (in other words, the environmental manifold) is complex and nonlinear. As a result, without the environmental manifold, we can only tell half the story in quantifying a species-environment relationship. For example, we may find from empirical data that, in a region of three-dimensional niche space, the occurrence probability of our species is close to zero. This may reflect an intolerance of the species to this set of environmental conditions - but equally, it may represent a region of this niche space which is not geographically realised. The knowledge to discern between these two very different realities is provided by the environmental manifold, which has heretofore been absent in attempts to quantify the ecological niche (such as Broennimann et al. 2012; Swanson et al. 2015; Holt 2009; Blonder et al. 2014).

We have shown in this paper that random forest and smoothing are capable of handling these nonlinearities which arise in the realised niche, in contrast with the GLM which assumes a symmetry and isotropy of the underlying predictor space. With the smoothing function and the environmental manifold, deeper insight into the ecological niche of a species is possible, and we hope this will form the basis for exciting future work. The possibility of delving into the mysteries of the environmental manifold and its connections with other ecological questions lay beyond the scope of this study. However, the environmental manifold is a concept which applies to contexts much wider than that

studied in this paper; we believe that the combination of its underlying ubiquity in questions pertaining to species-environment relationships, together with its mathematical richness and depth, makes it a tantalising subject of further study - both as an accomplice in other ecological explorations, and also as a matter in its own right.

Conclusion

The simulation framework is useful in exploring the effectiveness of different modeling methods, and in studying the influences of factors such as sampling design, sample size, complexity of response and environmental covariance on the performance of such methods. Our results show that random forest and smoothing are robust to the complexity of both species response and environmental covariance, with the smoothing function proving more accurate in every case; in contrast, the GLM is strongly affected on the functional complexity of the species response. Furthermore, the GLM can also be highly impacted by certain shapes of the environmental manifold. This suggests that, given the expected nonlinear and interactive relationships in empirical ecological data, methods like random forest and smoothing which are equipped to handle complex and a priori unknown responses are likely to be preferable and more capable than parametric models like the GLM.

Moreover, this study shows that applying a kernel smoothing function to data in environmental space is a promising and powerful way to explore and observe the patterns of species-environment relationships, with particular relevance to scale dependence in environmental space, which heretofore has not been well explored. The smoothing function provides a conceptually simple and mathematically transparent approach to recover responses of any complexity, in a manner which fits well into an ecological setting by inherently accounting for sampling bias. In addition to the smoothing function, we have introduced the environmental manifold, a new and mathematically rich area of exploration in ecology with deep implications for ecological practice and theory, particularly in regards to understanding species distributions and modeling the Hutchinsonian niche.

Finally, we would like to contextualise this work within the broader realm of scientific discourse in relation to the natural world, by acknowledging the vast and growing body of literature which discusses and explores, among other things, the importance of language and mythos employed in quantitative studies such as in this paper (Abram 2010; Ingold 2000; Cronon 1997; Kimmerer 2013). When discussing the species-environment relationship in this work, it has been regarding the response of a species with respect to the gradients of environmental variables, both of which are gathered from

empirical data and studied through a quantifiable viewpoint. Such a viewpoint, and its resulting numerical analyses, have tremendous value in many areas of ecology and the life sciences. But we would like to emphasise that the work in this paper is not motivated by an effort to ultimately explain the myriad depths to the living, breathing relationship between a species and the more-than-human earth through purely quantitative means (Abram 1996; Snyder 1990). Instead, we have sought to develop and test more powerful methods to aid us in the observation and prediction of the quantitative patterns arising from these relationships. We hope that the findings of our simulation study, together with the concepts of the smoothing function and the environmental manifold, provide helpful tools for future explorations in ecology.

A Details on smoothing in environmental space

A.1 Nonlinear correlations from linear responses

We present here some visual examples of the phenomenon discussed in Section 2.1, which explained why, even after accounting for sampling bias with the smoothing function, the observed correlations of the distribution of species occurrence along environmental gradients may very much not reflect the species underlying response, precisely due to the nonlinearities of the environmental manifold. In particular, highly nonlinear and multimodal correlations may spuriously arise even from very simple species-environment relationships, such as when the response to some or all environmental variables is linear.

