SORT 41 (2) July-December 2017, 277-296 DOI: 10.2436/20.8080.02.60

Bayesian hierarchical models for analysing the spatial distribution of bioclimatic indices

Xavier Barber*,[1], David Conesa[2], Antonio López-Quílez[2], Asunción Mayoral[1], Javier Morales[1] and Antoni Barber[3]

Abstract

A methodological approach for modelling the spatial distribution of bioclimatic indices is proposed in this paper. The value of the bioclimatic index is modelled with a hierarchical Bayesian model that incorporates both structured and unstructured random effects. Selection of prior distributions is also discussed in order to better incorporate any possible prior knowledge about the parameters that could refer to the particular characteristics of bioclimatic indices. MCMC methods and distributed programming are used to obtain an approximation of the posterior distribution of the parameters and also the posterior predictive distribution of the indices. One main outcome of the proposal is the spatial bioclimatic probability distribution of each bioclimatic index, which allows researchers to obtain the probability of each location belonging to different bioclimates. The methodology is evaluated on two indices in the Island of Cyprus.

MSC: 62F15, 62M30, 62P10, 62P12, 86A32.

Keywords: Bioclimatology, geostatistics, parallel computation, spatial prediction.

1. Introduction

Bioclimatology is an ecological science that studies the relationship between climate and the distribution of the living species on Earth, particularly the distribution of vegetation. It aims to determine the relationship between certain numerical values of temperature and precipitation and the areas in which single plant species and plant communities are geographically distributed. The spatial distribution of the species and the relationship between climate and vegetation allows us to better manage plant resources and landscape, as well as to forecast the production of agricultural and forestry resources to combat hunger and determine future vegetation scenarios in certain geographic areas through the study of vegetation borders.

[1] Centro de Investigación Operativa. Universidad Miguel Hernández de Elche. xbarber@umh.es

[2] Dpt. Estadística i Investigació Operativa. Universitat de València.

[3] IDENTIA Institute.

Received: June 2016
Accepted: March 2017

As an ecological science, the distribution of the spatial structure of species and its relationship with environmental factors having high spatial dependence has been an important subject of study for several years. Osborne et al. (2000); Britton et al. (2001); Cheddadi, Guiot and Jolly (2001); Tasser and Tappeiner (2002); Legendre, Borcard and Peres-Neto (2005); Dostálek, Frantík and Šilarová (2014); Baltensperger and Huettmann (2015) are examples of studies applying these ideas to analyse land-use changes and distribution of terrestrial vegetation.

Bioclimatic Classification Systems have been introduced to assign bioclimates to a region under study by means of what are known as bioclimatic indices. But more importantly, these bioclimates allow us to identify the geographical limits of the main types of vegetation in the region under study. As a result, having a good spatial representation of the bioclimatic indices is key to describing the relationship between climate and the distribution of vegetation.

Information about bioclimatic indices is usually available only in meteorological stations, not in the whole region of study. It is therefore important to be able to construct maps from these data. Until now, many studies have used only standard geographical information system (GIS) techniques. Geostatistics has also been proposed as a way to explain bioclimatic indices (Robertson, 1987; Rossi et al., 1992; Burrough, 2001; Garzón-Machado, Otto and del Arco Aguilar, 2014), although this approach can present certain obstacles such as spatial scale problems (Atkinson and Tate, 2000).

Our main interest in this research is twofold. Firstly, we present another way to model the spatial distribution of bioclimatic indices. Specifically, we propose a hierarchical Bayesian model to predict (in non-sampled locations) the bioclimatic index values by incorporating the altitude and spatial features of each sampled location. As usual in Bayesian approaches, we also explain how to select prior distributions in this context. But more importantly, we secondly describe the two main outcomes of the modelling, i.e., the posterior predictive distribution of bioclimatic indices and the probability maps for the bioclimates, which provide more realistic geographical limits. As the resulting hierarchical model has no closed expression for the posterior distribution of all the parameters, we also present how to perform inference by MCMC methods, and how to predict on non-observed locations by means of distributed programming, reducing the computation time by more than 80% in comparison to standard R packages.

