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 and time (ref). Over the last two decades, ecologists have developed multiple distribution models to try to untangle the factors that play a role in defining such distributions (Guisan & Zimmermann, 2000). These models estimate species' realized niches using several covariates, including environmental variables (?), species ecological traits' (Pollock *et al.*, 2012) and phylogenetic relations (Ives & Helmus, 2011). More recently, some of the focus have shifted towards approaches that estimate and account for biotic factors, such as competitive or facilitative relationships between species (Ovaskainen *et al.*, 2017). The idea is that by untangling the ways in which such biotic and abiotic factors shape species' distributions, we can gain a mechanistic understanding on how ecological communities are established and change over time. However, while these factors can increase the predictive performance of some of the models (Norberg *et al.*, 2019), the interpretation of the corresponding parameter estimates has been often questioned (Gotelli & Ulrich, 2010; Harris, 2016; Thurman *et al.*, 2019). This was best illustrated by Blanchet *et al.* (2020), who used basic statistical arguments to highlight the artefactual nature of the link between co-occurrence and species' ecological interactions drawn by some distribution models.

The value of gaining a mechanistic understanding of species' distributions is unquestionable (ref), with several studies highlighting the importance of factors such as biotic interactions and dispersal ability in setting species' range limits (Wisz *et al.*, 2013; Pollock *et al.*, 2014; Neuschulz *et al.*, 2018). That said, a lot can be learned from taking a phenomenological approach, focussing instead on the description of basic properties of species' realized niches. For example, the study of species' range sizes along environmental gradients can reveal general biodiversity patterns that are crucial from a conservation and management perspective (Stevens, 1992). Differences in species' responses to the environment could shed light on how climatic processes and historical contingencies have shaped their distributions (Rohde, 1992; **?**). Other properties, such as the skewness of species' distributions, can also reveal general underlying processes regarding species' physiological tolerance to different environmental conditions (Kaufman, 1995). More generally, understanding the shape of species' realized niches and the extend to which these vary across species is a crucial issue in ecology and biogeography (ref); however, we do not have an effective way to parsimoniously compare the realized niches of many species. Indeed, there is no general agreement on the shape of species' distributions (ref).

Many ecological textbooks (Krebs, 1972) assume the shape of species distributions to be unimodal and symmetric, but some have warned that empirical distributions can take many different forms (Austin, 1987; **?**). In practice, distribution frameworks often use logistic regressions with a linear relationship between covariates (but see XX and YY). This is useful because it simplifies the optimization process, but it comes with several statistical shortcomings. First and foremost, such response curve and the linear relationship between covariates often comes with a set of implicit mathematical constrains that might not be biologically justified. From a purely statistical perspective, if all that we are willing to assume is that species occupy finite geographic ranges—i.e. their probability distributions have finite variance—the most conservative statistical approach is to model these as a Gaussian distributions (Frank, 2009). This is rarely the starting point in most statistical frameworks that study general biodiversity patterns (but see ref), choosing to use instead Gaussian-logit response curves (refs). Other factors might then condition species distributions to showcase heavy-tails or a skewed shapes, revealing interesting ecological processes shaping biodiversity patterns (Austin, 1976; Minchin, 1987). The starting point, nevertheless, should be the one that makes the fewest assumptions (i.e. the maximum entropy distribution **?**), and every new shape will imply a hypotheses on how communities are distributed (D'Amen *et al.*, 2017). Second, the aforementioned structural constrains also limit our ability to include any prior information to our parameter estimates. Observations on species' geographic variation and optimal climatic conditions have long been documented, with extensive databases compiled by botanists and field ecologists documenting basic knowledge on species' realized niches (e.g. Landolt *et al.* 2010). That said, this information is rarely accounted for in most modelling approaches, mainly because there is not a straightforward way to feed this information into the parameters of a linear model (Scherrer & Guisan 2019; but see ter Braak & Looman 1986). Finally, and perhaps most importantly, a direct biological interpretation of parameter estimates in linear models becomes increasingly difficult as one moves from unimodal and symmetric distributions (Jamil & ter Braak, 2013; ter Braak & Looman, 1986) to skewed distributions (Huisman *et al.*, 1993), making the tests of hypothesis on global biodiversity patterns particularly challenging. For example, Huisman *et al.* (1993) proposed several non-linear models to characterize several features of species' response curves; however, species' environmental indicator values, range size or distribution skewness are difficult to understand altogether following these model structures.