To illustrate this, on the same 484 empirical camera trap locations mentioned in Section 2.3, we used the techniques of Section 2.3 to simulate four different species-environment relationships, as a function of the three environmental variables (elevation, human footprint and forest cover) discussed in Section 2.2, which we label here as X, Y, Z respectively. These four functions were $z_1 = X$, $z_2 = X + 5Y$, $z_3 = X + 3Z$, and $z_4 = X + 2Y^2 - 3Z^2$, which are all linearly increasing with respect to the elevation variable X. We may thus expect that, as elevation increases, the occurrence probability of our simulated species also increases. However, in Figure A.1 we see that the correlations which emerge after accounting for sampling bias show this only to be roughly the case for $z_1 = X$, which is a function solely of elevation; when the species responds also to human footprint and forest cover, such as in z_2, z_3 and z_4, we may observe correlations which do not at all reflect the linearity (or even monotonicity) of its response to elevation.

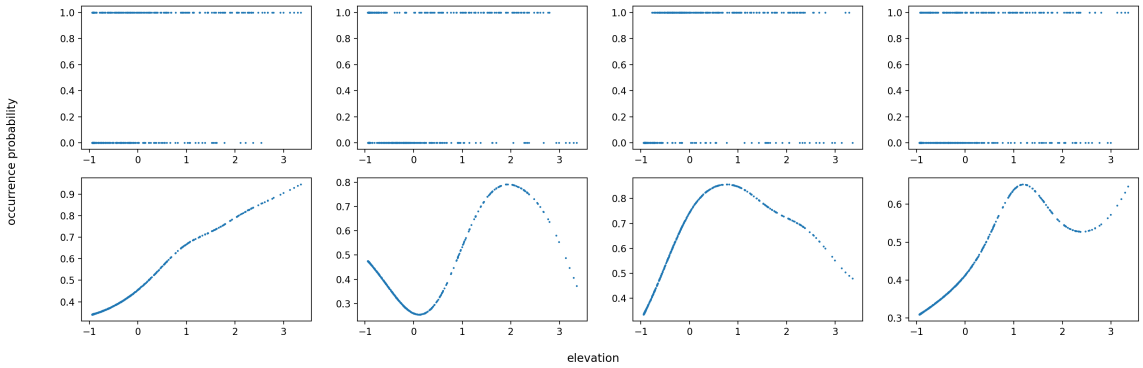

Figure A.1: Nonlinear correlations arising spuriously from linear relationships, due to nonlinear environmental covariance. From left to right, we see the emergent correlations for the four species-environment relationships z_1, z_2, z_3 and z_4 respectively, with elevation on the horizontal axis. For each relationship, the upper graph displays the binary presence-absence data, and the lower graph shows the occurrence probability which emerges when applying the smoothing function to account for sampling bias ($\sigma = 0.5$). Though all relationships are linearly increasing with respect to elevation, the only observed correlation which approximately reflects this is that of z_1, which is a function of elevation only.

These nonlinearities arise from the linear response to elevation precisely because of the complexity with which the three environmental variables co-vary across these locations, as discussed in Sections 2.1 and 2.2. Indeed, if we instead sample our locations uniformly randomly across environmental space, Figure A.2 shows that when the environmental covariance is isotropic (which reflect assumptions regarding the Hutchinsonian niche), then all of the smoothed graphs in Figure A.1 would have instead displayed a linearly increasing shape. We can thus conclude that the environmental manifold, which gives us the precise information of how our environmental variables co-vary in geographical space, plays a fundamental role in understanding the correlations which emerge from the response of a species to its environment. This is a finding of great importance for both ecological theory and practice: in terms of theory, this shows that one cannot effectively study niche structures without considering the environmental manifold shape; and in practice, it is important that policy and conservation decisions are to be based on predictions of species distributions which accurately reflect their underlying response to the environment.

