Model simplicity breeds contempt: using simple models to answer basic questions on species' distributions

Bernat Bramon Mora[1,*] and Jake M. Alexander[1]

[1]Institute of Integrative Biology, ETH Zürich, Zürich, Switzerland; [*]bernat.bramon@gmail.com

Abstract

We know a lot about the factors that could theoretically influence species' distributions, and a rapidly growing body of research have been primarily focused on trying to untangle some of such biotic and abiotic predictors—with an increasing effort placed in improving the predictive power of statistical models. However, much less is known about how species' distributions compare to each other. Here, we use a conceptually more conservative approach to instead understand and compare basic aspects regarding the shape of species' distribution along environmental gradients.

Introduction

One of the central goals of ecology is to understand the ways species are distributed across space (ref). Over the years, researchers have developed multiple distribution models to try to untangle the factors that play a role in defining such distributions (ref). These models estimate species' realized niches using several covariates, including environmental variables (ref), species ecological traits' (ref) and phylogenetic relations (ref). Recent work on these approaches has focused on estimating (ref) and accounting for (ref) biotic factors, such as competitive or facilitative relationships. The idea is that by understanding how such factors shape species' distributions, we will gain a mechanistic understanding on how ecological communities are established and change over time. Unfortunately, while some of this approaches can increase the predictive performance of distribution models (ref), the nature of some of the estimates have been theoretically questioned (ref).

Increasing efforts have been devoted to improving the ability of statistical models to predict the presence/absence of species across ranges (ref). Accurately predicting how species are distributed across ranges is crucial for understanding the impacts and effect of global climate change. With that said, accounting for the mechanisms can come at the cost of modelling noise... However, much less attention is paid to how species' distributions compare to each other.

Increasing efforts have been devoted to improving the ability of statistical models to predict the presence/absence of species across ranges. However, much less attention is paid to how species' distributions compare to each other.

A lot can be learned from the basic properties of species' realized niches.

Accounting for expert knowledge on species' environmental preferences to understand general distribution patterns.

Here, we...Bayesian framework... This allow us to account for... as well as to tackle long-standing hypothesis regarding basic aspects of species distributions.

In this work, we first....

There is no general agreement the shape of species distributions. While many ecological textbooks (Begon et al., 1990, Giller, 1984, Krebs, 1994) assume this to be unimodal and symmetric, some have warned that empirical distributions can take many different forms (Austin, 2002). There is not an easy way to untangle the true shape of species' distributions, as this shape is likely to showcase idiosyncrasies at the species level and across systems. The aim of this work, it is not to answer these questions nor to provide a general approach that accommodates such idiosyncrasies. Instead, we want to use a model that is solely constrained by the empirical information that we truly have regarding a particular system, relaxing as much as possible the structural constrains of the statistical framework. Then, we want to use this model to answer basic aspects regarding the way systems of many species are distributed along an environmental gradient.

To decide among modelling approaches, we first needs to agree on what we know about the system. We know that species occupy a geographic range; therefore, we know that their distributions have finite variance. Indeed, observations on species' geographic variation and optimal climatic conditions have been long documented, with extensive databases compiled by botanists and field ecologists documenting basic knowledge on species' distributions. One could point out that we also know that many other factors might influence species' presence/absence—e.g. the influence of biotic interactions among species. However, we do not necessarily have an intuition of how exactly these factors will influence the shape of species' distributions. As a result, if all we truly knew about a species' distribution was that they have finite variance, the most conservative assumption and the safest bet—i.e. the one with the largest entropy—is that such distribution is a Gaussian.

Scarce data and little to no attempt to account for uncertainty in the predictions. Similar to rapoport's rule, we can also ask other questions regarding general geographical patterns of species distributions.

Methods

Empirical data

We studied the distribution of alpine plant communities along an elevation gradient. To do so, we combined two different datasets: i) one describing the co-occurrence of species across multiple open grasslands in the Swiss Alps, and ii) an extensive floristic database containing environmental and physiological traits for all vegetation across Switzerland (Landolt *et al.*, 2010).

