Analysis on Climate and Biodiversity Suitability for Spanish Dehesa Species

Eric Bataller, Danilo Méndez

December 22, 2021

1 Introduction

As climate change is affecting the way we live all across the globe it has become more and more important to focus on solutions or mitigation for the well being of humans and the planet that we inhabit. One of such solutions is the use of regenerative agricultural practices [4] to mitigate the effects on climate change on the food chain while restoring fertility and being a sustainable enterprise. One of these regenerative agricultural practices is the Dehesa, which is a silvopasture production system in Spain and Portugal where a system of naturally occurring leguminous trees (*Fagaceae*) and grasses (*Poaceae*) are maintained by grazing animals and human intervention. The dehesa allows for a great deal of biodiversity and production form different sources such as food from grain, tree crops and animal crops or wood and cork from trees.

Given this sustainable system that can also help mitigate climate change, it is important to understand more about it. In the first part of this project we will try to understand the underlying model that "truly" affects the presence of the most representative species in Spanish dehesas [6]. We used occurrence data to model species presence (1 for species present in a section and 0 for a species not present in a section) in certain sections of Spain that are determined by bio-climatic variables and elevation. By using Bayesian Model Selection and Bayesian Model Averaging we got a deeper view on what "truly" affects the presence of certain species in sections of Spain.
Lastly, in the second part of this document, we further analyse the presence of these plants across different regions in Spain to seek for biogeographical patterns that can turn out useful both, from a biodiversity analysis standpoint, and from a practical perspective for someone looking to build a Dehesa. Latent Dirichlet Allocation (LDA) was used on spatial data, revealing gradual changes in community structure by delineating overlapping groups of species.

The analysis can help people understand if their land has a positive or negative effect for species in a dehesa and what kind of dehesa is appropriate for the region.

2 Related Work

2.1 Part 1 - Model Selection

A lot of work has been done in ecosystem suitability. This predicts the presence of a species in a specific region given occurrence data [11] and in some cases, where better data is at hand using absence data as well. Other researchers have indicated that occurrence only data does not imply species abundance [1] which makes the estimation of population numbers meaningless without the absence data. Others have avoided that approach and for tree species have used cover change over time to estimate the ecological health of a region. These are all used in the estimation of habitat suitability or ecosystem health, two subjects that are very appropriate for dehesa modelling. We wanted to understand what bio-climatic variables have a higher impact on species presence, meaning that we are not interested in regression, like the articles have mentioned, but in model selection.

2.2 Part 2 - Topic Modelling

Understanding how species composition varies across space and time is fundamental to ecology. While multiple methods have been created to characterize this variation through the identification of groups of species that tend to co-occur, most of these methods unfortunately are not able to represent gradual variation in species composition (e.g., k-means, hierarchical clustering, network methods, and model-based approaches). To our knowledge, there are only two papers that make use of LDA for biodiversity purposes, and they are both relatively new ([9], [10]).

3 Data sets

The data used for analysis comes from 2 data sets accessed through an API using the *rgbif* package. GBIF (Global Biodiversity Information Facility) is a repository of databases of occurrences from different sources (mainly researchers and public institutions). Form this page we accessed two occurrence databases:

1. Tercer Inventario Forestal Nacional (IFN3)[7] - *Fagaceae*
2. CSIC-Real Jardín Botánico-Anthos. Sistema de Información de las Plantas de España [8] - *Poaceae*

We chose these occurrence data sets because they had research grade observations of our species of interest and were recorded at approximately the same time period; 1997-2007 for the IFN3 and 1990 to today for the CSIC-Real Jardín Botánico. These are long-term programs that monitors the status and trend of plant populations in Spain. In brief, data on the occurrences of species are collected throughout each decade by trained participants along randomly established forest areas, keeping track of the forest inventory.

3.1 Additional data for part 1 - Model Selection

On top of this we accessed the *worldclim* [3] API to obtain the 19 bio-climatic and elevation variables for each observation. The bio climatic variables and elevation variables are uniform for a range of 10 arc minutes, for Spain's latitude this is approximately 13x13km^2 sections.

Once we merged the 19 bio-climatic and elevation variables to the occurrence data via the coordinates, we obtained a large data set of 176,317 rows and 24 columns. We grouped the data by these 13x13km^2 sections and obtained a column for species presence in the given section, this is how we obtained our response variable. We reduced the number of predictor variables to 18 which corresponds to the 17 bio-climatic variables and one for the elevation.

3.2 Additional data for part 2 - Topic Modelling

The polygon data of the different administrative units of Spain required in part 2 of this document can be downloaded from the spanish National Center for Geographic Information, in the area called "Líneas límite municipales" [2].

4 Part 1 - Model Selection

For the first part of this project we are interested in understanding what bio-climatic factors *"truly"* affects the presence of a species in a given section of Spain. That is we don't want to predict or regress, but try to find the underlying model for species presence taking into account only climate and elevation. In [5] the method they used for obtaining probability of presence was using occurrences and background data. This is locations of known sightings and locations where the species are not known to occur. This way, only having the data for the occurrences we can construct the background data from the "missing" occurrences in all sections.

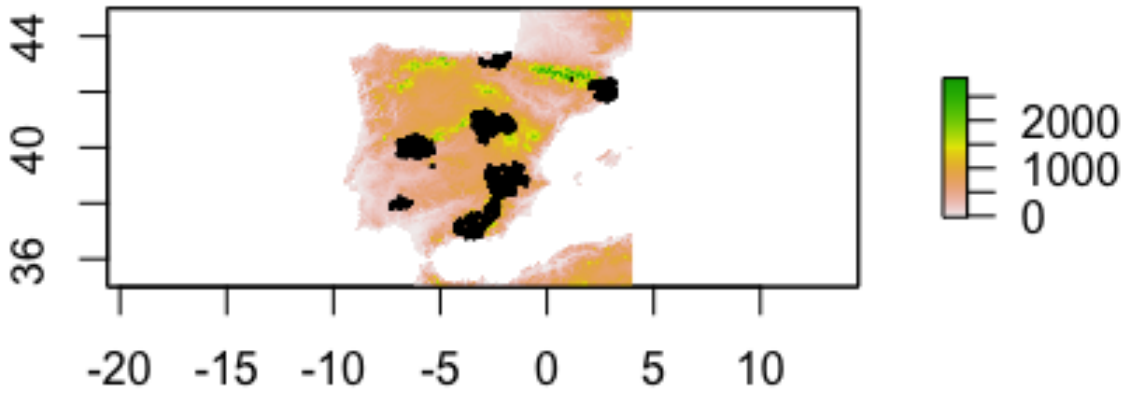


Figure 1: Elevation Plot of 1500 *Fagaceae* occurrences

We now have the presence and background data for each species and for each section. We differ from other strategies in the sense that we are not looking to predict presence of a species given certain parameters. We are interested in understanding which parameters have an effect and how can it be interpreted. For this we use 2 methods the Bayesian Model Selection (BMS through its connection with the BIC) to understand which variables have "true" effects and then we use Bayesian Model Averaging (BMA) to estimate parameters and interpret these effects.