The remainder of this article is organised as follows. After this introduction, Section 2 presents a general Bayesian hierarchical spatial model of the bioclimatic indices. In Section 3, we describe how to select prior distributions, while Section 4 explains how to perform inference and prediction for these indices. In Section 5, we apply this methodology in a real setting, we obtain the predictive distributions of two bioclimatic indices on the island of Cyprus, using the altitude and the climate information (temperatures and rainfall) from 59 meteorological stations. Finally, Section 6 concludes and presents some future lines of research.

2. Modelling bioclimatic indices

In what follows, we first introduce three bioclimatic indices of the Worldwide Bioclimatic Classification System by Rivas-Martínez (Rivas-Martínez, 1994; Rivas-Martínez et al., 2002; Rivas-Martínez and Rivas-Saenz, 2016), one of the most popular Bioclimatic Classification Systems available. This classification encompasses five macrobioclimates (Tropical, Mediterranean, Temperate, Boreal and Polar), which are in turn subdivided into twenty-seven bioclimates and five bioclimatic variants. It is worth noting that all the results presented here could also be applied to any other bioclimatic index from any classification selected. After defining the bioclimatic indices, Section 2.2 describes the Bayesian hierarchical spatial model for each one of them.

2.1. Bioclimatic indices

As previously mentioned, the procedure for constructing bioclimatic maps is based on the bioclimatic indices. In general, these indices are values obtained by simple mathematical expressions that combine certain climatic parameters and factors such as altitude or latitude, and which are commonly used to characterise the climate of a region. This makes it possible to recognise climatically homogeneous areas that may have similar vegetation types (species, communities, series).

One of the most important bioclimatic indices is the *Ombrothermic Index* (OI), which relates the rainfall and the temperature in an area using an average of the last n years (usually at least 25 years), and it is defined by

$$OI = \frac{10}{n}\sum_{j=1}^{n}\left(\frac{P_{p,j}}{T_{p,j}}\right), \tag{1}$$

where P_p is the sum of the average rainfall (in mm.) of the months whose average temperature is above zero degrees Celsius, and T_p is the sum of monthly average temperatures above zero degrees Celsius, expressed in tenths of a degree.

The variation of temperature (thermicity) over the seasons in an area is one of the most influential factors in the characterisation of climate, since the vegetation distribution is greatly affected by the area's thermicity. Hence, another important bioclimatic index is the *Thermicity Index* (TI) of the last n years, defined as

$$TI = \frac{10}{n}\sum_{j=1}^{n}(T_j + m_j + M_j), \tag{2}$$

where T is the sum of the annual mean temperature in decimal degrees, m is the average of the minimum temperature of the coldest month and M is the average of the maximum temperature of the coldest month.

This Thermicity Index has some problems of definition in extratropical regions (North and South of latitude 23 N and S respectively). The Compensated Thermicity Index (TIc) avoids these problems by weighting the Thermicity Index value (TI) by adding or subtracting the Compensation Value, C_i, in those places where the Continentality Index (CI), defined as the annual oscillation variation of temperature $CI = T_{\max} - T_{\min}$) takes extreme values:

$$TIc = \begin{cases} TI & \text{if } 8 \leq CI \leq 18, \\ TI + C_i & \text{if } CI < 8 \text{ or } CI > 18, i = 0, \ldots, 4 \end{cases} \tag{3}$$

Note that all the temperatures are in Celsius, and periods are 25 years, the minimum recommended period.

2.2. Bayesian hierarchical model for bioclimatic indices

After presenting the bioclimatic indices, we now introduce a way of modelling them by means of a Bayesian hierarchical spatial model. If $\boldsymbol{Y} = [Y(s_i)]_{i=1}^{n}$ represents the vector of values of the bioclimatic index in a subset of locations $\boldsymbol{s} = (s_1, \ldots, s_n)$ in the region D, then the usual geostatistical assumption is that $\boldsymbol{Y}$ is multivariate normal:

$$\boldsymbol{Y} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}), \tag{4}$$

where $\boldsymbol{\mu}$ denotes the mean vector of the process, and $\boldsymbol{\Sigma}$ represents the covariance matrix between locations. This matrix can be re-written separately as spatial and non-spatial covariances matrices

$$\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_w + \boldsymbol{\Sigma}_r, \tag{5}$$