Distribution data

We studied the distribution of 798 species across 912 sites covering most of the mountain region of the Western Alps in the Canton de Vaud (Switzerland; Scherrer & Guisan 2019). Each of these sites is a 8×8 m plot placed somewhere along an elevation range from 375 m to 3210 m. In all sites, presence/absence data as well as Braun-Blanquet abundance-dominance classes were recorded for all species. Additionally, following 30 years (1961–1990) of meteorological data from national weather stations, Scherrer & Guisan (2019) calculated multiple climatic variables for each site at high spatial resolution (25 m). Here, we focussed on 9 climatic variables, including: daily minimum, maximum and average temperature; sum of growing degree-days above 5°C; mean temperature of wettest quarter; annual precipitation, precipitation seasonality, and precipitation of driest quarter.

Floristic data

To complement the aforementioned distribution data, we used a floristic database of most vegetation across Switzerland. This database was build based on expert knowledge and field experience of botanists and ecologists, and contains information regarding species' environmental preferences and physiological traits. Species' environmental preferences in this database can be used to inform distribution models—e.g. as an informative prior in a Bayesian framework. These are characterized following the ecological indicator values developed by Landolt *et al.* (2010), providing both an estimate of the average conditions in which a species can be found and a broad description of their range of variation. These values are provided for a range of 10 climatic variables, including temperature, continentality, light conditions, as well as moisture, acidity and nutrient content of the soil (see a full list and description of the ecological indicators in the Supplementary Methods; Landolt *et al.* 2010). On the other hand, the information regarding species' physiological traits represent general descriptions of species' growth and life strategies—examples include their growth forms, nature of the storage organs, dispersal ability and pollinator agents. In total, we identify more than 120 binary traits that characterize the physiology of species (see a full list and description of the ecological indicators in the Supplementary Methods; Landolt *et al.* 2010).

[Trait data]

This could be Tom's data if we end up using it.

Distribution model

There is a long list of model structures well suited to characterize species' distributions (see XX for a review); however, we were interested in a model that explicitly incorporates all information regarding plant's environmental preferences found in the floristic database. More specifically, we wanted to account for the climatic indicator values and range of variation registered for all plants in our dataset. These two values provide basic information regarding plant's optimal environmental conditions and width of their distributions. Therefore, we first formulated a baseline model that directly accounts for such prior information.

Baseline model

Given y_{ij} the presence/absence of any species i in any given site j, and a set of k environmental variables x_{jk}, we estimate species' distributions as:

$$\begin{aligned}
y_{ij} &\sim \text{Binomial}\left(1, p_{ij}\right) \\
\log\left(p_{ij}\right) &= -\alpha_i - \sum_k \lambda_{ik}\left(x_{jk} - \beta_{ik}\right)^2 \\
\log(\alpha) &\sim \text{MVNormal}\Big(\hat{\alpha}, \Sigma^{\alpha}\Big) \\
\beta_{ik} &\sim \text{MVNormal}\left(\hat{\beta}_k, \Sigma^{\beta_k}\right) \\
\log(\lambda_{ik}) &\sim \text{MVNormal}\left(\hat{\lambda_k}, \Sigma^{\lambda_k}\right) \\
\hat{\alpha}, \hat{\lambda^k}, \hat{\beta^k} &\sim \text{Normal}\left(0, 1\right)
\end{aligned} \tag{1}$$

Notice that this model structure assumes all plants to have a uni-modal distributions along each environmental axis (see the model's behaviour in Supplementary Figure XX), where parameters α_i, β_i^k, and λ_i^k describe amplitude of the probability p_{ij}, species' average climatic suitability and range of variation along the different environmental gradients, respectively[†]. While potentially sacrificing predictive accuracy, this model structure allows us to explicitly incorporate all prior knowledge that we have regarding species' distributions via Σ^{α}, Σ^{β_k} and Σ^{λ_k}. More specifically, we express β_i^k and $\log\left(\lambda_i^k\right)$ as multivariate normal distributions—i.e. Gaussian processes—such that Σ^{β_k} and Σ^{λ_k} are variance-covariance matrices describing species' similarity in terms of their average climatic suitability and range of variation along the different environmental gradients, respectively. Likewise, $\log\left(\alpha\right)$ is characterized as a Gaussian Process, where the corresponding variance-covariance matrix Σ^{α} is designed to also incorporate some of the prior information that we have with regards to species' physiological traits.

[†] I'll rewrite the likelihood function to an ordered categorical as soon as I get things to work properly with count data.