Given the vast amount of data that we have on many different species, a truly in depth analysis of each species presence is unfeasible for this project. This is why we reduced our analysis approach to just the species that have at least 20% presence (that is, the species are present in at least 400 sections). This reduces our analysis space to just 12 different species, of which 4 are *Fagaceae* (legume trees) and 8 are *Poaceae* (grasses).

For the analysis we first need to state the family from which the response variable belongs to. Since the response variable can only be $y \in \{0, 1\}$ and it can take the value 1 with probability θ where θ depends on the data. This makes y a binomial random variable. It is now very straight forward to use the *mombf* package to use the model selection function on the 12 species and get the BMS and BMA for each. Before we present the results we will provide context for the two methods we will use for analysis.

4.1 Bayesian Model Selection

We wish to obtain the best possible model given the observations. For this we allow a positive probability that every one of our parameters is zero. We can set priors on models (γ) and choose the one with the highest posterior mode (highest posterior probability). The result is the choice of variables that are non zero such that the posterior probability is the highest possible. This means that the choice for model prior is important. We chose the binomial prior (since d << n) for this analysis which has the property that choosing the highest posterior probability is asymptotically equivalent to choosing the best model γ with the best BIC_γ.

We applied this to all 12 species and we got the following results:

Species	γ Vector
Arrhenatherum elatius	[1, 1, 1, 1, 1, 1, 0, 1, 1, 1, 1, 1, 0, 1, 1, 1, 1, 1]
Brachypodium sylvaticum	[1, 1, 1, 1, 1, 1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1]
Bromus hordeaceus	[1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1, 1, 1, 1, 0]
Dactylis glomerata	[1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1]
Helictochloa bromoides	[1, 1, 1, 1, 1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1]
Holcus lanatus	[1, 1, 1, 1, 1, 1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1]
Hordeum murinum	[1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1, 1, 1, 1, 0]
Poa bulbosa	[1, 1, 1, 1, 1, 1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1]
Quercus coccifera	[1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1]
Quercus faginea	[1, 1, 1, 1, 1, 0, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1]
Quercus pyrenaica	[1, 1, 1, 1, 1, 1, 0, 1, 1, 1, 1, 1, 0, 1, 1, 1, 1, 1]
Quercus rotundifolia	[1, 1, 1, 1, 1, 1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1]

For the above table we have that the gamma vector corresponds to the following variable vector where a 1 in any position of γ corresponds to that variable being chosen by Bayesian Model Selection in the following order.

- (Annual Mean Temperature, Mean Diurnal Range, Temperature Seasonality, Max Temperature of Warmest Month, Min Temperature of Coldest Month, Mean Temperature of Wettest Quarter, Mean Temperature of Driest Quarter, Mean Temperature of Warmest Quarter, Mean Temperature of Coldest Quarter, Annual Precipitation, Precipitation of Wettest Month, Precipitation of Driest Month, Precipitation Seasonality, Precipitation of Warmest Quarter, Precipitation of Wettest Quarter, Precipitation of Driest Quarter, Precipitation of Coldest Quarter, Elevation)

For example: for Hordeum murinum Elevation is not a relevant variable for the model with the highest posterior probability. We can see that the vectors are quite dense, meaning that they don't contain many zeros. This leads us to believe that almost all variables are have a "true" effect on the presence of the corresponding species. We also confirm that the Gibbs sampling converges for both model size and log-posterior model probabilities as seen in the Appedix.

4.2 Bayesian Model Averaging

We want to interpret the meaning behind our results, we know that some variables "truly" affect the presence of species, but we also want to know *how* it affects the presence. For this we can use point estimates which is very direct. We take the posterior under each possible model γ, and weight them based on the posterior probability $p(\gamma|y)$ of that model given the data y. We get an immediate result for each parameter:

$$E(\beta|y) = \sum_{\gamma} E(\beta|\gamma, y)p(\gamma|y) \tag{1}$$

We can also obtain quite easily the lower and upper bound:

$$P(\beta_j \in [l, u]\,|y) = 0.95$$

We have done this for all 12 species and here are the results.

We can see from figures [2, 3 and 4]that the results are very different from the picture presented by BMS. Most species have many β_j that are very close to zero while other show promise of being far from zero. This gives us a better picture of the true underlying model since we have averaged many different models to obtain the β_j and the confidence intervals. The main takeaway from these graphs is that the true effect of some of these variables is effectively zero for all the species in question.

It is also very clear that variables 1, 6 and 8 have a big impact, sometimes positive and sometimes negative on all species presence. These variables correspond to Annual Mean Temperature, Mean