The field of ecology has quickly moved towards mechanistic and process-based approaches to understand species' distributions (Warton *et al.*, 2015). This has resulted in a plethora of models accounting for several biotic and abiotic factors into the predictions of species co-occurrence. Here, we instead rethink traditional modelling approaches and develop a conceptually simple—and yet statistical and computationally complex—statistical framework to revisit some classic hypothesis in ecology and biogeoraphy. In particular, we develop a Bayesian hierarchical model that accounts for all prior information that we have regarding the distribution of alpine plant species along an elevation gradient in the Swiss Alps, including expert knowledge on species environmental indicator values, range sizes, and plant physiology. We start by considering species' response curves as Gaussian distributed, and then we adapt our model to allow for skewed and long-tailed distributions. Using this statistical framework, we are able to compare the basic properties of the realized niches of multiple species, testing for the existence of general biogeographical patterns. First, we test for the Rapopor's rule, which predicts a positive relationship between range size and elevation (Stevens, 1992). While this pattern has been largely studied for multiple systems and across gradients (McCain & Knight, 2013); contrasting evidence suggests this rule not to be pervasive across species (Ribas & Schoereder, 2006; Bhattarai & Vetaas, 2006; McCain & Knight, 2013). Our results not only allow us to properly test the existence of this geographical pattern, but they also showcase variation in how different types of species, such as native or neophytes, might respond to an environmental gradient. Second, we study whether or not species' distributions show steeper declines towards stressful conditions, testing the so-called abiotic stress limitation hypothesis (ref). Normand *et al.* (2009) tested this for vegetation data using Huisman *et al.*'s statistical models for several independent species, finding no clear support for such a hypothesis. Our results are able to shed light on this geographical pattern as well as to highlight the degree to which different species will showcase different levels of decline towards stressful conditions. Specifically, we are able to link plant physiological traits to the skewness of their distributions. Overall, we use models that are solely constrained by the empirical information that we truly have regarding our system, relaxing as much as possible the structural constrains of the statistical framework. Using these models, we are able uncover the true shape of empirical plant distributions and answer fundamental questions regarding the way systems of many species are distributed along environmental gradients.

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 used data describing 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, we used meteorological data provided by Scherrer & Guisan (2019), containing multiple variables characterizing the climate in each site at high spatial resolution (25 m). This dataset was compiled based on 30 years (1961–1990) of records from national weather stations. Since most of the data is highly correlated, we calculated the main axes of variation of the following variables: 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 (see Supplementary Methods; Supplementary Fig. 1).

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 as well as 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 represents 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).

Baseline model

There is a long list of model structures well suited to characterize species' distributions (see Norberg *et al.* 2019). As a baseline model, however, we were interested in a hierarchical model that does not make any assumptions regarding the shape of the distributions, and yet explicitly incorporates all information that we have regarding plant's environmental preferences. More specifically, we wanted to account for the climatic indicator values and range of variation registered in the floristic database for all plants in our dataset. These two values provide basic information regarding plant's optimal environmental conditions and width of their distributions.

Response curve

To choose an appropriate response curve, we first need to agree on what we truly know about the system. Given the prior information that we have about the system, we know that species occupy specific geographic ranges; therefore, we know that their distributions have finite variance. While we could also assume that many other factors might influence species' presence in a given site—e.g. the biotic interactions among specie in the site—we do not necessarily have an *a priori* expectation of how exactly these factors will influence the shape of species' distributions. Therefore, for this baseline model, if all that we are willing to assume about species' realized niches is that these have finite variance, the most conservative assumption and the safest bet—i.e. the one with the largest entropy—is that they follow a Gaussian distribution. That is, given the presence/absence or abundance y_{ij} of any species i in any given site j, and an environmental variable x_j, we define can species' responses to the environment as

$$\begin{aligned} y_{ij} &\sim F\left(p_{ij}\right) \\ \log\left(p_{ij}\right) &= -\alpha_i - \gamma_i\left(x_j - \beta_i\right)^2, \end{aligned} \tag{1}$$

where F is the likelihood function, and α_i, β_i^k, and γ_i describe amplitude of the probability p_{ij}, species' average climatic suitability and range of variation along the an environmental gradient, respectively. Notice that F characterizes a Bernoulli distribution when considering binary data, and it characterizes an ordered categorical likelihood function when we consider Braun-Blanquet abundance-dominance classes as response variables (see the full description of both models in the Supplementary Methods). For the sake of simplicity, we use only one environmental variable to characterize species' probability distribution. That said, this model can easily be generalized to account for multiple predictors (see Supplementary Methods).