In all cases, all variance-covariance matrices are defined as follows:

$$\Sigma^{\chi}_{ij} = \eta_{\chi} \exp\left(-\rho_{\chi} {D^{\chi}_{ij}}^{2}\right) + \delta_{ij}\sigma_{\chi}, \tag{2}$$

where Σ^{χ}_{ij} describes the covariance between any pair of species i and j for any given parameter α_i, β^k_i, and λ^k_i. Following this expression, such covariance declines exponentially with the square of the different D^{χ}_{ij}, which are distance measures computed using the prior information that we have regarding species' distributions. Specifically, given α_i, β^k_i, and λ^k_i, the distance measures are calculated using plants' physiological traits, ecological indicator values and range of variation, respectively (see below for further details). For each covariance matrix, the hyperparameter ρ_{χ} determines the rate of decline of the covariance between any two species, and η_{χ} defines its maximum value. The hyperparameter σ_{χ} describes the additional covariance between the different observations for any given species. For any given hyperparameter, we choose adaptive priors across covariance structures. That is, and taking ρ_{χ} as an example, we choose a prior $\log(\rho_{\chi}) \sim \text{Normal}(\hat{\rho}, \sigma_{\rho})$ such that $\hat{\rho} \sim \text{Normal}(0, 1)$ and $\sigma_{\rho} \sim \text{Exponential}(1)$. Similar priors were chosen for both η_{χ} and σ_{χ}. We generated the posterior samples for the Bayesian models with the help of the R package 'rstan' to (Team *et al.*, 2019).

Distance matrices

The missing component in the description of model (1) is the distance matrices D^{χ} used to define the covariance matrices Σ^{α}, Σ^{β_k} and Σ^{λ_k}. In this model, such distance matrices characterize differences between plant species. In the floristic data, however, the prior information that we have for these differences is represented by a set of ordinal and categorical traits. More specifically, both the ecological indicator values and range of variation—which define the prior information that we have for β^k_i, and λ^k_i, respectively—are ordinal traits specified for all species. In contrast, the plants' physiological data—shaping the prior for the parameters α_i—are characterized by categorical data containing multiple missing entries. Therefore, we need to carefully compile this data into distance matrices in order to be able to feed this prior information into the model.

More generally, we want to understand the way N species are characterized by M categorical traits. One way to frame this problem is by using a network representation. Following the ideas presented by Godoy-Lorite *et al.* (2016), we assume that species can be connected to each of these traits by an interaction (i, j) that can be of any type $r \in R$. Notice that this provides as with multiple ways to account for the information—and lack thereof—contained in the different categorical and ordinal traits M. That is, the R types of interactions can represent the lack of information for a particular link (i, j), the absence or presence of such interaction, and any type of association between i and j.

Given a set of interactions R^* between N and M, we use a Mixed Membership Stochastic Block Model (MMSBM) to characterize these. In particular, we consider that plants and traits can be classified into K and L groups, respectively. For every species i, we assume that there is a probability $\theta_{i\alpha}$ for it to belong to any of the K species groups. Likewise, we also assume that any trait j has a probability $\phi_{j\beta}$ of belonging to any of the L trait groups. Finally, we define $p_{\alpha\beta}\left(r\right)$ as the probability of a species from group α interacting with a trait from group β by an association type r. Putting these together, the probability of an interaction (i, j) of type r can be calculated as:

$$Pr[r_{ij} = r] = \sum_{\alpha\beta} \theta_{i\alpha}\phi_{j\beta}p_{\alpha\beta}\left(r\right) \tag{3}$$

Following this definition, we want to find the group memberships that maximize the likelihood $P\left(R^*|\theta, \phi, p\right)$. Doing so is difficult optimization problem; however, it has been shown that one can estimate the different $\theta_{i\alpha}$, $\phi_{j\beta}$, and $p_{\alpha\beta}\left(r\right)$ parameters by maximizing the likelihood using an expectation-maximization algorithm (Godoy-Lorite *et al.*, 2016; Tarrés-Deulofeu *et al.*, 2019). In simple terms, one can iteratively find multiple local minima for the likelihood, and average over the estimated the parameter values (Godoy-Lorite *et al.*, 2016)[†].