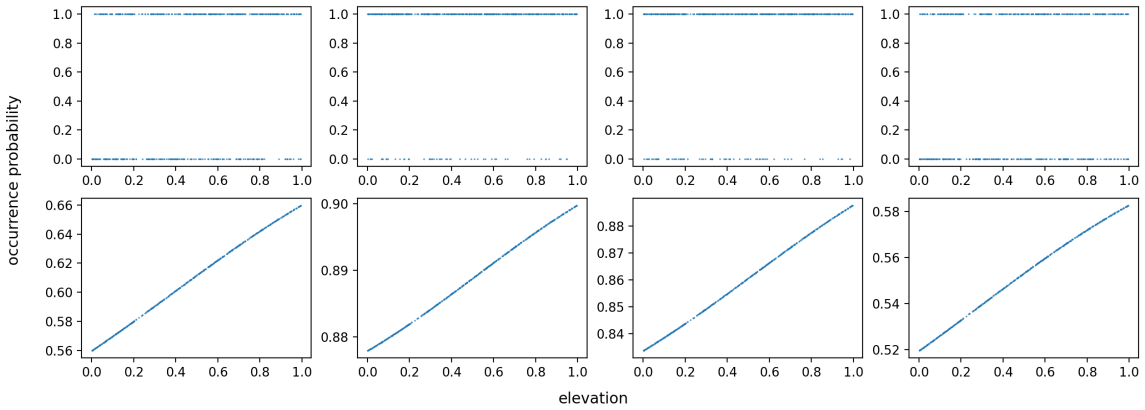

Figure A.2: Emergent correlations for the same relationships z_1, z_2, z_3, z_4 as before, again with elevation on the horizontal axis, but now with the 484 data points instead sampled uniformly randomly from three-dimensional environmental space. These smoothed graphs show the true linear nature of the underlying response to elevation, which now becomes visible precisely because of the uniformity of the environmental manifold.

A.2 Sample code for the smoothing function

First, let us recall the formula for the smoothing function, as defined in Section 2.1:

$$S(x) = \frac{\sum K_\sigma(x, x') \cdot V(x')}{\sum K_\sigma(x, x')}, \quad \text{where} \quad K_\sigma(x, x') = e^{-\frac{(x-x')^2}{2\sigma^2}}$$

As noted before, the sum is taken over all points x' in the data set, but if the data set is much larger than those used in this analysis, then the kernel can be truncated to improve run times. We now present the sample code for the smoothing function used in our simulation, implemented in Python 3.7:

```
import numpy as np
from scipy.spatial import distance


def S(x, V, sigma):
    D = distance.cdist(x, x, 'sqeuclidean')
    K = np.exp(- D / (2 * sigma ** 2))
    return np.sum(K * V, axis = 1) / np.sum(K, axis = 1)
```

The input data consist of: `x`, an array consisting of the coordinates of the points in environmental

space, with each row giving the coordinates of one point; `V`, a vector of binary presence-absence values at each point; `sigma`, the smoothing parameter for the Gaussian kernel K_σ defined in Section 2.1. The smoothing function, `S`, first computes the matrix of squared distances `D` for the occurrence points in environmental space (whose coordinates are given by `x`), then calculates the kernel `K`, and finally returns the vector of smoothed `V`-values. For example, in our simulation on the environmental manifold in three-dimensional space, `x` is a 10,000×3 numpy array and `V` is a vector of length 10,000.

B Mathematical aspects

In this section, we touch on some of the more mathematical aspects of the environmental manifold. We begin by noting that our environmental space $\mathbb{E}$ can be thought of as the m-dimensional space of real numbers $\mathbb{R}^m$, where m is the number of environmental variables used for the axes in $\mathbb{E}$. Consider the construction of the environmental manifold from Section 2.2, which involves the three environmental variables of elevation, human footprint and forest cover. We label them here as e_1, e_2, e_3 respectively. Mathematically speaking, in constructing the environmental manifold, we think of the GIS polygon representing Borneo as a subset U of two-dimensional space $\mathbb{R}^2$, parameterised by latitude and longitude. We then look at the image $M = f(U)$ of the map $f : U \to \mathbb{R}^3$ defined by $f(x, y) = (e_1(x, y), e_2(x, y), e_3(x, y))$ for (x, y) in U. In practice, the function f is defined on discrete data, namely the pixels of the GIS polygon. However, it can be thought of as continuous and piecewise-linear: it takes the flat GIS polygon of Borneo and maps it to a piecewise-linear surface, by 'joining up the dots' of the pixels of the GIS polygon in environmental space $\mathbb{E}$ (as was seen in Figure 7 for a small portion of the GIS polygon). The resulting position of these pixels in $\mathbb{E}$ is provided by the values of those pixels in the empirical GIS layers of the environmental variables, as follows: first, we obtained these layers as described in Section 2.2; then, we used the Python package `rasterio` (Gillies, Ward, and Peterson 2013) to import the empirical GIS layers into Python as matrices; finally, we defined the mapping f by sending each element (x, y) of a meshgrid to the point in $\mathbb{R}^3$ given by $(e_1(x, y), e_2(x, y), e_3(x, y))$ as mentioned above, where the e_i are the matrices obtained from the environmental GIS layers and $e_i(x, y)$ is the (x, y)-entry of that matrix.