Model priors

The model structure described above allows us to explicitly incorporate all prior knowledge that we have regarding species' distributions contained in the floristic database. To do so, we define the prior distributions for the parameters in model (1) as:

$$\begin{aligned} \beta_i &\sim \text{MVNormal}\left(\hat{\beta}, \Sigma^{\beta}\right) \\ \log(\gamma_i) &\sim \text{MVNormal}\left(\hat{\gamma}, \Sigma^{\gamma}\right) \\ \log(\alpha_i) &\sim \text{Normal}\left(\hat{\alpha}, \sigma_{\alpha}\right) \\ \hat{\beta}, \hat{\gamma}, \hat{\alpha} &\sim \text{Normal}\left(0, 1\right) \\ \sigma_{\alpha} &\sim \text{Exponential}\left(1\right) \end{aligned} \tag{2}$$

where parameters γ_i and β_i are expressed as multivariate normal distributions—i.e. Gaussian processes—such that Σ^{β} and Σ^{γ} are variance-covariance matrices describing species' similarity in terms of their average climatic suitability and range of variation along the different environmental gradients, respectively. We define these variance-covariance matrices as follows:

$$\Sigma_{ij} = \eta \exp\left(-\rho {D_{ij}}^2\right) + \delta_{ij}\sigma, \tag{3}$$

where Σ_{ij} characterizes the covariance between any pair of species i and j. Notice that such a covariance structure declines exponentially with the square of a distance matrix D_{ij}, which characterize differences between species computed using our prior information. In the floristic database, this information is represented by the set of ordinal specified for the different species. While there are many different ways to turn ordinal data into distance matrices, we choose to use a mixed-membership stochastic block model because it allows us to deal with cases of missing data (see Supplementary Methods for extended details; Godoy-Lorite *et al.* 2016). In each covariance matrix, the hyperparameter ρ determines the rate of decline of the covariance between any two species, and η defines its maximum value. The hyperparameter σ describes the additional covariance between the different observations for any given species. For all these hyperparameters, we choose weekly informative priors such that $\sigma, \eta \sim \text{Exponential}\,(1)$ and $\rho \sim \text{Exponential}\,(0.5)$.

Sampling the posterior

We generated the posterior samples for the Bayesian models with the help of the R package 'rstan' to (Stan Developent Team, 2021). Sampling models like the ones described above can be computationally very expensive. This is especially true when using ordered categorical likelihood functions (see Stan Development Team 2021). Therefore, we focus on those species for which we have more than 30 occurrences when modelling ordinal data, which is the case for the majority of the results of this work. When using presence/absence data, we limit our study to those species for which have more than 10 occurrences.

To test the performance of the model as well as our choice of prior distributions, we modelled simulated data and compared the sampled posterior distributions to the data-generating parameters (see Supplementary Methods; Supplementary Fig. 2). Notice that using the link function in Eq. 1 could cause problems when sampling the model, and some adjustments need to be made when specifying the model (see Supplementary Methods and the Code Availability section).

Modifying the baseline model

We proposed a baseline model that is naive in terms of the structural assumptions regarding the data, and yet accounts for all information we truly have about the system. Modifying this model, we can now test hypotheses regarding the properties of empirical species' distributions. To propose new species' response curves, however, we want to ensure two key conditions: (i) the probability distribution must have defined variance, and (ii) the Gaussian shape must be a special case of the probability distribution.

Heavy-tail response curve

Skewed response curve

Heavy-tail and skewed response curve

Alternative variance-covariance structures

The model structure defined in Eq. (9) allows us to test the effect of adding new information. Specifically, we can do this by modifying Eq. (10). 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. (10) 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.

