**Title:** A hierarchical Bayesian approach for modeling changes in species composition

Filipe S. Dias1,2, Michael Betancourt3, Patricia María Rodríguez-González4, Luís Borda de Água1,2

**Affiliations**

1. CIBIO/InBio, Centro de Investigação em Biodiversidade e Recursos Genéticos, Laboratório Associado, Universidade do Porto, Campus Agrário de Vairão, 4485-661 Vairão, Portugal

2. CIBIO/InBio, Centro de Investigação em Biodiversidade e Recursos Genéticos, Laboratório Associado, Instituto Superior de Agronomia, Universidade de Lisboa, Tapada da Ajuda, 1349-017 Lisbon, Portugal

3. Symplectomorphic, LLC., New York, USA

4. Centro de Estudos Florestais, Instituto Superior de Agronomia, Universidade de Lisboa, Tapada da Ajuda, 1349-017 Lisboa, Portugal

**Abstract**

Understanding the factors associated with changes in species composition is a critical issue in ecology. A typical modeling approach consists of calculating beta diversity indices between pairs of sites using beta diversity indices and assessing the relationship between those indices and environmental covariates. Beta diversity indices are paired comparisons, which means that indices calculated with the same sample are not independent. The most common solution is to transform community data and environmental covariates into distance matrices, perform row and column permutations, and calculate correlations (e.g., Mantel tests and Generalized Dissimilarity Modeling). Here we introduce a novel hierarchical Bayesian approach that deals with this dependence by including two varying intercepts, one per sample, that capture the heterogeneity corresponding to that sample. Our modeling approach is highly flexible and customizable, adding novel features that are not available in any existing software packages. It allows the relationship between beta diversity indices and covariates to change across different regions by including varying intercepts and slopes and provides a simple path for performing model validation and model improvement.

**Keywords:** Beta diversity; community similarity; Mantel tests; riparian vegetation; Sorensen index;

**1. Introduction**

Species composition changes across space due to environmental factors, neutral factors (e.g. ecological drift and random dispersal) or a combination of both (Hubbell 2001, Leibold et al. 2004, Gravel et al. 2006). Quantifying and understanding these changes is crucial for understanding ecological processes,for assessing human activities' impacts on natural ecosystems, and developing conservation plans (Legendre et al. 2005).

A common approach for modeling species composition changes consists of 1) calculating community similarity between pairs of sites using beta diversity indices (e.g. Sorensen index, Morisita Horn index, etc.) and 2) assess the relationship between those indices and environmental covariates. The main difficulty when analyzing this type of data is that community similarity indices are paired comparisons, which means that indices calculated with the same sample are not independent. The most common strategy for dealing with this is to transform community composition data and environmental covariates into distance matrices and using matrix correlations or regression techniques (Legendre et al. 2005). Mantel tests and partial Mantel tests (Mantel 1967) are some of the most popular techniques that employ this approach. The Mantel test examines the linear correlation between two distance matrices, a community similarity matrix and a covariate distance matrix. The resulting Mantel statistic can be regarded as a correlation coefficient, but its significance is assessed by randomly permuting the rows' order and columns of one matrix. When using large datasets, it is possible to transform the Mantel statistic into a t-statistic, whose significance is obtained from an asymptotic approximation of the t-test. The partial Mantel test is an extension of the original Mantel test, where a third matrix (or more) is held constant while the relationship between the other two is determined (Smouse et al. 1986). This test is performed by regressing the first two matrices' elements onto the third matrix and using the residuals from the regressions as the input for the standard Mantel test (Legendre 2000). Mantel tests' main drawback is the assumption of a linear relationship between community similarity and covariates (Legendre et al. 2015). In fact, the use of Mantel tests has been widely criticized in the literature, with studies suggesting that the test can produce biased or invalid estimates in a wide range of scenarios (e.g., Guillot and Rousset 2013). Ferrier et al. (2007) developed the Generalized Dissimilarity Modeling (GDM), which is a technique that can accommodate non-linear relationships between changes in community similarity and covariates, as well as variation in the rate of compositional turnover at different positions along environmental gradients. This technique is a form of a linear (matrix) regression between all possible community similarity measures and covariates that uses random permutations to deal with the dependence in the data (Manly 1986, Legendre et al. 1994).