More generally, we could use any number m of environmental variables, and consider the image $f(U)$ in $\mathbb{R}^m$. This map f would be equivalent (up to isometry of $\mathbb{R}^m$) to considering the metric $||.||_{\mathbb{E}}$ on U, where $d(p, q) := ||p - q||_{\mathbb{E}}$ is the Euclidean distance between $f(p)$ and $f(q)$ in m-dimensional environmental space $\mathbb{E}$. Note that the environmental manifold may not be a 'manifold' strictly in

the mathematical sense, and that the metric $||.||_{\mathbb{E}}$ could fail positive-definiteness, since it could cross over itself; this would happen precisely when two points in the region U of geographical space have the same value for each of the m environmental variables considered in $\mathbb{E}$. In light of this, we have not shortened the name 'environmental manifold' to 'manifold' at any instance in this work.

One exciting aspect of the environmental manifold is that its setting provides a fertile ground for mathematical insights in an ecological context, particularly those of a differential, geometric or topological nature. The geometric explanation below of the failure of the GLM in particular cases provides a basic example of this, in which we utilised notions of the geometry of intersections of certain shapes with the environmental manifold. In particular, this geometric viewpoint afforded the insight that an important role is played by the shape of the environmental manifold (which is precisely the realisable niche) in predicting species distributions.

There is much to explore with this new geometric ecological tool. One example which arises in this paper is that the occurrence probability of a species over some geographical region U can be thought of as a map from U to the unit interval $[0, 1]$. In our case, we obtained this as the composition of the three maps $\phi \circ z \circ f : U \to f(U) \to \mathbb{R} \to [0, 1]$. Recall that the map ϕ was given by the logistic transform $\phi(z) = e^z/(1 + e^z)$ of a function $z : \mathbb{E} \to \mathbb{R}$. As a result, we can view the 'thickened' level sets $c_\epsilon = \{\mathbf{x} \in f(U) \mid |\phi(\mathbf{x}) - c| < \epsilon\}$ of ϕ restricted to the environmental manifold $f(U)$ as those regions of $f(U)$ with occurrence probability approximately c, which will typically look like 1-dimensional subspaces of $f(U)$. Framing species abundance thresholds in terms of these level sets, and recruiting the smoothing function to help us recover these level sets from noisy data, could notions of gradient descent, Morse theory, and persistent homology (Ghrist 2014) be recruited for describing the realised niche?

Furthermore, among methods which seek to recover the underlying distribution of a species along an environmental gradient, the smoothing function (as expressed mathematically in Section 2.1) is unique in being a genuine mathematical function on the occurrence data points themselves, and so is open to mathematical analysis in a way in which models such as the GLM and random forest are not. For example, could we understand more deeply the shapes of the correlations seen in Figures 1 and 2 from the knowledge of: (1) the environmental manifold geometry, and (2) the smoothed species data on the environmental manifold? This would give us insight into not just a richer and more accurate description of the Hutchinsonian niche, but would also help practitioners in conservation and management effect policy which better reflects the underlying response of species to environmental gradients.

B.1 The geometry of intersections

We explore here the error of the GLM, when the environmental manifold is a plane P, in recovering the slope coefficients $(\beta_0, \beta_1, \beta_2, \beta_3)$ in the linear relationship $z = \beta_1 X + \beta_2 Y + \beta_3 Z + \beta_0$. The answer is of a geometric nature, and is due precisely to the shape of the environmental manifold. We denote the specific linear relationship $X + Y - Z - 1$ and environmental manifold plane $4X + Y - 2Z = 5$ from our earlier analysis by z_1 and P_1 respectively, to distinguish it from a general linear relationship z and environmental manifold plane P, discussed in this section.

First, recall that the GLM tries to predict the coefficients β_i defining z. So, from the GLM we obtain a prediction $\overline{z} = \overline{\beta}_1 X + \overline{\beta}_2 Y + \overline{\beta}_3 Z + \overline{\beta}_0$, where the $\overline{\beta}_i$ are the predictions for the β_i. This equation for $\overline{z}$ defines a set of planes in three-dimensional space as the value of $\overline{z}$ varies, corresponding to difference levels of occurrence probability. These planes of $\overline{z}$ move along the normal direction $\mathbf{n} = (\overline{\beta}_1, \overline{\beta}_2, \overline{\beta}_3)$, and thus intersect the environmental manifold plane P in a series of parallel straight lines.