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:

$$\begin{aligned}
y_{ij} &\sim \text{Binomial}\left(1, p_{ij}\right) \\
\log\left(p_{ij}\right) &= -\alpha_i - \sum_k \lambda_{ik}\left(x_{jk} - \beta_{ik}\right)^2 \\
\log(\alpha) &\sim \text{MVNormal}\Big(\hat{\alpha}, \Sigma^{\alpha}\Big) \\
\beta_{ik} &\sim \text{MVNormal}\left(\hat{\beta}_k, \Sigma^{\beta_k}\right) \\
\log(\lambda_{ik}) &\sim \text{MVNormal}\left(\hat{\lambda}_k, \Sigma^{\lambda_k}\right) \\
\hat{\alpha}, \hat{\lambda}^k, \hat{\beta}^k &\sim \text{Normal}\left(0, 1\right)
\end{aligned} \tag{5}$$

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.

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

$$\Sigma^{\chi}_{ij} = \eta_{\chi} \exp\left(-\rho_{\chi} D^{\chi\,2}_{ij}\right) + \delta_{ij}\sigma_{\chi}, \tag{6}$$

where Σ^{χ}_{ij} describes the covariance between any pair of species i and j for any given parameter α_i, β_i^k, and λ_i^k. Following this expression, such covariance declines exponentially

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

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 (**?**).

Covariance matrices from incomplete categorical and ordinal data

The prior information that we have regarding species' distributions is represented by the set of ordinal and categorical traits found in the floristic database. More specifically, both the ecological indicator values and range of variation are ordinal traits specified for all species, whereas plants' physiological data are characterized by categorical data containing multiple missing entries. These data could be directly used as covariates in any given distribution model; however, we want this information to be accounted for as a prior for the parameters of our Bayesian model. To do so, we need to compile the traits in the floristic database into variance-covariance matrices characterizing the *a priori* similarity between species.

We define these variance-covariance matrices are defined as follows:

$$\Sigma_{ij} = \eta \exp\left(-\rho {D_{ij}}^2\right) + \delta_{ij}\sigma, \tag{7}$$

where Σ_{ij} describes the covariance between any pair of species i and j. Following this expression, such covariance declines exponentially with the square of the different D_{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 (**?**).

—which define the prior information that we have for β_i^k, and λ_i^k, 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.

The missing component in the description of model (9) 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_i^k, and λ_i^k, 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:

$$Pr[r_{ij} = r] = \sum_{\alpha\beta} \theta_{i\alpha}\phi_{j\beta}p_{\alpha\beta}\left(r\right) \tag{8}$$

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)[†].

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

[†] 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.

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 $D^{\alpha}_{ij} = |\theta_{i\alpha} - \theta_{j\alpha}|$.

Distribution models

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:

$$\begin{aligned}
y_{ij} &\sim \text{Binomial}\left(1, p_{ij}\right) \\
\log\left(p_{ij}\right) &= -\alpha_i - \sum_k \lambda_{ik}\left(x_{jk} - \beta_{ik}\right)^2 \\
\log(\alpha) &\sim \text{MVNormal}\Big(\hat{\alpha}, \Sigma^{\alpha}\Big) \\
\beta_{ik} &\sim \text{MVNormal}\left(\hat{\beta}_k, \Sigma^{\beta_k}\right) \\
\log(\lambda_{ik}) &\sim \text{MVNormal}\left(\hat{\lambda_k}, \Sigma^{\lambda_k}\right) \\
\hat{\alpha}, \hat{\lambda^k}, \hat{\beta^k} &\sim \text{Normal}\left(0, 1\right)
\end{aligned} \tag{9}$$

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^k_i, and λ^k_i describe amplitude of the probability p_{ij}, species' average climatic suitability and range of variation along the different environmental gradients, respectively[†].

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

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, $\log(\alpha)$ 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.

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

$$\Sigma_{ij}^{\chi} = \eta_{\chi} \exp\left(-\rho_{\chi} {D_{ij}^{\chi}}^{2}\right) + \delta_{ij}\sigma_{\chi}, \tag{10}$$

where Σ_{ij}^{χ} describes the covariance between any pair of species i and j for any given parameter α_i, β_i^k, and λ_i^k. Following this expression, such covariance declines exponentially with the square of the different D_{ij}^{χ}, which are distance measures computed using the prior information that we have regarding species' distributions. Specifically, given α_i, β_i^k, and λ_i^k, 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 (**?**).