Here we introduce a hierarchical Bayesian approach that matches the capabilities of both Mantel tests and GDM and adds a few novel key features. We provide a detailed explanation of our approach and present a case study using riparian plant communities. Since Bayesian methods may not be familiar to all readers, we also provide a step by step guide for running the analysis and all the required codes (Appendix 1).

**2. Description of the Method**

**2.1 General overview**

We start by providing a general introduction to our method by using linear regression as a starting point. Suppose we are modeling the effects of a variable *C* on another variable *S*. The traditional linear regression assumes that *S* is normally distributed and that there is a linear relationship between the expected value of *S* and *C*:

*S* ~ Normal (***μ***, ***σ***)

***μ*** = ***α*** + ***β*.*C***

where *μ* and *σ* are the mean and the standard deviation of the normal distribution and α and β are the intercept and slope of the regression. One of the assumptions of regression models is that observed *S* values are independent of each other for fixed covariate values. When *S* values result from paired comparisons (e.g., number of species in common, genetic distance), this assumption does not hold. This happens because multiple *S* values are calculated using the same sample, which creates dependence among them. Our approach consists of adding a varying intercept *αs* for each of the samples that were involved in the calculation of *S*. Each *αs* captures the heterogeneity of the corresponding sample by taking the same value whenever that sample is used to calculate *S*. The model is, then:

*S* ~ Normal (***μ***, ***σ***)

***μ*** = ***α*** + ***αs***[sample1] + ***αs***[sample2]  + ***β****.****C***.

As with any regression model care is required with this approach. If we fit a model with just ***μ*** = ***α*** + ***αs***[sample1] + ***αs***[sample2]  the *αs* intercepts will absorb any variation associated with the observed samples, including confounding variation that is not related to the samples themselves, but rather to how the samples were selected during measurement. In order to ensure that ***αs***[sample1] and ***αs***[sample2] capture only the behavior common to each sample, and can actually be interpreted as sample parameters, we need to explicitly model any other systematic variations that might be present by adding other intercepts and slopes to the regression model.

**2.2 Model extensions**

**2.2.1 Change the response’s variable distribution**

We assumed that S followed a normal distribution in the previous section, but that assumption may not always be reasonable. We can exchange the normal distribution with any other family of density distributions with some location parameter (e.g., the mean), such as Student's-t or Beta (with the mean and sample size parameterization) when S is continuous or Poisson or negative binomial distribution when S values are counts. Below, we exemplify with the Student’s-t:

*S* ~ Student’s-*t* (***ν***, ***μ***, ***σ***)

***μ*** = ***α*** + ***αs***[sample1] + ***αs***[sample2]  + ***β****.****C***

where *ν* stands for degrees of freedom, *μ* for mean and *σ* for standard deviation:

**2.2.2 Varying intercepts and slopes**

This framework also permits the inclusion of varying intercepts (*αc*[cluster]) and slopes (*β*[cluster]), which allow the relationship between *S* and *C* to change across different clusters. We write such a model as follows:

*S* ~ Normal (***μ***, ***σ***)

***μ*** = ***α*** + ***αc***[cluster] + ***αs****[sample1]* + ***αs****[sample2]* + ***β***[cluster].*C*

Varying intercepts and slopes allow the model to capture cluster-specific behaviors that would otherwise be missed or inadvertently absorbed into other behaviors, thus providing better inferences.

**3. Case study**

We demonstrate our approach by modeling how the community similarity of riparian plant communities changes as a function of two covariates: network distance and precipitation difference. Our data consist of riparian vegetation inventories obtained between 2003 and 2006 (INAG 2008) in 338 sites distributed across continental Portugal, covering 11 river basins.