The key point is that the accuracy of the GLM in recovering the simulated linear relationship z_1 (Table 1) must mean that these parallel lines of intersection between the predicted planes of $\overline{z_1}$ and P_1 are almost exactly the same as those of z_1 and P_1. However, given these parallel lines of intersection, there is still one degree of freedom in defining the planes given by $\overline{z_1}$. With this information, the GLM was only recover z_1 up to a rotation along these parallel lines, as we see in Figure B.3. Indeed, the environmental manifold plane P_1, relationship plane z_1 and predicted plane $\overline{z_1}$ all differ by a rotation about the same axis of intersection with P_1, with the true relationship z_1 and the predicted relationship $\overline{z_1}$ differing by approximately 65 degrees along this axis.

To investigate this further, we explored rotating and translating the environmental manifold plane P (which, for this problem, would be equivalent to varying the coefficients β_i defining z), adding degrees of noise orthogonal to P, and adding nonlinearities orthogonal to P (such as sinusoidal surfaces). We found that, when the environmental manifold was a plane, the accuracy of the predicted slope coefficients was determined by the angle between the environmental manifold plane P and the relationship planes defined by z, where accuracy of $\overline{z}$ increases with orthogonality between z and P. When nonlinearities were added, the accuracy of the predicted coefficients depended on the manner in which the relationship place z intersected the environmental manifold. Some of these findings are displayed in Table B.1.

These examples demonstrate that when the environmental manifold plane P lies orthogonal to

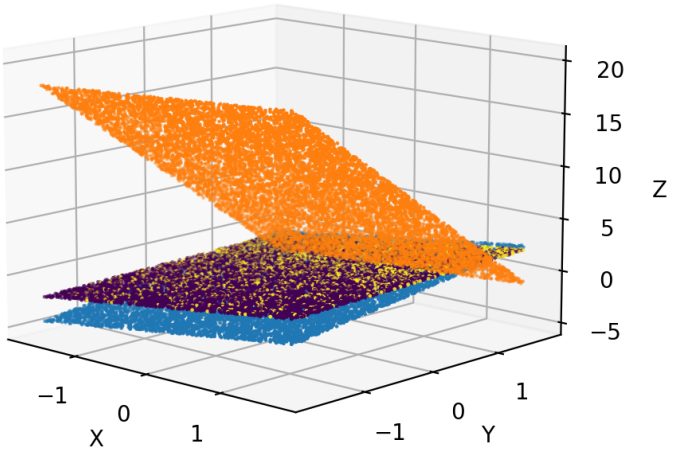

Figure B.3: The environmental manifold plane P_1 used in our simulation analysis, defined by the equation $4X + Y - 2Z - 5 = 0$. It is coloured by yellow and dark blue, corresponding to locations of presence and absence of the simulated species. The plane given by the linear relationship z_1 is coloured light blue, and the plane predicted by the GLM is coloured orange. The predicted plane has coefficients $(-0.063, 0.705, 0.113, -1.013)$, in contrast with the true relationship plane $(1, 1, -1, -1)$, and they differ by an angle of approximately 65 degrees.

Manifold	Predicted coefficients
P_1	(-0.063,0.705,0.113,-1.013)
Plane orthogonal to z_1	(0.996,1.013,-0.979,-0.958)
Plane approximately tangent to z_1	(0.010,0.014,0.017,0.007)
Plane defined by $Z = 0$	(1.013,1.010,0,-0.968)
P_1 with 0.1 orthogonal random noise	(2.148,1.263,-1.554,-2.334)
P_1 with 0.5 orthogonal random noise	(0.898,1.016,-0.951,-0.893)
Sinusoidal surface with 0.5 amplitude	(0.778,0.763,-0.008,-0.999)
Sinusoidal surface with 5 amplitude	(1.068,0.946,-0.991,-1.003)

Table B.1: The GLM's prediction for the slope coefficients $(1, 1-1, -1)$ of the linear relationship $z_1 = X + Y - Z - 1$, with six different shapes for the environmental manifold. Some are linear, and some are nonlinear to differing degrees. P_1 is the environmental manifold plane used in our simulation analysis.

a linear relationship z, the GLM's predicted coefficients are most accurate. In contrast, when P and z are approximately tangent, the predictions are extremely inaccurate. In the case of the plane P defined by the equation $Z = 0$, we find the coefficients of X, Y and the intercept accurately recovered, but the coefficient of Z is predicted to be 0. This is an example of the phenomenon that, if the environmental sampling is approximately constant along some environmental gradient, then the GLM cannot predict the coefficient for this environmental variable. The latter four cases in the table relax the constraint of nonlinearity, and demonstrate that if the environmental manifold is not sufficiently nonlinear, then the same errors can arise.