Distance matrices

The missing component in the description of model (9) 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_i^k, and λ_i^k, 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 $p_{\alpha\beta}(r)$ 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:

$$Pr[r_{ij} = r] = \sum_{\alpha\beta} \theta_{i\alpha}\phi_{j\beta}p_{\alpha\beta}(r) \qquad (11)$$

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)[†].

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. (9) allows us to test the effect of adding new information. Specifically, we can do this by modifying Eq. (10). 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

[†] 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.

modify Eq. (10) for a particular parameter—e.g. parameter α_i—such that

$$\Sigma_{ij}^{\alpha} = \eta_\alpha \exp\left(- \sum_k \rho_{\alpha k} {D_{ij}^{k}}^{2} \right) + \delta_{ij}\sigma_\alpha, \tag{12}$$

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_i^k and λ_i^k. Finally, for all hyperparameters and as described for the baseline model, we use adaptive priors across covariance structures.

Results

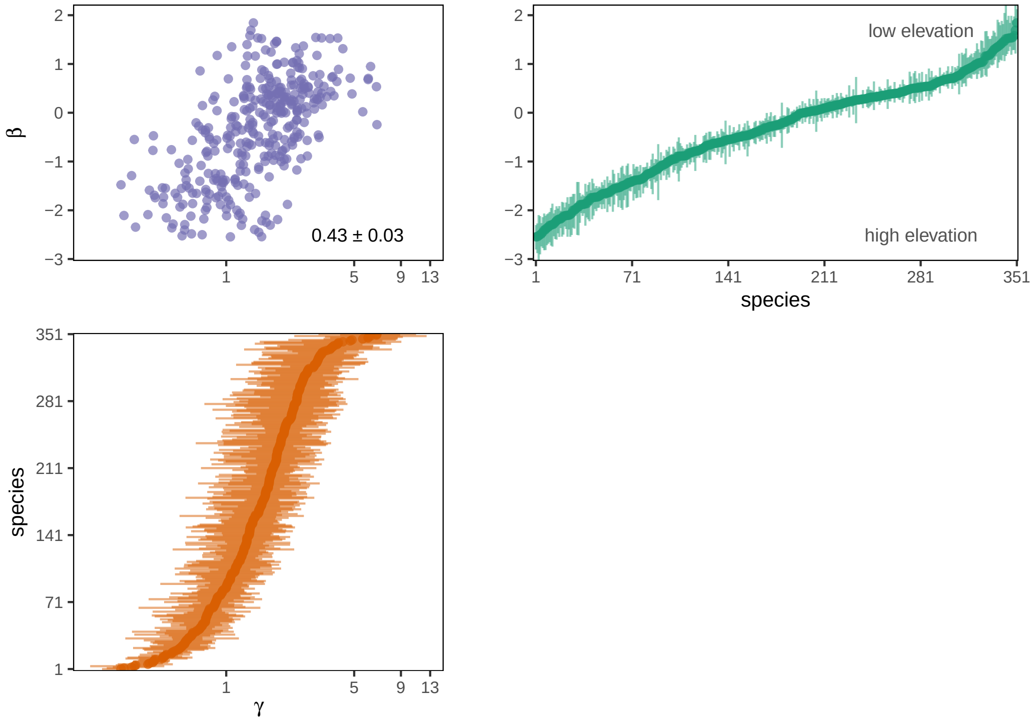

Figure 1: Relationship between mean and variance of species' distributions. These are the results for the main axis of variation for the climatic data (results for the second axis of variation presented in the Supplementary Fig. 2).