To measure community similarity between pairs of sites we selected the Sorensen similarity index

where *a* is the number of shared species, *b* is the number of unique species to one of the sites, and *c* is the number of unique species to the other site. In other words, the Sorensen index corresponds to the number of species in common between two plots divided by their average number of species. A Sorensen index of 1 indicates that communities have the same composition, while 0 indicates that communities have no species in common. We performed the calculations for obtaining the Sorensen index with the package “vegan” (Oksanen et al. 2019) using R 4.0.4 (R Core Team 2021).

In order to illustrate the method, we selected two covariates, the network distance and precipitation difference. The network distance is the distance between two sites along the river network. Precipitation difference is the absolute value of the difference in mean annual precipitation between two sites using data between 1960 and 1990 from Monteiro-Henriques et al. (2016).

To model Sorensen indices, we chose a Beta distribution because it is a continuous probability distribution defined on the interval [0, 1], which is the same range of values of the Sorensen index. We selected a parameterization of the Beta distribution based on the mean, *μ*, and sample size, *κ* (Stan Development Team 2020a). We wrote the model as follows:

**Likelihood**

Sorensen indexi ~ Beta distribution (***μ***, *κ*)

logit(*μ*) = ***α*** + ***αs***[sample1] + ***αs***[sample2] + ***αc***[basin] +

***β*1**[basin] Network distance + ***β*2**[*basin*] Precipitation difference

**Priors**

***α*** ~ Normal (0,0.3)

***αs*** ~ Normal (0,*σs*)

***σs***~ Exponential(1)

***αc***~ Normal (0,0.3)

***β*1**[*basin*] ~ Normal (***μ***distance, ***σ***distance)

***β*2**[*basin*] ~ Normal (***μ***precipitation, ***σ***precipitation)

***μ***distance, ***μ***precipitation ~ Normal (0,0.3)

***σ***distance, ***σ***precipitation ~ Exponential (1)

***κ*** ~ Normal (0,50)

To incorporate sensitivity to the covariates while ensuring that *μ* is bounded between 0 and 1 we used the logit of *μ* in a linear model of the covariates. The term *α* is the baseline for characterizing changes in Sorensen indices. The terms ***αs***[sample1] and ***αs***[sample2] are the additive varying intercepts that account for the dependence between observations. The term ***αc*** is a varying intercept with 11 levels corresponding to the river basins. The terms ***β*1**[*basin*] and ***β*2**[*basin*] are the varying slopes corresponding to network distance, and precipitation differences. These slopes were sampled from a higher order distribution (i.e., hyper prior) that generates parameters for each river basin. Parameter estimates obtained in this fashion are more precise at the river basin level and typically more robust to extreme observations (Betancourt 2020, McElreath 2020). We interpreted the posterior distribution of ***μ****distance* and ***μ****precipitation* (as seen in Fig.2) as the average effect of the covariate on Sorensen indices if we could go into the field to repeat data collection an infinite number of times.

To improve model fit we often need to transform independent variables. A common approach consists of subtracting the mean and dividing the result by its standard deviation. Here we adopt a different strategy. Instead of using the mean and standard deviation, we use values that are determined by our domain expertise which improves the interpretably and generalizability of the resulting inferences, as we show below. We transformed network distance values by subtracting 100 km to observed values and by dividing the resulting value by 100. Therefore, a slope of, for instance, -0.10 means that an increase of 1 km in the network distance beyond a baseline of 100 km will decrease Sorensen indices by -0.10. Concerning precipitation difference values, we applied a log(x+1) transformation because we detected a second maximum for larger values (see Appendix 1 – 4.1) . Afterwards, we subtracted 5.71 ( log(300+1) = 5.71) and divided the resulting value by 5.71. Therefore, a slope of -0.1 indicates that if the precipitation difference increases by 1 mm beyond 300 mm (log(300 +1) = 5.71), the Sorensen index change by – 0.10.