So, the GLM can only 'see' the species-environment relationship z where it intersects the environmental manifold. Thus, when the environmental manifold is a plane - or more generally, approximately constant along some linear combination of axes in environmental space - the GLM does not have enough information to recover the coefficients β_i defining a linear relationship z. This also explains the accuracy of $\overline{z}$ in the other three choices of environmental manifold in our simulation analysis, in which the GLM sees enough of the species response planes z via their intersections with the environmental manifold M to fix the remaining degree of freedom. For example, if the intersections are shaped like a circle (like the case in our simulation when M was a sphere), then M curves into enough dimensions of environmental space to fix the predicted planes $\overline{z}$.

Thus, in this simple example in three-dimensional space when the environmental data are constant along certain axes (as is the case for a plane), or more generally when the intersections of z with M are not sufficient to fix the coefficients $\overline{\beta}_i$, the GLM may dramatically fail to recover the parameters of a species-environment relationship, precisely because of the manner in which the environmental variables co-vary, a phenomenon which only becomes apparent with the help of the environmental manifold.

Acknowledgements

Siddharth Unnithan Kumar would like to warmly thank his family, and also Sung Hyun Lim, Alexandra Georges-Picot, Gonzalo Gonzalez de Diego, André Henriques, Jacob Socolar and Rafferty James Lindon for enriching conversations; for inspiration and wonder, a deep thanks to David Abram; for their compassion and support, Sandhya Patel and Andrew Teal; and for nourishment and sustenance, he bows to the Buddha, Dharma and Sangha, to Port Meadow, Binsey Lane, and above all, our more-than-human earth. This work was kindly supported by Oxford University's Pembroke College-Mathematical Institute PhD scholarship.

References

[1] David Abram. *The Spell of the Sensuous: Perception and Language in a More-Than-Human World*. Pantheon books, 1996.

[2] David Abram. *Becoming Animal: An Earthly Cosmology*. Vintage, 2010.

[3] Eric Ash, David W Macdonald, Samuel A Cushman, Adisorn Noochdumrong, Tim Redford, and Żaneta Kaszta. *Optimization of spatial scale, but not functional shape, affects the performance of habitat suitability models: a case study of tigers (Panthera tigris) in Thailand*. 2021. DOI: https://doi.org/10.1007/s10980-020-01105-6.

[4] Luciano Atzeni, Samuel A Cushman, Defeng Bai, Jun Wang, Pengju Chen, Kun Shi, and Philip Riordan. *Meta-replication, sampling bias, and multi-scale model selection: A case study on snow leopard (Panthera uncia) in western China*. 2020. DOI: https://doi.org/10.1002/ece3.6492.

[5] Benjamin Blonder, Christine Lamanna, Cyrille Violle, and Brian J Enquist. *The n-dimensional hypervolume*. 2014. DOI: https://doi.org/10.1111/geb.12146.

[6] Leo Breiman. *Bagging predictors*. 1996. DOI: https://doi.org/10.1007/BF00058655.

[7] Leo Breiman. *Random forests*. 2001. DOI: https://doi.org/10.1023/A:1010933404324.

[8] Olivier Broennimann, Matthew C Fitzpatrick, Peter B Pearman, Blaise Petitpierre, Loïc Pellissier, Nigel G Yoccoz, Wilfried Thuiller, Marie-Josée Fortin, Christophe Randin, Niklaus E Zimmermann, et al. *Measuring ecological niche overlap from occurrence and spatial environmental data*. 2012. DOI: https://doi.org/10.1111/j.1466-8238.2011.00698.x.

[9] Richard Chandler and Jeffrey Hepinstall-Cymerman. *Estimating the spatial scales of landscape effects on abundance*. 2016. DOI: https://doi.org/10.1007/s10980-016-0380-z.