[†] While this averaging is trivial for the estimated probabilities $Pr[r_{ij} = r]$, it is non-trivial if one wants to find averages for the group memberships. The reason for this is related to the stochastic nature of the expectation-maximization algorithm. This algorithm initially assigns random group memberships to both species and traits. While this random labelling is irrelevant when studying the probabilities $Pr[r_{ij} = r]$, it is instead crucial for averaging $\theta_{i\alpha}$, $\phi_{j\beta}$, and $p_{\alpha\beta}\left(r\right)$. Therefore, before averaging the group membership estimates, one needs to find the bijective relationship for the labellings of different iterations of the optimization algorithm. In a nutshell, for every iteration, I do this by using a simulated annealing algorithm on the estimated $p_{\alpha\beta}\left(r\right)$, matching the corresponding labelling to a reference iteration.

The average estimates for the group memberships provide us with a different scale to classify species based on the traits these have. In short, for any species i, we can estimate a K-dimensional vector $\vec{\theta}_i$ that describes the extend to which i belong to each group membership—i.e. the extend to which a species is of one type or another. This classification is useful because it can be used to compare species, defining a way to measure the distance between species based on an arbitrary—and potentially incomplete—set of categorical or ordinal traits M. The simplest case is to define the distance as $D_{ij} = |\vec{\theta}_i - \vec{\theta}_j|$. Alternatively, one could also define K distance matrices based on the different group memberships $D^{\alpha}_{ij} = |\theta_{i\alpha} - \theta_{j\alpha}|$.

Modifying the variance-covariance structures

The model structure defined in Eq. (1) allows us to test the effect of adding new information. Specifically, we can do this by modifying Eq. (2). For example, imagine that we have multiple matrices D^k characterizing species' differences along different axis of variation—i.e. two matrices characterizing ecological and environmental traits, or multiple matrices resulting from the different group memberships estimated using the MMSBM. One could modify Eq. (2) for a particular parameter—e.g. parameter α_i—such that

$$\Sigma^{\alpha}_{ij} = \eta_{\alpha} \exp\left(-\sum_{k} \rho_{\alpha k} {D^{k}_{ij}}^{2}\right) + \delta_{ij}\sigma_{\alpha}, \tag{4}$$

where now $\rho_{\alpha k}$ are separate relevance hyperparameters for each distance matrix in the total variance of α_i. Notice that the same is true for the covariance of parameters β^k_i and λ^k_i. Finally, for all hyperparameters and as described for the baseline model, we use adaptive priors across covariance structures.

Results

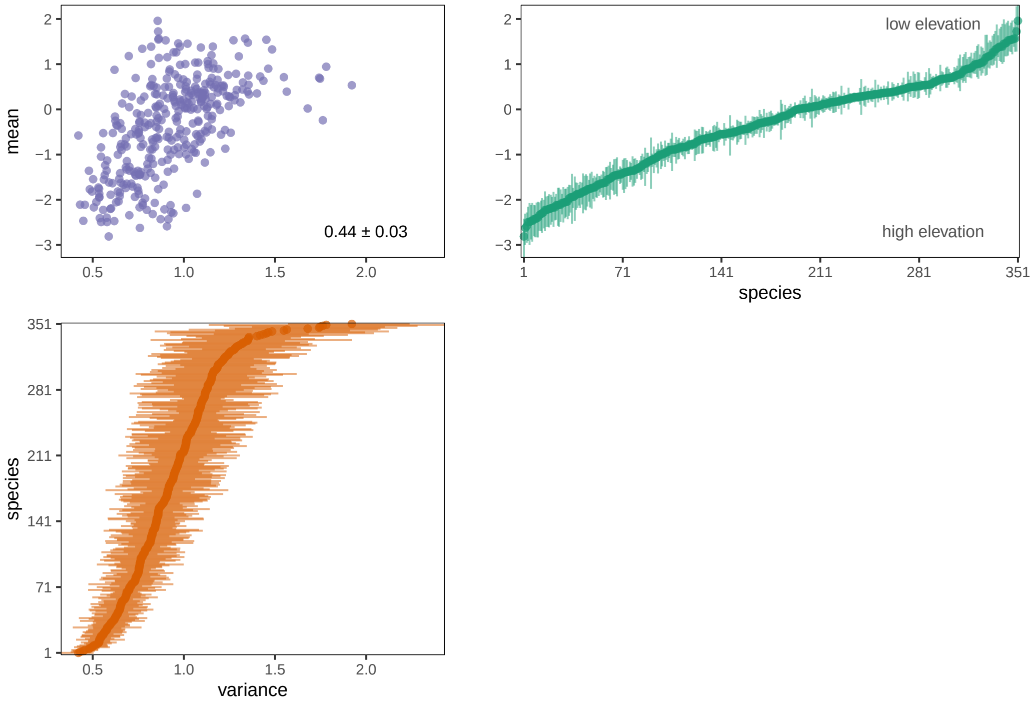

Figure 1: Relationship between mean and variance of species' distributions. These are the results for the main axis of variation for the climatic data (results for the second axis of variation presented in the Supplementary Fig. 2).