Equally important in a Bayesian analysis is to assess the validity of our choice of priors, i.e., the prior model (Betancourt 2020). The prior model expresses our beliefs about how community similarity changes across river basins before we consider the evidence provided by the observed data. Based in our expertise on riparian habitats, we believe most vegetation samples will share between 30 and 60% of species (Rodríguez-González et al. 2008). Conversely, few samples will share less than 30% or more than 60% of the species. As for the effects of covariates, we used weakly informative priors that allow for both positive and negative effects on community similarity. To check if the prior model reflects our beliefs, we generated and plotted simulated distributions of Sorensen indices from the prior predictive distribution.

Concerning the full model, to validate whether it captures the data's relevant structure, we compared the observed distribution of the Sorensen indices with the posterior distribution of Sorensen indices and checked how well they matched. Then, 1) we plotted the differences between the posterior distribution and the observed Sorensen indices (i.e., error distributions) conditional on covariates and 2) we plotted the posterior distribution of Sorensen indices against the covariates. We checked for any systematic deviations indicating structure in the data that our model was unable to capture. We assessed the relative importance of the covariates using two criteria, 1) by determining whether 95% of the parameter's posterior distribution was above or below 0 and 2) by assessing the parameter's magnitude.

We performed all calculations with Stan via the R package “rstan” version 2.18.3 (Stan Development Team 2020b). We run the models with four independent Markov chains to sample from, with 1000 warmup iterations and 2000 sampling iterations. To check if our Markov chains were stationary and enabled reasonable posterior expectation value estimators, we performed both qualitative and quantitative diagnostics. In addition, to spot checking traceplots we also verified that the split potential scale reduction factor, often called “Rhat”, was consistent with 1 for all functions of interest and verified that there were no divergent transitions or Markov chains that saturated the maximum tree depth.

Finally, a note on terminology. Throughout the text we use the term “retrodictive” instead of “predictive” whenever we are referring to the process of comparing predicted results with observed data (Betancourt 2020).

**3. Results**

**3.1 Prior predictive checks**

We generated 1000 simulations of Sorensen indices from the prior model. From these, we observe that most probability mass is around 0.5, which reflects our prior belief that most vegetation samples share approximately 30 to 60% of the species. Still, there is some probability mass close to 0 and 1 which reflects our belief that the number of samples that share very few or very high numbers of species is low, but not zero. Based on Figure 2 we conclude that our prior model is consistent with our knowledge of community similarity across river basins.

**3.2 Model validation**

The chains were stationary and well mixing with Rhat values of ~1. No iterations ended with divergences or saturated the maximum tree depth (Appendix 1 - 4.4.1).

The posterior retrodictive distribution of Sorensen indices closely matched the observed distribution of Sorensen indices except for values below 0.05, which are slightly overestimated, and values above 0.62, which are slightly underestimated (Fig. 3). We found no evidence of systematic deviations between our data and the model (Annex 1 – 4.4.2).

**3.3 Effects of covariates on community similarity**

Network distance was negatively associated with Sorensen indices, with mean slope estimates ranging between -0.16 for the Douro and -0.76 for the Vouga basin. In the case of Mira and Minho basins, a small part of the 95% credibility intervals crosses zero, which implies a small probability of the slope being zero or slightly positive (Fig. 4a). The mean estimate for ***μ***distance is -0.33 with a 95% credibility interval [-0.50, -0.15], which means that on average, network distance has a negative effect on Sorensen indices.

Higher differences in precipitation are associated with lower levels of community similarity. The mean estimates for precipitation difference slopes were mostly negative, with mean values between -0.58 and -0.12 (Fig. 4b). However, in six out of the eleven basins, credibility intervals crossed zero, indicating the effect could also be positive with varying degrees of probability. For instance, over 30% of Mira's basin parameter distribution is on the right side of zero.

In Appendix 1 we provide a step by guide with both R and Stan codes for reproducing these results.