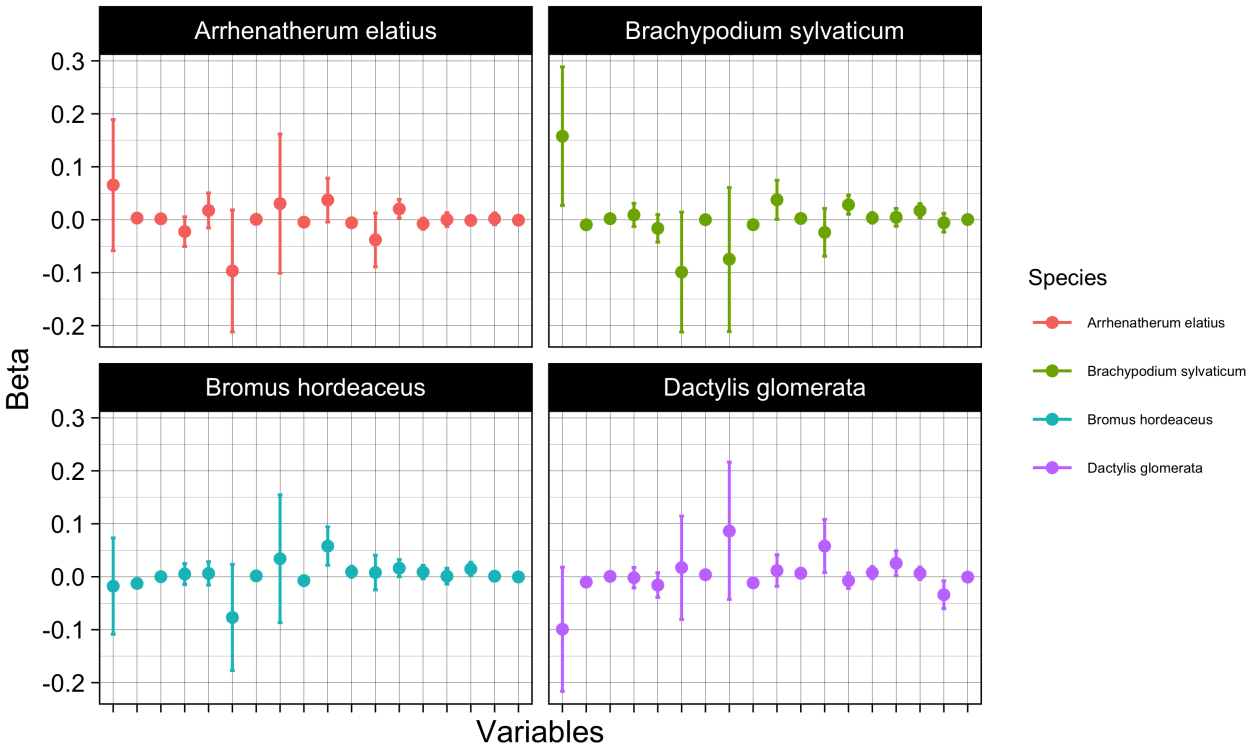


Figure 2: Bayesian Model Averaging with confidence intervals for 4 grass species

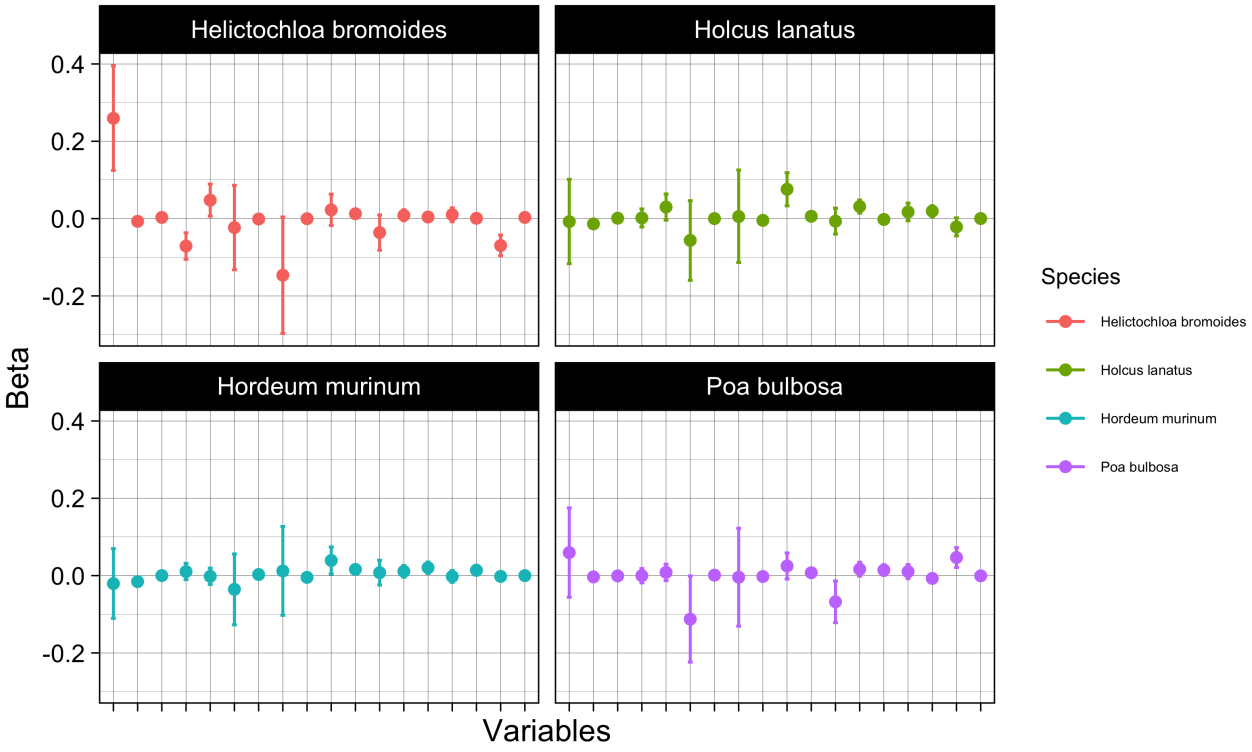


Figure 3: Bayesian Model Averaging with confidence intervals for 4 new grass species

Temperature of Wettest Quarter and Mean Temperature of Warmest Quarter (the same representation as in the previous subsection) which leads us to believe that temperature plays a very key role.