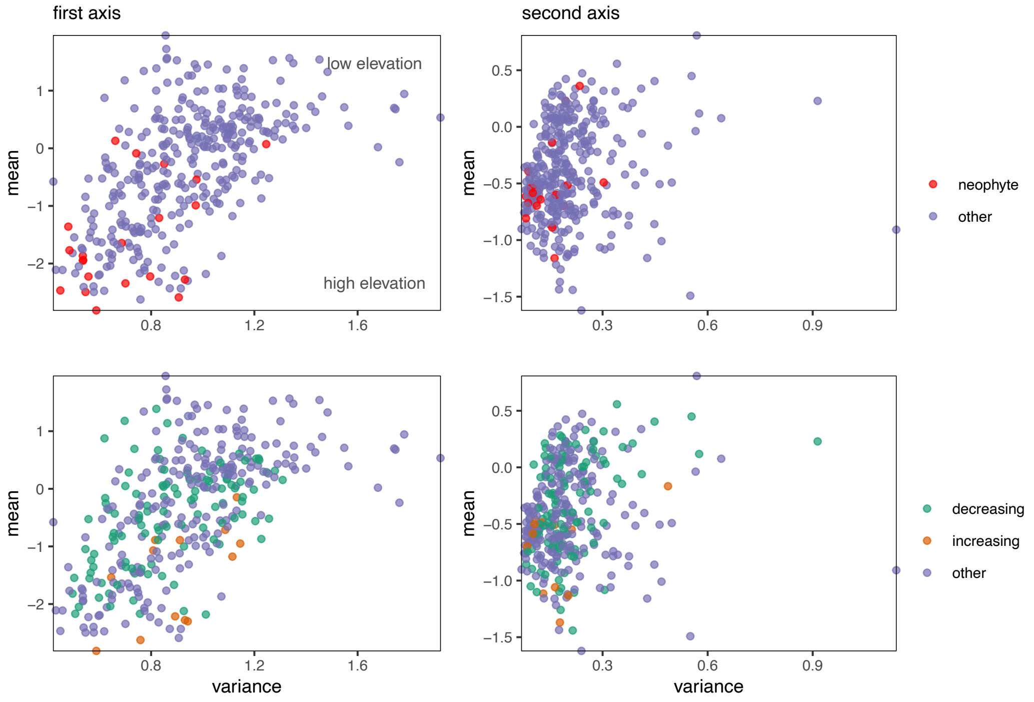

Figure 2: Are there clear geographical patterns for neophytes and for species with decreasing or increasing abundance?

References

Austin, M. (2002). Spatial prediction of species distribution: an interface between ecological theory and statistical modelling. *Ecological modelling*, 157, 101–118.

Godoy-Lorite, A., Guimerà, R., Moore, C. & Sales-Pardo, M. (2016). Accurate and scalable social recommendation using mixed-membership stochastic block models. *PNAS*, 113, 14207–14212.

Landolt, E., Bäumler, B., Ehrhardt, A., Hegg, O., Klötzli, F., Lämmler, W., Nobis, M., Rudmann-Maurer, K., Schweingruber, F. H., Theurillat, J.-P., Urmi, E., Vust, M. & Wohlgemuth, T. (2010). *Flora indicativa: Okologische Zeigerwerte und biologische Kennzeichen zur Flora der Schweiz und der Alpen*. Haupt, Bern. ISBN 978-3-258-07461-0.

Scherrer, D. & Guisan, A. (2019). Ecological indicator values reveal missing predictors of species distributions. *Scientific Reports*, 9, 1–8.

Tarrés-Deulofeu, M., Godoy-Lorite, A., Guimerà, R. & Sales-Pardo, M. (2019). Tensorial and bipartite block models for link prediction in layered networks and temporal networks. *Phys. Rev. E*, 99, 032307.

Team, S. D. *et al.* (2019). RStan: the R interface to Stan. R package version 2.19.1.