**4. Discussion**

In this report, we presented a hierarchical Bayesian approach for modeling changes in community similarity as a function of covariates. This approach deals with the dependence between similarity indices calculated with the same samples by adding two varying intercepts, one for each sample, that captures the heterogeneity of the corresponding sample. This approach has two main advantages when compared to Mantel tests and Generalized Dissimilarity Modelling. The first is that it can be easily adapted to incorporate new response variable distribution and varying intercepts and slopes, which is not possible within the frameworks provided by Mantel tests and GDM. Our approach leverages the power and flexibility of Stan’s probabilistic programming language which allows us to write and modify models without having to rely on multiple software packages. The second is that our approach (and implementation in Stan) provides a clear path for performing model validation. Bayesian models are generative, meaning that we can obtain predicted data from the posterior distribution and compare it with the observed data. This procedure allows us to determine if the model captures the data's relevant structure and provides key insights for improving the model when that is not the case.

We believe our approach provides a solid framework for analyzing paired comparisons data that can be of interest to both ecologists who seek to model changes in community composition, and scientists from other fields that work with similar type of data, such a anthropologists and molecular biologists.

**6. Acknowledgments**

FD, LBA and PMRG were financed by the Fundação para a Ciência e a Tecnologia (FCT). FD through project POCI-01-0145-FEDER-028729, LBA under the Norma Transitória – DL57/2016/CP1440/CT0022, and PMRG through Investigador FCT (IF/00059/2015) and CEEC (2020.03356.CEECIND).

7. **References**

Betancourt, Michael (2020). Hierarchical Modeling. Retrieved from https://github.com/betanalpha/knitr\_case\_studies/tree/master/hierarchical\_modeling, commit 27c1d260e9ceca710465dc3b02f59f59b729ca43

Ferrier, S., G. Manion, J. Elith, and K. Richardson. 2007. Using generalized dissimilarity modelling to analyse and predict patterns of beta diversity in regional biodiversity assessment. Diversity and Distributions 13:252–264.

Gravel, D., C. D. Canham, M. Beaudet, and C. Messier. 2006. Reconciling niche and neutrality: the continuum hypothesis: Reconciling niche and neutrality. Ecology Letters 9:399–409.

Guillot, G., and F. Rousset. 2013. Dismantling the Mantel tests. Methods in Ecology and Evolution 4:336–344.

Hubbell, S. 2001. The Unified Neutral Theory of Biodiversity and Biogeography (MPB-32). Princeton University Press.

INAG. 2008. Manual para a avaliação biológica da qualidade da água em sistemas fluviais segundo a DQA - Protocolo de amostragem e análise para os macrófitos. Lisboa, Portugal.

Legendre, P. 2000. Comparison of permutation methods for the partial correlation and partial mantel tests. Journal of Statistical Computation and Simulation 67:37–73.

Legendre, P., D. Borcard, and P. R. Peres-Neto. 2005. Analyzing Beta Diversity: Partitioning the Spatial Variation of Community Composition Data. Ecological Monographs 75:435–450.

Legendre, P., M.-J. Fortin, and D. Borcard. 2015. Should the Mantel test be used in spatial analysis? Methods in Ecology and Evolution 6:1239–1247.

Legendre, P., F.-J. Lapointe, and P. Casgrain. 1994. Modeling brain evolution from behavior: a permutational regression approach. Evolution; International Journal of Organic Evolution 48:1487–1499.

Leibold, M. A., M. Holyoak, N. Mouquet, P. Amarasekare, J. M. Chase, M. F. Hoopes, R. D. Holt, J. B. Shurin, R. Law, D. Tilman, M. Loreau, and A. González. 2004. The metacommunity concept: a framework for multi-scale community ecology. Ecology Letters 7:601–613.

Manly, B. F. J. 1986. Randomization and regression methods for testing for associations with geographical, environmental and biological distances between populations. Researches on Population Ecology 28:201–218.

Mantel, N. 1967. The Detection of Disease Clustering and a Generalized Regression Approach. Cancer Research 27:209–220.

McElreath, R. 2020. Statistical Rethinking: A Bayesian Course with Examples in R and Stan. 2 edition. Chapman and Hall/CRC, Boca Raton.