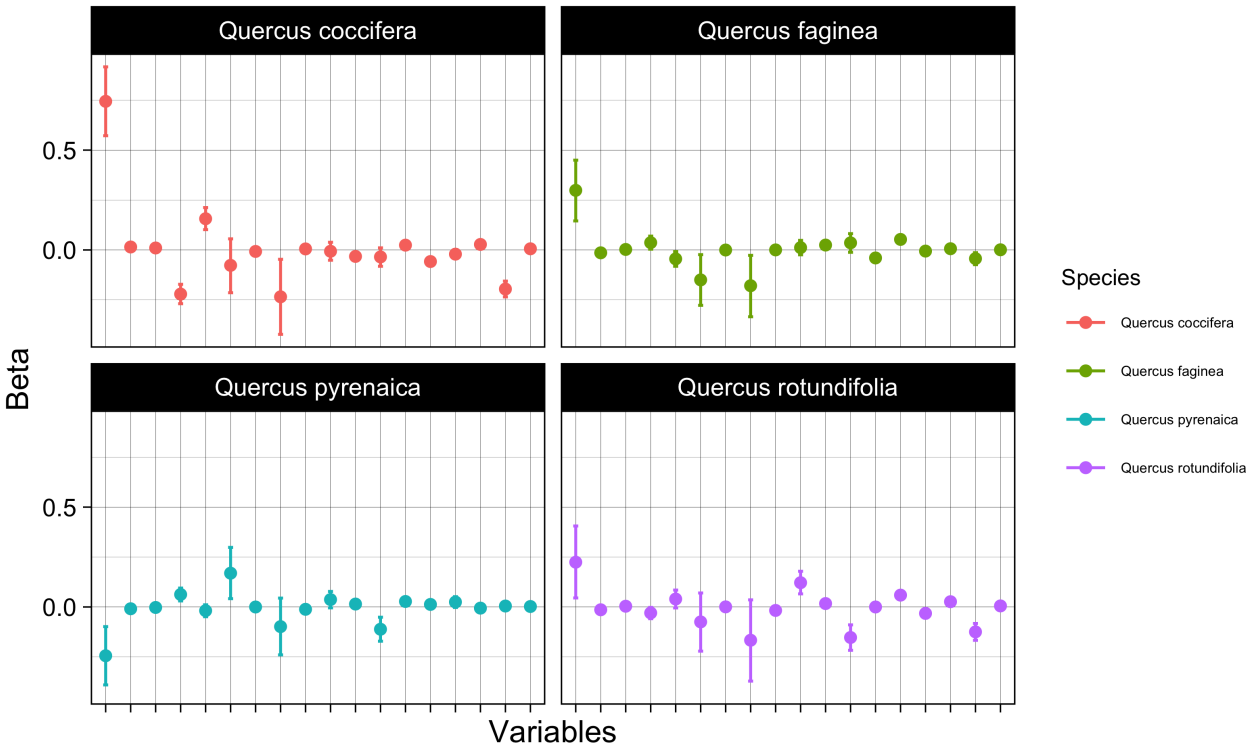


Figure 4: Bayesian Model Averaging with confidence intervals for 4 tree species

We interpret positive values of β as the bigger the positive β the stronger influence the variable in question has on species presence. In the case of temperature, the higher the mean annual temperature (for some species) the more likely a species is present. Analogously the lower the temperature the less likely a species is present (though it would still have a positive effect, just not as large). We also find that elevation does not have a big impact in general which is unexpected since we think of elevation as a determinant factor in other types of species distribution, although elevation is highly correlated with temperature.

More insights could probably be extracted by a agronomist, farmer or agricultural extension officer that has knowledge on how to implement the findings in this section. These are meant to be tools for people in the regenerative agriculture industry to use them to improve on their knowledge and extract expert insights that only experts in their field could know.That is why this work is limited only to model selection, giving the professionals a tool to see which climatic factors matter the most for the region that they are in and their choice of dehesa.

5 Part 2 - Topic Modelling

5.1 Model description

The overall goal of our method is to identify the major patterns of species co-occurrence in the data, each of which we define to be a distinct bioregion (given that these major co-occurrence patterns have to have a strong spatial pattern in the case of trees and other plants). These groups can overlap in space, and proportion of groups can change through time. This would be the equivalent to a topic in classic LDA applied to Natural Language Processing. More specifically, our method characterizes each sampling unit j in terms of the proportion of the different bioregions (parameter vector θ_j) and characterizes each group k in terms of the probability of the different species (parameter vector β_k). For example, given a location j, a vector $\theta_j = [0.2, 0.8, 0.0]$ indicates that the second bioregion/topic dominates unit j and that the third group is absent. This example also illustrates that a given sam-

pling unit can be a combination of multiple groups, which explains why we are able to model gradual variation between regions. In the same way, $\beta_k = [0.0, 0.3, 0.7]$ indicates that species 2 and 3 (but not species 1) are important species of group k. Note that a given species can have a high probability in more than one group.

Just like in classic LDA, our data consists of a corpus of sample units (documents), made of all the different occurrences (the words) that were registered in that specific region. The variables x_i are the occurrences in the sample units and they are the observable variables, corresponding each of them to a specific species. Latent Dirichlet allocation is a generative probabilistic model of a corpus. The underlying assumption of our model is that sample units are represented as random mixtures over K latent bioregions (topics), where each bioregion is characterized by a distribution over V words.
We denote by $Dir_V(\eta)$ the Dirichlet distribution of dimension V with parameter vector $\eta = (\eta_1, ..., \eta_V)$ and $Dir_V(\alpha)$ is the Dirichlet distribution of dimension K with parameter vector $\alpha = (\alpha_1, ..., \alpha_K)$. As it's commonly done, a symmetric Dirichlet distribution was used, setting all $\eta = 1/K$ and $\alpha = 1/V$. The reason is that we don't have any prior knowledge on the importance of each species across each group, nor do we know anything about the importance of each bioregion in Spain. Nonetheless, we highlight that this could be an interesting path to exploit for those that don't have a lot of data on the occurrences of different species and can provide some prior insight on the structure of the Flora (and Fauna) of their region of study (for example, one could argue that for a continental region elephants are not very commonly seen and therefore $alpha_{elephant}$ could be priorly set to a small value). This would hopefully make the estimated distribution more accurate given that we provide useful insight.

We now have all the ingredients to devise our generative model for any corpus x:

1. For each bioregion $k = 1, ..., K$:
   (a) Draw a distribution of species $\beta_k \sim Dir_V(\eta)$
2. For each sample unit $i = 1, ..., n$:
   (a) Draw a random vector of bioregions (topics) proportions $\theta_i \sim Dir_K(\alpha)$.
   (b) For each occurrence $j = 1, ..., d_i$:
      i. Draw a bioregion (topic) assignment $z_{ij} \sim Multinomial(\theta_i)$, $z_{ij} \in 1, ..., K$
      ii. Draw a species $x_{ij} \sim Multinomial(\beta_{z_{ij}})$, $x_{ij} \in 1, ..., V$

A graphical representation of such model can be found in 5.

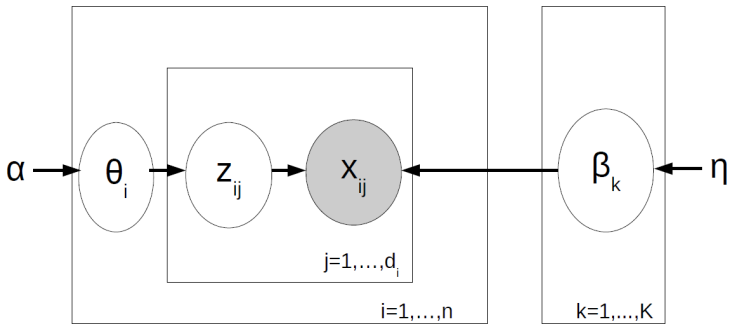


Figure 5: Graphical representation of the LDA model

5.2 Analysis

In an era of global change, understanding the biodiversity of each region and its dynamics across time and spatial gradients is key for the control of species and the efficiency of regenerative agricultural practices. This understanding can be partially provided by LDA as we demonstrate in the following

section. More specifically, we assessed which species of fagacea and poaceae tend to coexist together, and therefore make a good fit as a "cluster", given that they share similar characteristics in terms of the type of bioregion they tend to live in. The analysis is based on the occurrences data available for each municipality in Spain. Such granularity is desired to detect sudden changes in the biodiversity data (e.g changes associated with tall mountains or very dry areas), but it also comes at a risk of reducing the available occurrences to a point in which they are not representative of the true species distribution (a illustration of the granularity of our analysis can be seen in Fig.6). This method generates biologically interpretable results because it decomposes each sampling unit (municipality) into distinct component communities; and characterises each of these component communities in terms of the relative abundance of species. Furthermore, the model might adequately represents the uncertainty associated with its estimates given that it's quite good at capturing the gradual variation across regions.
An important feature of our method is that it is able to detect relatively subtle temporal changes in species composition. Although our purpose throughout this project was merely exploratory, the periodic repetition of such analysis across time would accurately reveal pattern changes in the dynamics of each bioregion.