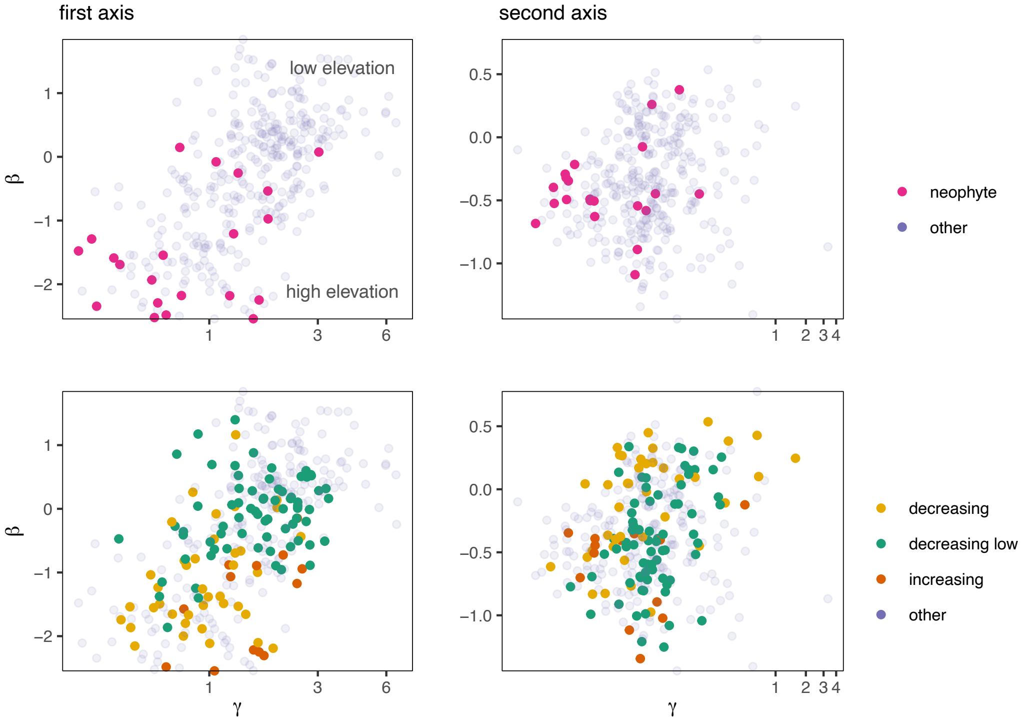

Figure 2: Are there clear geographical patterns for neophytes and for species with decreasing or increasing abundance?

References

Austin, M. (1976). On non-linear species response models in ordination. *Vegetatio*, 33, 33–41.

Austin, M. P. (1987). Models for the analysis of species' response to environmental gradients. *Vegetatio*, 69, 35–45.

Bhattarai, K. R. & Vetaas, O. R. (2006). Can Rapoport's rule explain tree species richness along the Himalayan elevation gradient, Nepal? *Diversity and Distributions*, 12, 373–378.

Blanchet, F. G., Cazelles, K. & Gravel, D. (2020). Co-occurrence is not evidence of ecological interactions. *Ecology Letters*, 23, 1050–1063.

D'Amen, M., Rahbek, C., Zimmermann, N. E. & Guisan, A. (2017). Spatial predictions at the community level: From current approaches to future frameworks. *Biological Reviews*, 92, 169–187.

Frank, S. A. (2009). The Common Patterns of Nature. *Journal of evolutionary biology*, 22, 1563–1585.

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

Gotelli, N. J. & Ulrich, W. (2010). The empirical Bayes approach as a tool to identify non-random species associations. *Oecologia*, 162, 463–477.

Guisan, A. & Zimmermann, N. E. (2000). Predictive habitat distribution models in ecology. *Ecological Modelling*, 135, 147–186.

Harris, D. J. (2016). Inferring species interactions from co-occurrence data with Markov networks. *Ecology*, 97, 3308–3314.

Huisman, J., Olff, H. & Fresco, L. F. M. (1993). A hierarchical set of models for species response analysis. *Journal of Vegetation Science*, 4, 37–46.

Ives, A. R. & Helmus, M. R. (2011). Generalized linear mixed models for phylogenetic analyses of community structure. *Ecological Monographs*, 81, 511–525.

Jamil, T. & ter Braak, C. J. F. (2013). Generalized linear mixed models can detect unimodal species-environment relationships. *PeerJ*, 1, e95.

Kaufman, D. M. (1995). Diversity of New World Mammals: Universality of the Latitudinal Gradients of Species and Bauplans. *Journal of Mammalogy*, 76, 322–334.

Krebs, C. J. (1972). *Ecology: The Experimental Analysis of Distribution and Abundance/by Charles J. Krebs.* 4th edn. Harper & Row, New York.

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.

McCain, C. M. & Knight, K. B. (2013). Elevational Rapoport's rule is not pervasive on mountains. *Global Ecology and Biogeography*, 22, 750–759.

Minchin, P. R. (1987). An evaluation of the relative robustness of techniques for ecological ordination. *Vegetatio*, 69, 89–107.