Monteiro-Henriques, T., M. J. Martins, J. O. Cerdeira, P. Silva, P. Arsénio, Á. Silva, A. Bellu, and J. C. Costa. 2016. Bioclimatological mapping tackling uncertainty propagation: application to mainland Portugal. International Journal of Climatology 36:400–411.

Oksanen, J., F. G. Blanchet, M. Friendly, R. Kindt, P. Legendre, D. McGlinn, P. R. Minchin, R. B. O’Hara, G. L. Simpson, P. Solymos, M. Henry H. Stevens, E. Szoecs, and Helene Wagner. 2019. vegan: Community Ecology Package. R package version 2.0-10, <http://CRAN.R-project.org/package=vegan>.

Rodríguez-González, P. M., M. T. Ferreira, A. Albuquerque, D. E. Santo, and P. R. Rego. 2008. Spatial variation of wetland woods in the latitudinal transition to arid regions: a multiscale approach. Journal of Biogeography 35:1498–1511.

R Core Team. 2021. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

Smouse, P. E., J. C. Long, and R. R. Sokal. 1986. Multiple Regression and Correlation Extensions of the Mantel Test of Matrix Correspondence. Systematic Zoology 35:627–632.

Stan Development Team. 2020a. Stan Functions Reference Version 2.25.

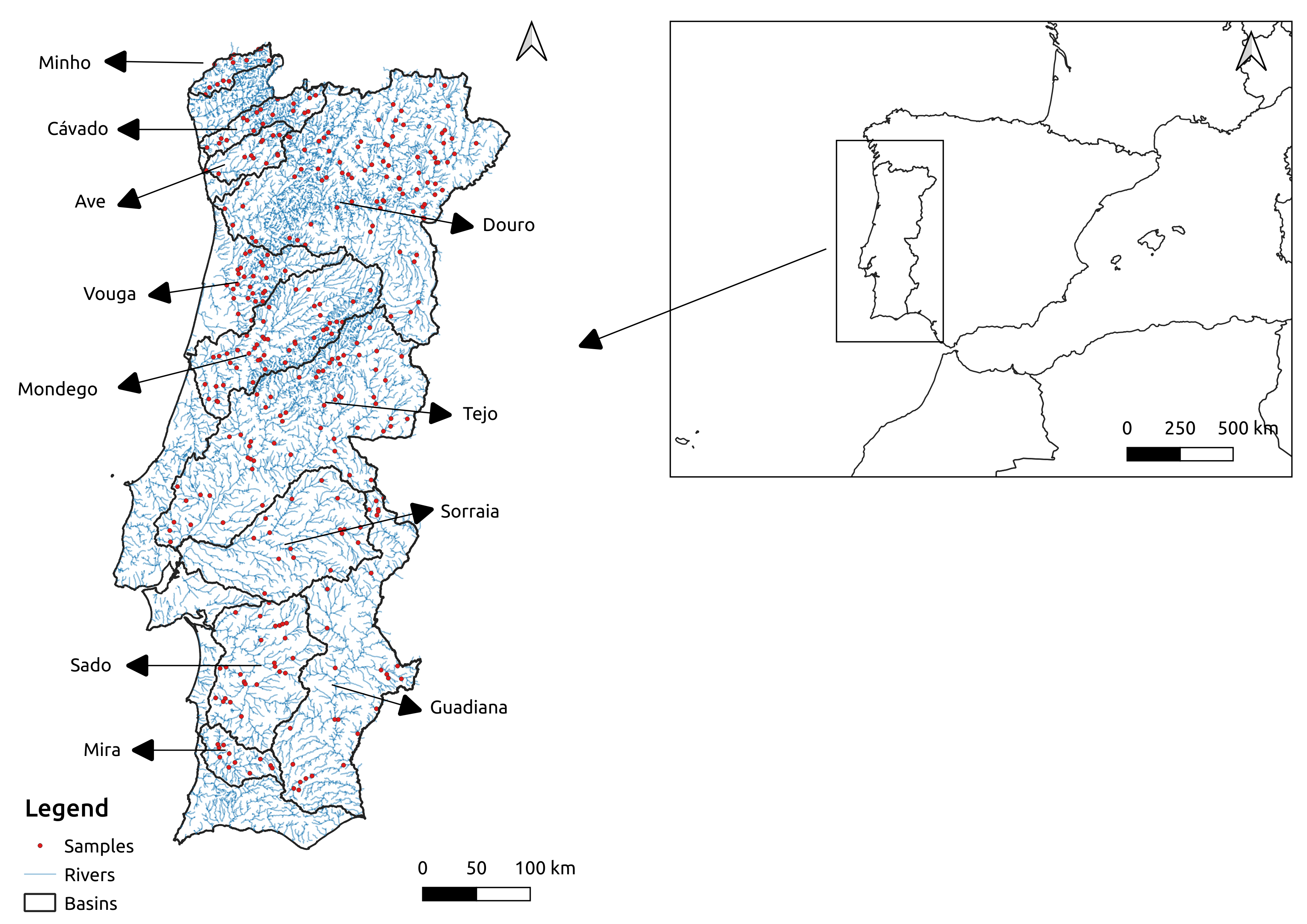
Stan Development Team. 2020b. RStan: the R interface to Stan. R package version 2.21.2.