To detect the patterns, we map the coordinates of each occurrence to the belonging municipality, these form our documents from which we are able to count the abundance of each species and then we fit the model. We set the maximum number of groups to 8 for our case study as it seems to be a good amount so that the algorithm can learn meaningful structures within the species communities.

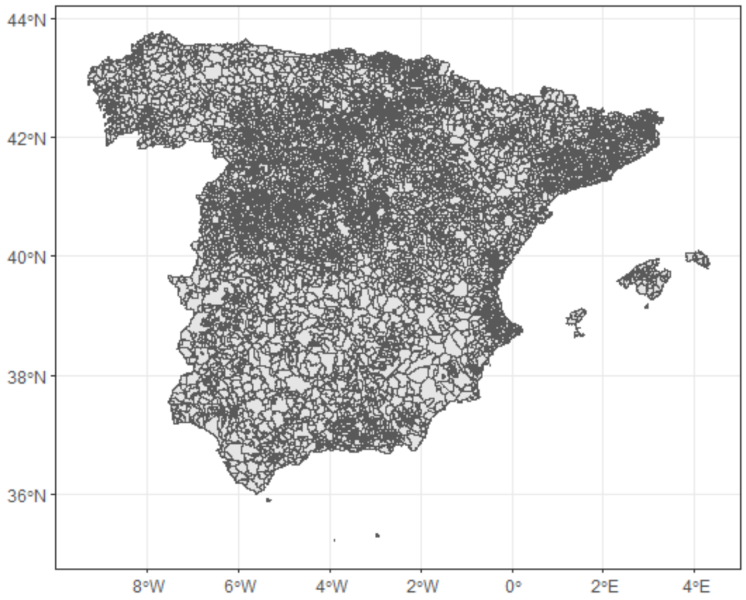


Figure 6: Map of Spain delimited by municipalities (sample units).

5.3 Results

Our results reveal that the algorithm accurately estimates the proportion of the different groups in each location. An illustration of the distribution of a specific bioregion (number 6, to be precise) across Spain can be seen in Fig.7
Overall, we identified 4 main plants groups (of a maximum of 8) after eliminating groups that were very common throughout all regions (Fagaceae Quercus Rotundifolia specifically, given its robustness across multiple climates). An important test for any unsupervised method is if it is able to retrieve patterns that are widely acknowledged to exist by experts. Going back to bioregion number 6, we observe strong presence around the Galicia zone (showing strong spacial correlation that the algorithm as able to learn even when no spatial data was provided). Further more, we see that the decrease in proportion across space is relatively smooth (contrary to the result one could obtain by means of

hard clustering algorithms). If we further inspect the composition of bioregion number 6 (Appendix B.1), we will find that the top species that define it are greatly related in terms of the ecosystems they usually live in. For example, Quercus robur, Castanea sativa, Fagus sylvatica share significant requirements: a humid atmosphere (precipitation well distributed throughout the year and frequent fogs) and well-drained soil (it cannot handle excessive stagnant water).

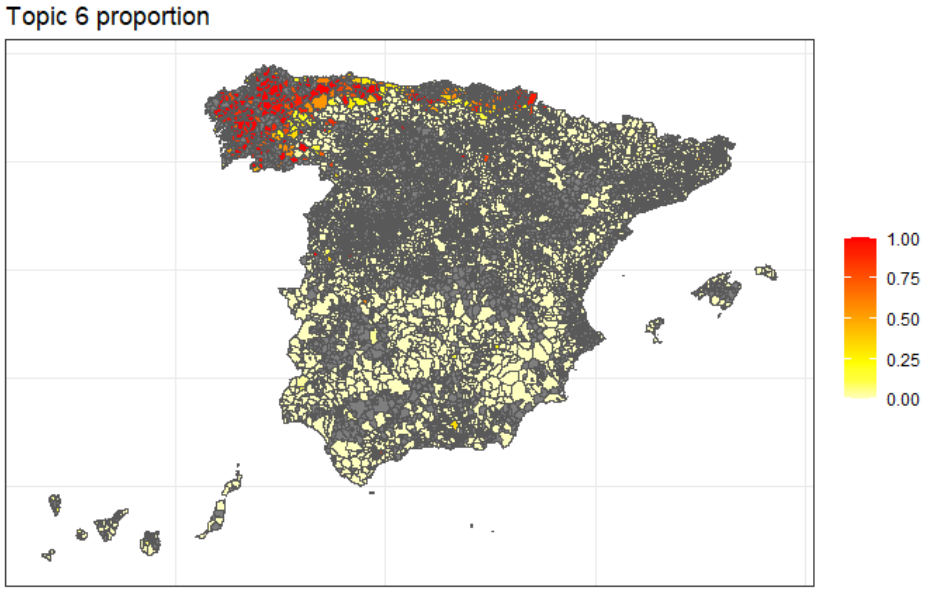


Figure 7: Distribution of topic (bioregion) number 6 across Spain's municipalities. Note that the grey areas are those municipalities for which there's no data available.

Similar insights can be gathered by looking at the rest of bioregions (Appendix B.2):

- Quercus suber, Quercus coccifera and Quercus canariensis (also known as Andalusian oak) all have low soil demands and also grow in poor, dry or rocky locations. Consequently, they were assigned together in bioregion number 5, and are mainly located in the southern part of Spain.

- Quercus faginea, Quercus coccifera and other Poaceae species typicaly live in Mediterranean type of forests, and thus, they were assigned together at bioregion number 7.

- Quercus pyrenaica and Deschampsia flexuosa are famous for being able to exist over 1,200 metres above sea level. They have a preference for acidic, free-draining soil, and avoids chalk and limestone areas. They were grouped together in bioregion number 8, were we can see they are predominant in the Cantabrian mountains and the Pyrinees.

The rest of groups don't seem to be predominantly concentrated around any area, but rather spasly across the territory (in the case of bioregion number 2 we can't really observe the true distribution, as most of the predominant area seem to be missing due to a lack of data).

5.4 Discussion

Here, we have substantially developed the standard LDA model to enable the analysis of presence/absence data and we have demonstrated that novel insights can be gained using our method when applied to a continental-scale wildlife dataset. An important limitation of the method that we have presented is that the proposed model does not take into account imperfect detection, a pervasive issue for wildlife sampling (specially outside the plants kingdom), and occurrences data where observations might correlated with the populated density of the regions. Unfortunately, we were limited by the scarcity of

our data. Although we initially had more than 170k occurrences, most of them were gathered around certain key areas, leaving huge empty holes across all territory (as it is the case of Castile and León). Thus, it's important to have a minimum sample size available across all territory in order to accurately estimate the presence of different communities and the species structures on those.