Neuschulz, E. L., Merges, D., Bollmann, K., Gugerli, F. & Böhning-Gaese, K. (2018). Biotic interactions and seed deposition rather than abiotic factors determine recruitment at elevational range limits of an alpine tree. *Journal of Ecology*, 106, 948–959.

Norberg, A., Abrego, N., Blanchet, F. G., Adler, F. R., Anderson, B. J., Anttila, J., Araújo, M. B., Dallas, T., Dunson, D., Elith, J., Foster, S. D., Fox, R., Franklin, J., Godsoe, W., Guisan, A., O'Hara, B., Hill, N. A., Holt, R. D., Hui, F. K. C., Husby, M., Kålås, J. A., Lehikoinen, A., Luoto, M., Mod, H. K., Newell, G., Renner, I., Roslin, T., Soininen, J., Thuiller, W., Vanhatalo, J., Warton, D., White, M., Zimmermann, N. E., Gravel, D. & Ovaskainen, O. (2019). A comprehensive evaluation of predictive performance of 33 species distribution models at species and community levels. *Ecological Monographs*, 89, e01370.

Normand, S., Treier, U. A., Randin, C., Vittoz, P., Guisan, A. & Svenning, J.-C. (2009). Importance of abiotic stress as a range-limit determinant for European plants: Insights from species responses to climatic gradients. *Global Ecology and Biogeography*, 18, 437–449.

Ovaskainen, O., Tikhonov, G., Norberg, A., Blanchet, F. G., Duan, L., Dunson, D., Roslin, T. & Abrego, N. (2017). How to make more out of community data? A conceptual framework and its implementation as models and software. *Ecology Letters*, 20, 561–576.

Pollock, L. J., Morris, W. K. & Vesk, P. A. (2012). The role of functional traits in species distributions revealed through a hierarchical model. *Ecography*, 35, 716–725.

Pollock, L. J., Tingley, R., Morris, W. K., Golding, N., O'Hara, R. B., Parris, K. M., Vesk, P. A. & McCarthy, M. A. (2014). Understanding co-occurrence by modelling species simultaneously with a Joint Species Distribution Model (JSDM). *Methods in Ecology and Evolution*, 5, 397–406.

Ribas, C. R. & Schoereder, J. H. (2006). Is the Rapoport effect widespread? Null models revisited. *Global Ecology and Biogeography*, 15, 614–624.

Rohde, K. (1992). Latitudinal Gradients in Species Diversity: The Search for the Primary Cause. *Oikos*, 65, 514–527.

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

Stan Developent Team (2021). RStan: The R interface to Stan.

Stan Development Team (2021). Stan Modeling Language Users Guide and Reference Manual.

Stevens, G. C. (1992). The Elevational Gradient in Altitudinal Range: An Extension of Rapoport's Latitudinal Rule to Altitude. *The American Naturalist*, 140, 893–911.

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. *Physical Review E*, 99, 032307.

ter Braak, C. J. F. & Looman, C. W. N. (1986). Weighted averaging, logistic regression and the Gaussian response model. *Vegetatio*, 65, 3–11.

Thurman, L. L., Barner, A. K., Garcia, T. S. & Chestnut, T. (2019). Testing the link between species interactions and species co-occurrence in a trophic network. *Ecography*, 42, 1658–1670.

Warton, D. I., Blanchet, F. G., O'Hara, R. B., Ovaskainen, O., Taskinen, S., Walker, S. C. & Hui, F. K. C. (2015). So Many Variables: Joint Modeling in Community Ecology. *Trends in Ecology & Evolution*, 30, 766–779.

Wisz, M. S., Pottier, J., Kissling, W. D., Pellissier, L., Lenoir, J., Damgaard, C. F., Dormann, C. F., Forchhammer, M. C., Grytnes, J.-A., Guisan, A., Heikkinen, R. K., Høye, T. T., Kühn, I., Luoto, M., Maiorano, L., Nilsson, M.-C., Normand, S., Öckinger, E., Schmidt, N. M., Termansen, M., Timmermann, A., Wardle, D. A., Aastrup, P. & Svenning, J.-C. (2013). The role of biotic interactions in shaping distributions and realised assemblages of species: Implications for species distribution modelling. *Biological Reviews*, 88, 15–30.