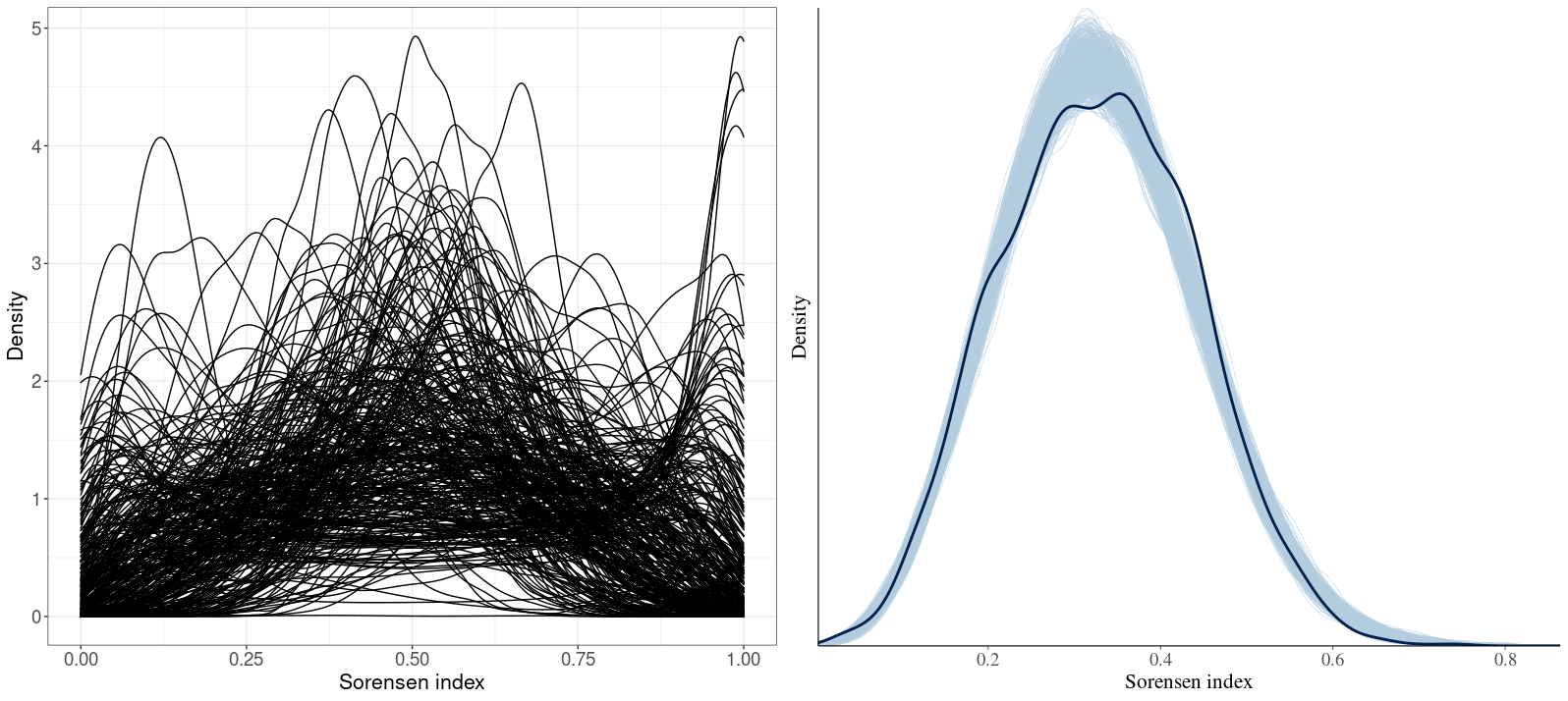
**Figure 1** –Continental Portugal and the location of the vegetation plots in the eleven river basins studied (red dots).