6 Conclusion and future work

6.1 Part 1

For the first part of the analysis we used Bayesian Model Selection and Bayesian Model Averaging. Of these two it is clear that BMA was more efficient in demonstrating the magnitude of the effect and the number of variables that are very close to zero. BMA had a more conservative choice of variables that are not zero, which related to temperature in all species. We believe that this is because BMS only chooses the mode as the best model, if the mode is not close to 1 then there could be other very plausible models that the BMS ignores. BMA fixes this problem. For future work we would have gotten more occurrences from more reputable sources to get a finer grain definition for the sections of Spain. Instead of 10 arc minutes a 30 arc seconds could provide more variability. Other articles mentioned adding more covariates like distance from closest human settlement, distance from closest road, soil type, orientation and many more. The quality of the data type is of great importance, adding absence data could lead to dramatic improvements on the real model at work. Future attempts should include more variables and a wider range of data sources and data types.

6.2 Part 2

Topic modeling was used for the second part of this project. Meaningful insights can be drawn from the final classification and the conditional distribution of each region $p(z = k|x)$, proving that LDA can be applied successfully outside of the scope of Natural Language processing, and providing a key tool for mixed-membership analysis over spatial data.

The Latent Dirichlet Allocation model is a useful model for ecologists because it can more faithfully represent community dynamics and the impact of environmental change through the estimation of mixed-membership sites. The standard LDA requires abundance data but, for many taxa, reliably estimating abundance is often very hard and costly. For these reasons, presence/absence data are typically much more ubiquitous than abundance data, often enabling analysis at much larger spatial and temporal scales than that afforded by abundance data. LDA is based on a fully probabilistic generative model that allows for straightforward quantification of uncertainty. We didn't dive into the quantitative formulation of uncertainty (given that it's outside of the scope of this project), but it is an important characteristic of our model because it allows scientists to judge if the observed changes in species composition (e.g due to global change) are greater than the uncertainty associated with these results.

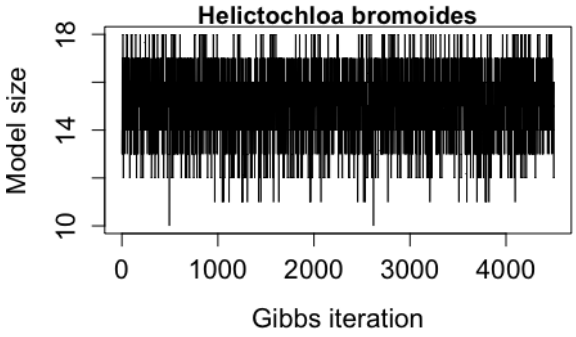
A Appendix A: Bayesian Model Selection Convergence Criteria

A.1 Model Size

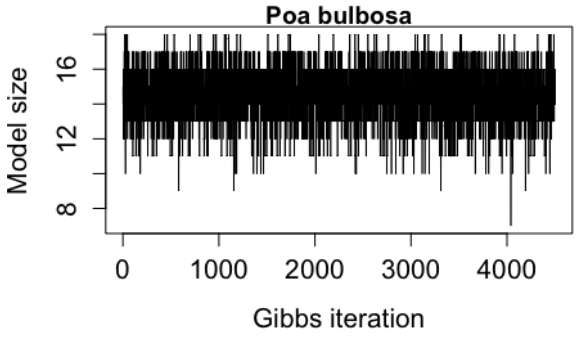
Here are all the trace of the number of variables at each MCMC iteration for a sample of species. The rest are in the R code.

A.2 Log-posterior Model Probabilities

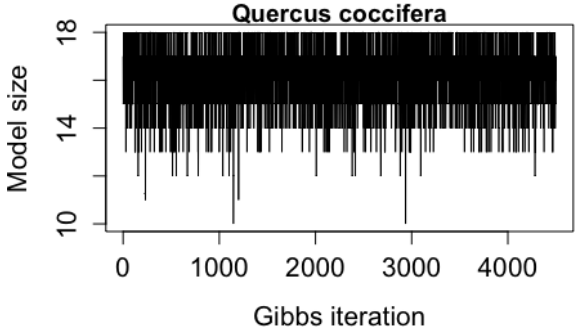
Trace plot for $log[p(y|\gamma)p(\gamma)]$ which is equal to log-posterior model probabilities $logp(\gamma y)$ up to a normalizing constant for a sample of species. The rest are in the R code.