[10] Luca Chiaverini, Ho Yi Wan, Beth Hahn, Amy Cilimburg, Tzeidle N Wasserman, and Samuel A Cushman. *Effects of non-representative sampling design on multi-scale habitat models: flammulated owls in the Rocky Mountains.* 2021. DOI: https://doi.org/10.1016/j.ecolmodel.2021.109566.

[11] William Cronon. *Uncommon Ground: Rethinking the Human Place in Nature*. W. W. Norton & Company, 1997.

[12] Samuel A Cushman, Ewan A Macdonald, Erin L Landguth, Yadvinder Malhi, and David W Macdonald. *Multiple-scale prediction of forest loss risk across Borneo*. 2017. DOI: https://doi.org/10.1007/s10980-017-0520-0.

[13] Samuel A Cushman and Kevin McGarigal. *Hierarchical, multi-scale decomposition of species-environment relationships*. 2002. DOI: https://doi.org/10.1023/A:1021571603605.

[14] Samuel A Cushman and Tzeidle N Wasserman. *Landscape applications of machine learning: comparing random forests and logistic regression in multi-scale optimized predictive modeling of American marten occurrence in northern Idaho, USA*. 2018. DOI: https://doi.org/10.1007/978-3-319-96978-7_9.

[15] Samuel A. Cushman and Falk Huettman. *Spatial Complexity, Informatics, and Wildlife Conservation*. Springer, Tokyo, 2010.

[16] Samuel A. Cushman and Erin L. Landguth. *Spurious correlations and inference in landscape genetics*. 2010. DOI: https://doi.org/10.1111/j.1365-294X.2010.04656.x.

[17] D Richard Cutler, Thomas C Edwards Jr, Karen H Beard, Adele Cutler, Kyle T Hess, Jacob Gibson, and Joshua J Lawler. *Random forests for classification in ecology*. 2007. DOI: https://doi.org/10.1890/07-0539.1.

[18] Jeffrey S Evans, Melanie A Murphy, Zachary A Holden, and Samuel A Cushman. *Modeling species distribution and change using random forest*. 2011. DOI: https://doi.org/10.1007/978-1-4419-7390-0_8.

[19] Robert Ghrist. *Elementary Applied Topology*. Createspace, 2014.

[20] S. Gillies, B. Ward, and A.S. Peterson. *Rasterio: Geospatial raster I/O for Python programmers*. 2013. URL: https://github.%20com/mapbox/rasterio.

[21] Matthew C Hansen, Peter V Potapov, Rebecca Moore, Matt Hancher, Svetlana A Turubanova, Alexandra Tyukavina, David Thau, SV Stehman, Scott J Goetz, Thomas R Loveland, et al. *High-resolution global maps of 21st-century forest cover change*. 2013. DOI: 10.1126/science.1244693.

[22] Andrew J Hearn, Samuel A Cushman, Joanna Ross, Benoit Goossens, Luke TB Hunter, and David W Macdonald. *Spatio-temporal ecology of sympatric felids on Borneo. Evidence for resource partitioning?* 2018. DOI: https://doi.org/10.1371/journal.pone.0200828.

[23] Troy M Hegel, Samuel A Cushman, Jeffrey Evans, and Falk Huettmann. *Current state of the art for statistical modelling of species distributions*. 2010. DOI: https://doi.org/10.1007/978-4-431-87771-4_16.

[24] Robert D Holt. *Bringing the Hutchinsonian niche into the 21st century: ecological and evolutionary perspectives*. 2009. DOI: https://doi.org/10.1073/pnas.0905137106.

[25] Robert D Holt. *Reflections on niches and numbers*. 2020. DOI: https://doi.org/10.1111/ecog.04828.

[26] Johan van den Hoogen et al. *Soil nematode abundance and functional group composition at a global scale*. 2019. DOI: https://doi.org/10.1038/s41586-019-1418-6.

[27] G. Evelyn Hutchinson. *Concluding remarks*. 1957. DOI: 10.1101/SQB.1957.022.01.039.

[28] Tim Ingold. *The perception of the environment: essays on livelihood, dwelling and skill*. Routledge, 2000. DOI: https://doi.org/10.4324/9780203466025.

[29] Andy Jarvis, Hannes Isaak Reuter, Andy Nelson, and Edward Guevara. *[dataset] Hole-filled seamless SRTM data V4*. 2008. URL: https://srtm.csi.cgiar.org/.