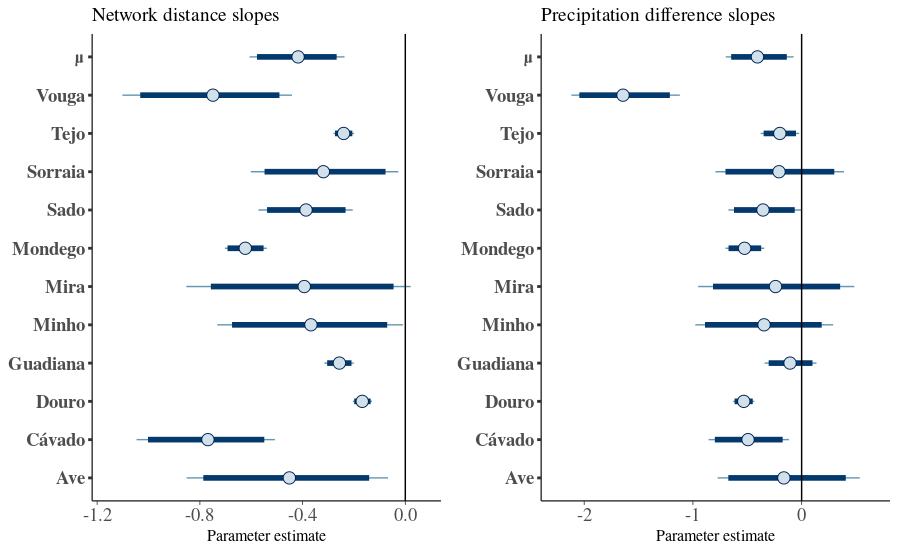
**Figure 2** – Density plot showing prior predictive distribution (light blue lines).

**Figure 3** – Density plot showing the observed distribution of Sorensen indices (thick line) against 1000 posterior distributions (thin lines).

**Figure 4** – Posterior estimates of slope parameters for the covariates network distance (left) and precipitation difference (right) for each river basin. The parameter *μ* is the mean of the normal distribution where slopes are sampled from. Dark blue lines represent 95% credibility intervals. The thin light blue line represents the complete distribution of the parameters. The dot represents the marginal posterior mean.

**Figure 1** –Continental Portugal and the location of the vegetation plots in the eleven river basins studied (red dots).

**Figure 2** – Density plots showing prior predictive distribution (left) and the observed distribution of Sorensen indices (thick line) against 1000 posterior distributions (thin lines) (right).



**Figure 4** – Posterior estimates of slope parameters for the covariates network distance (left) and precipitation difference (right) for each river basin. The parameter *μ* is the mean of the normal distribution where slopes are sampled from. Dark blue lines represent 95% credibility intervals. The thin light blue line represents the complete distribution of the parameters. The dot represents the marginal posterior mean.