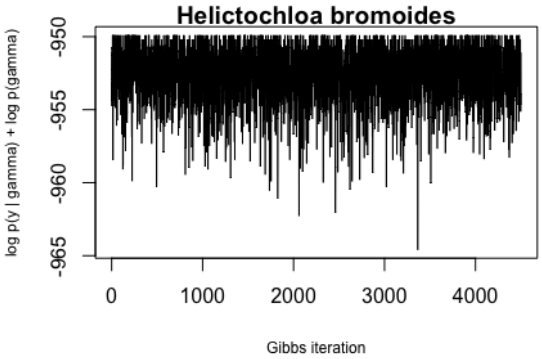
(a) Converges between 13 and 18



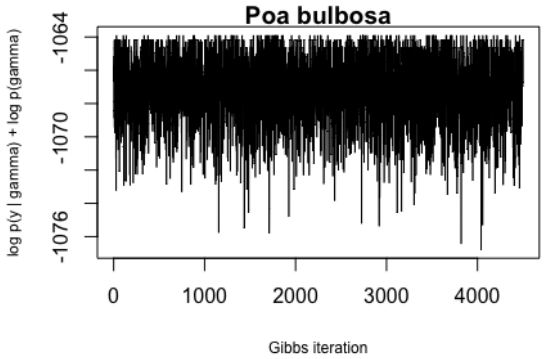
(b) Converges between 11 and 18



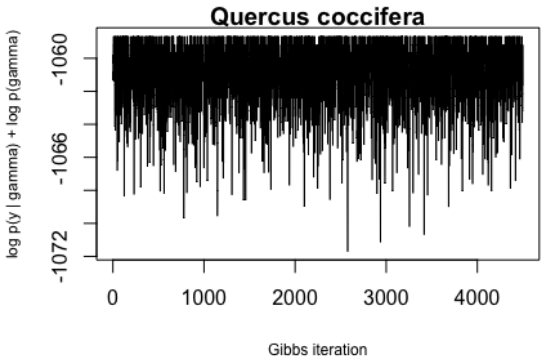
(c) Converges between 13 and 18



(a) Stabilized



(b) Stabilized



(c) Stabilized

	[,1]	[,2]	[,3]	[,4]
[1,]	"fagaceae-quercus-coccifera"	"fagaceae-quercus-pyrenaica"	"poaceae-dactylis-glomerata"	"fagaceae-quercus-robur"
[2,]	"fagaceae-quercus-faginea"	"poaceae-helictochloa-marginata"	"poaceae-brachypodium-retusum"	"fagaceae-castanea-sativa"
[3,]	"fagaceae-quercus-pubescens"	"fagaceae-quercus-faginea"	"poaceae-phragmites-australis"	"fagaceae-fagus-sylvatica"
[4,]	"fagaceae-quercus-petraea"	"poaceae-dactylis-glomerata"	"poaceae-brachypodium-distachyon"	"fagaceae-quercus-pyrenaica"
[5,]	"poaceae-avena-sativa"	"poaceae-glyceria-declinata"	"poaceae-hordeum-murinum"	"fagaceae-quercus-petraea"
[6,]	"poaceae-festuca-gracilior"	"poaceae-agrostis-castellana"	"fagaceae-quercus-coccifera"	"poaceae-arrhenatherum-longifolium"
[7,]	"poaceae-poa-compressa"	"poaceae-aira-caryophyllea"	"poaceae-cynodon-dactylon"	"poaceae-agrostis-curtisii"
[8,]	"poaceae-glyceria-notata"	"poaceae-arrhenatherum-elatius"	"poaceae-avena-barbata"	"poaceae-holcus-mollis"
[9,]	"poaceae-festuca-paniculata"	"poaceae-trisetaria-ovata"	"poaceae-bromus-hordeaceus"	"poaceae-agrostis-capillaris"
[10,]	"poaceae-achnatherum-calamagrostis"	"poaceae-holcus-lanatus"	"poaceae-agrostis-stolonifera"	"fagaceae-quercus-suber"
[11,]	"poaceae-festuca-marginata"	"poaceae-periballia-involucrata"	"poaceae-hyparrhenia-hirta"	"poaceae-holcus-lanatus"
[12,]	"poaceae-puccinellia-fasciculata"	"poaceae-bromus-hordeaceus"	"poaceae-brachypodium-phoenicoides"	"poaceae-agrostis-durieui"
[13,]	"poaceae-arrhenatherum-elatius"	"poaceae-brachypodium-sylvaticum"	"poaceae-catapodium-rigidum"	"poaceae-anthoxanthum-odoratum"
[14,]	"poaceae-festuca-gautieri"	"poaceae-poa-bulbosa"	"poaceae-lolium-rigidum"	"fagaceae-quercus-rubra"
[15,]	"poaceae-agrostis-castellana"	"poaceae-corynephorus-canescens"	"poaceae-bromus-rubens"	"poaceae-festuca-rubra"
[16,]	"poaceae-echinochloa-colona"	"poaceae-poa-trivialis"	"poaceae-bromus-diandrus"	"poaceae-brachypodium-sylvaticum"
[17,]	"poaceae-festuca-ovina"	"poaceae-agrostis-capillaris"	"poaceae-helictochloa-bromoides"	"poaceae-glyceria-fluitans"
[18,]	"poaceae-echinaria-capitata"	"poaceae-festuca-paniculata"	"poaceae-poa-bulbosa"	"poaceae-avenella-flexuosa"
[19,]	"poaceae-poa-pratensis"	"poaceae-avenella-flexuosa"	"poaceae-lolium-arundinaceum"	"poaceae-brachypodium-pinnatum"
[20,]	"poaceae-agrostis-stolonifera"	"poaceae-nardus-stricta"	"poaceae-briza-maxima"	"poaceae-paspalum-dilatatum"

	[,5]	[,6]	[,7]	[,8]
[1,]	"fagaceae-quercus-suber"	"fagaceae-quercus-pyrenaica"	"fagaceae-quercus-coccifera"	"fagaceae-fagus-sylvatica"
[2,]	"fagaceae-quercus-coccifera"	"fagaceae-quercus-faginea"	"fagaceae-quercus-faginea"	"poaceae-agrostis-capillaris"
[3,]	"fagaceae-quercus-faginea"	"fagaceae-fagus-sylvatica"	"poaceae-helictochloa-bromoides"	"poaceae-dactylis-glomerata"
[4,]	"fagaceae-quercus-canariensis"	"fagaceae-quercus-petraea"	"poaceae-brachypodium-retusum"	"poaceae-anthoxanthum-odoratum"
[5,]	"fagaceae-castanea-sativa"	"poaceae-festuca-hystrix"	"poaceae-festuca-hystrix"	"poaceae-briza-media"
[6,]	"fagaceae-quercus-pubescens"	"poaceae-helictochloa-bromoides"	"poaceae-helictotrichon-filifolium"	"poaceae-koeleria-vallesiana"
[7,]	"poaceae-agrostis-pourretii"	"poaceae-sesleria-argentea"	"poaceae-festuca-plicata"	"poaceae-avenella-flexuosa"
[8,]	"poaceae-brachypodium-sylvaticum"	"poaceae-agrostis-curtisii"	"poaceae-agrostis-nevadensis"	"poaceae-poa-alpina"
[9,]	"poaceae-gastridium-ventricosum"	"poaceae-stipa-offneri"	"poaceae-poa-ligulata"	"poaceae-festuca-gautieri"
[10,]	"poaceae-poa-bulbosa"	"poaceae-periballia-involucrata"	"poaceae-agrostis-nebulosa"	"poaceae-festuca-nigrescens"
[11,]	"poaceae-melica-minuta"	"poaceae-glyceria-fluitans"	"poaceae-festuca-capillifolia"	"poaceae-poa-nemoralis"
[12,]	"poaceae-agrostis-tenerrima"	"poaceae-stipa-atlantica"	"poaceae-piptatherum-paradoxum"	"poaceae-festuca-rubra"
[13,]	"poaceae-festuca-caerulescens"	"poaceae-piptatherum-paradoxum"	"poaceae-stipa-tenacissima"	"poaceae-bromus-erectus"
[14,]	"poaceae-glyceria-declinata"	"poaceae-alopecurus-myosuroides"	"poaceae-koeleria-vallesiana"	"poaceae-nardus-stricta"
[15,]	"poaceae-briza-maxima"	"poaceae-catabrosa-aquatica"	"poaceae-festuca-scariosa"	"fagaceae-quercus-petraea"
[16,]	"poaceae-festuca-ampla"	"poaceae-poa-bulbosa"	"poaceae-arrhenatherum-elatius"	"poaceae-arrhenatherum-elatius"
[17,]	"poaceae-agrostis-reuteri"	"poaceae-bromus-commutatus"	"poaceae-melica-minuta"	"poaceae-helictotrichon-sedenense"
[18,]	"poaceae-psilurus-incurvus"	"poaceae-milium-effusum"	"poaceae-stipa-offneri"	"poaceae-helictochloa-pratensis"
[19,]	"poaceae-poa-annua"	"poaceae-echinaria-capitata"	"poaceae-trisetum-velutinum"	"poaceae-holcus-lanatus"
[20,]	"poaceae-vulpia-fontquerana"	"poaceae-helictochloa-pratensis"	"poaceae-brachypodium-phoenicoides"	"poaceae-brachypodium-sylvaticum"

Figure 10: Top species that form each bioregion.