[30] Żaneta Kaszta, Samuel A Cushman, Andrew J Hearn, Dawn Burnham, Ewan A Macdonald, Benoit Goossens, Senthilvel KSS Nathan, and David W Macdonald. *Integrating Sunda clouded leopard (Neofelis diardi) conservation into development and restoration planning in Sabah (Borneo)*. 2019. DOI: https://doi.org/10.1016/j.biocon.2019.04.001.

[31] Robin Wall Kimmerer. *Braiding Sweetgrass: Indigenous Wisdom, Scientific Knowledge, and the Teachings of Plants*. Milkweed Editions, 2013.

[32] Simon A Levin. *The problem of pattern and scale in ecology: the Robert H. MacArthur award lecture*. 1992. DOI: https://doi.org/10.2307/1941447.

[33] Andy Liaw, Matthew Wiener, et al. *Classification and regression by randomForest*. 2002. URL: https://cogns.northwestern.edu/cbmg/LiawAndWiener2002.pdf.

[34] David W Macdonald, Luca Chiaverini, Helen M Bothwell, Żaneta Kaszta, Eric Ash, Gilmoore Bolongon, Özgün Emre Can, Ahimsa Campos-Arceiz, Phan Channa, Gopalasamy Reuben Clements, et al. *Predicting biodiversity richness in rapidly changing landscapes: climate, low human pressure or protection as salvation?* 2020. DOI: https://doi.org/10.1007/s10531-020-02062-x.

[35] Kevin McGarigal, Ho Yi Wan, Kathy A Zeller, Brad C Timm, and Samuel A Cushman. *Multi-scale habitat selection modeling: a review and outlook*. 2016. DOI: https://doi.org/10.1007/s10980-016-0374-x.

[36] Chunrong Mi, Falk Huettmann, Yumin Guo, Xuesong Han, and Lijia Wen. *Why choose Random Forest to predict rare species distribution with few samples in large undersampled areas? Three Asian crane species models provide supporting evidence*. 2017. DOI: 10.7717/peerj.2849.

[37] John Ashworth Nelder and Robert WM Wedderburn. *Generalized linear models*. 1972. DOI: https://doi.org/10.2307/2344614.

[38] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. *Scikit-learn: Machine learning in Python*. 2011. URL: https://www.jmlr.org/papers/volume12/pedregosa11a/pedregosa11a.pdf?source=post_page----------.

[39] Andrea Raya Rey and Falk Huettmann. *Telecoupling analysis of the Patagonian Shelf: A new approach to study global seabird-fisheries interactions to achieve sustainability*. 2020. DOI: https://doi.org/10.1016/j.jnc.2019.125748.

[40] Skipper Seabold and Josef Perktold. *Statsmodels: Econometric and statistical modeling with python*. 2010. URL: https://pdfs.semanticscholar.org/3a27/6417e5350e29cb6bf04ea5a4785601d5a.pdf.

[41] AJ Shirk, SA Cushman, and EL Landguth. *Simulating pattern-process relationships to validate landscape genetic models*. 2012. DOI: https://doi.org/10.1155/2012/539109.

[42] B. W. Silverman. *Density Estimation for Statistics and Data Analysis*. Chapman & Hall, 1986.

[43] Gary Snyder. *The Practice of the Wild*. North Point Press, 1990.

[44] Jorge Soberón. *Grinnellian and Eltonian niches and geographic distributions of species*. 2007. DOI: https://doi.org/10.1111/j.1461-0248.2007.01107.x.

[45] Heidi K Swanson, Martin Lysy, Michael Power, Ashley D Stasko, Jim D Johnson, and James D Reist. *A new probabilistic method for quantifying n-dimensional ecological niches and niche overlap*. 2015. DOI: https://doi.org/10.1890/14-0235.1.

[46] WCS and CIESIN. *[dataset] Last of The Wild data version 2, (LTW-2): global human footprint dataset*. 2005. DOI: https://sedac.ciesin.columbia.edu/data/collection/wildareas-v2.

[47] Mark J Whittingham, Philip A Stephens, Richard B Bradbury, and Robert P Freckleton. *Why do we still use stepwise modelling in ecology and behaviour?* 2006. DOI: https://doi.org/10.1111/j.1365-2656.2006.01141.x.

[48] John A Wiens. *Spatial scaling in ecology*. 1989. DOI: https://doi.org/10.2307/2389612.