B Appendix B: Topic Modelling

B.1 Top species per bioregion

B.2 Topic distributions across Spain

References

[1] Tad A. Dallas and Alan Hastings. Habitat suitability estimated by niche models is largely unrelated to species abundance. *Global Ecology and Biogeography*, 27(12):1448–1456, 2018.

[2] Organismo Autónomo Centro Nacional de Información Geográfica. Centro de descargas del cnig (ign).

[3] S.E. Fick and R.J. Hijmans. Worldclim 2: new 1km spatial resolution climate surfaces for global land areas, 2017.

[4] Hannah Gosnell, Susan Charnley, and Paige Stanley. Climate change mitigation as a co-benefit of regenerative ranching: insights from australia and the united states. *Interface Focus*, 10(5):20200027, Aug 2020.

[5] Wenkai Li, Qinghua Guo, and Charles Elkan. Can we model the probability of presence of species without absence data? *Ecography*, 34(6):1096–1105, Feb 2011.

[6] Teodoro Marañon. Plant species richness and canopy effect in the savanna-like "dehesa" of s.-w. spain. 1986.

[7] Occdownload Gbif.Org. Occurrence download, 2021.

[8] Occdownload Gbif.Org. Occurrence download, 2021.

[9] Denis Valle, Benjamin Baiser, Christopher W. Woodall, and Robin Chazdon. Decomposing biodiversity data using the latent dirichlet allocation model, a probabilistic multivariate statistical method. *Ecology Letters*, 17(12):1591–1601, 2014.

[10] Denis Valle, Gilson Shimizu, Rafael Izbicki, Leandro Maracahipes, Divino Vicente Silverio, Lucas N. Paolucci, Yusuf Jameel, and Paulo Brando. The latent dirichlet allocation model with covariates (ldacov): A case study on the effect of fire on species composition in amazonian forests. *Ecology and Evolution*, 11(12):7970–7979, 2021.

[11] Guiming Zhang, A-Xing Zhu, Steve K. Windels, and Cheng-Zhi Qin. Modelling species habitat suitability from presence-only data using kernel density estimation. *Ecological Indicators*, 93:387–396, 2018.

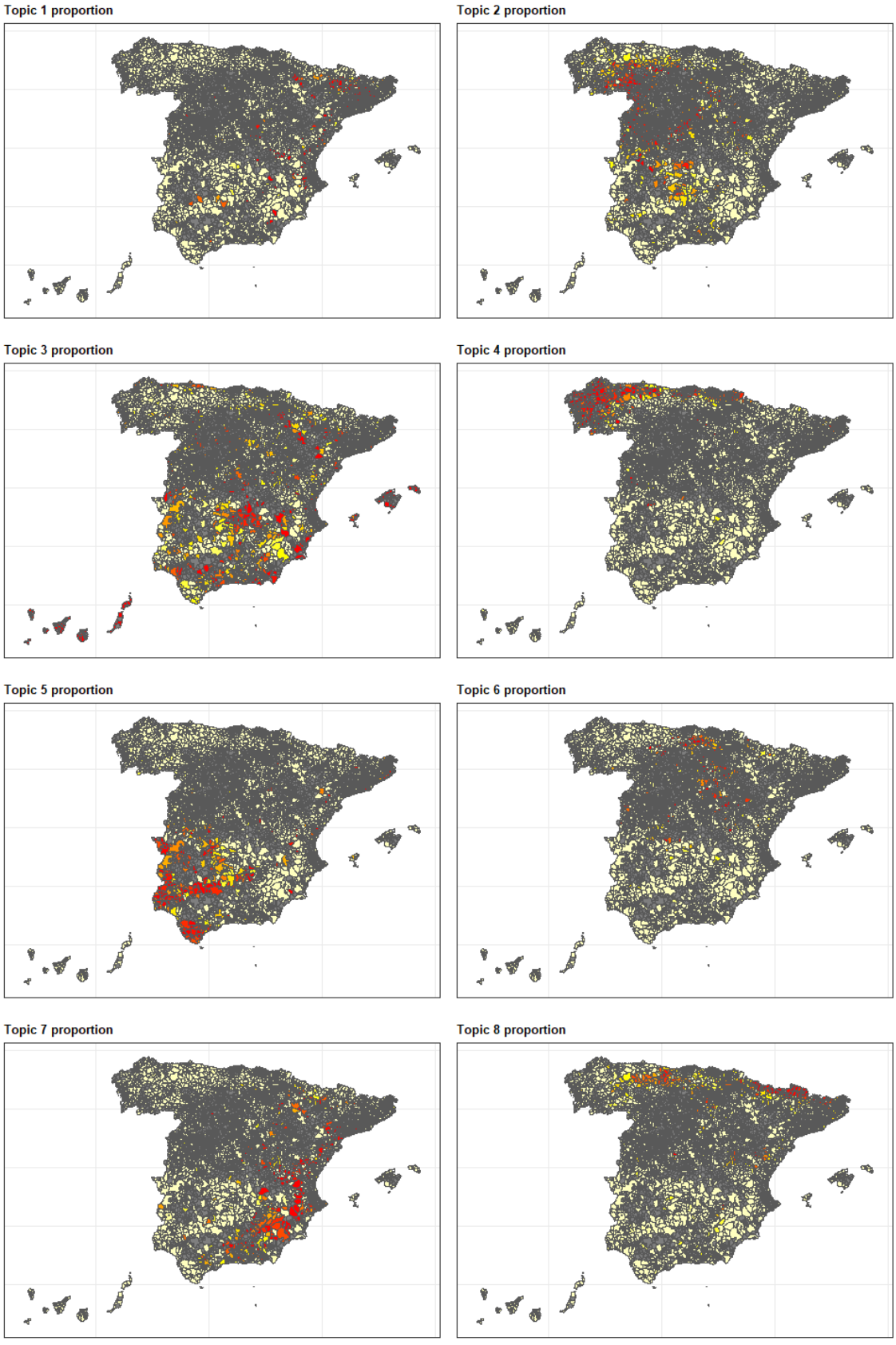


Figure 11: Topic distributions across Spain.