which, assuming that the observations are conditionally independent given the spatial process, can also be expressed as

$$\boldsymbol{\Sigma}_w = \sigma^2 \boldsymbol{H}(\boldsymbol{\theta}); \text{ and } \boldsymbol{\Sigma}_r = \tau^2 \boldsymbol{I}, \tag{6}$$

where $\boldsymbol{H}(\boldsymbol{\theta})$ is the Matérn correlation matrix between locations (Matérn, 1986), which depends on two parameters $\boldsymbol{\theta} = (\phi, \nu)$; the scale parameter $\phi > 0$ and the shape parameter $\nu > 0$. It is worth noting that the Matérn is a really flexible and general family of correlation generalising many of the most-used covariance models in spatial statistics (exponential and Gaussian among them).

The mean vector of the process can be related with covariates (in our case, altitude), and so the bioclimatic index is expressed as

$$\boldsymbol{Y}|\boldsymbol{\beta},\boldsymbol{W},\tau^2 \sim \mathcal{N}\left(X\boldsymbol{\beta}+\boldsymbol{W},\tau^2\boldsymbol{I}\right), \tag{7}$$

where $X\boldsymbol{\beta}$ represents the linear predictor associated with the covariates at the locations $\boldsymbol{s}=(s_1,\ldots,s_n)$.

Hence, the Bayesian hierarchical model corresponding to geostatistical homogeneous Gaussian process data for a bioclimatic index is expressed in three levels of information as

$$\begin{array}{ll} (I) & \boldsymbol{Y}|\boldsymbol{\beta},\boldsymbol{W},\tau^2 \sim \mathcal{N}\left(X\boldsymbol{\beta}+\boldsymbol{W},\tau^2\boldsymbol{I}\right) \\ (II) & \boldsymbol{W}|\sigma^2,\boldsymbol{\theta} \sim \mathcal{N}\left(0,\sigma^2\boldsymbol{H}(\boldsymbol{\theta})\right) \\ (III) & p(\boldsymbol{\beta},\sigma^2,\tau^2,\boldsymbol{\theta}), \end{array} \tag{8}$$

where the first level is the Gaussian process, the second level shows the information on the spatial effect and the third level specifies the prior distribution parameters and hyper-parameters.

Following Yan et al. (2007), and in order to avoid the identifiability problem of spatial and non-spatial variability, we reparametrise (8) as

$$\begin{array}{l} (I)\,\boldsymbol{Y} \sim \mathcal{N}\left(X\boldsymbol{\beta},\xi^2\left[(1-\kappa)\boldsymbol{H}(\boldsymbol{\theta})+\kappa\boldsymbol{I}\right]\right) \\ (II)\,p(\boldsymbol{\beta},\xi^2,\kappa,\boldsymbol{\theta}), \end{array} \tag{9}$$

where $\xi^2=\sigma^2+\tau^2$ now represents the total variability of the random effects, and $\kappa=\tau^2/\xi^2$ stands for the proportion of the non-spatial variability with respect to the total variability.

Once the model is determined, the next step is to estimate its parameters. As we are using the Bayesian paradigm, we have to select the prior distribution for the vector of parameters involved in the model.

3. Selection of prior distributions

Making use of previous information is considered one of the most useful characteristics of Bayesian statistics. A subjective approach involves defining prior distributions for unknown parameters according to personal experience and impression, recognising that the expert opinion is better than no knowledge. In contrast, objective Bayesians defend the idea that no other information should be considered apart from that introduced during model specification, although finding that prior distribution which contains only that knowledge can sometimes be tricky. In the context of spatial geostatistics models, the case in hand, it must be taken into account that using non-informative priors can lead to improper posterior distributions (De Oliveira, 2007).

A usual assumption when expressing prior knowledge is to consider prior independence of the parameters, that is,

$$p(\boldsymbol{\beta},\xi^2,\kappa,\boldsymbol{\theta}) = p(\boldsymbol{\beta})p(\xi^2)p(\kappa)p(\boldsymbol{\theta})\,.$$

In order to express our knowledge for each of these parameters, we must elicit both their distributions and the values of their hyperparameters. As mentioned above, for the latter the choice of their values can come under the “complete ignorance” premise, although we can also include the information available about them in order to improve the final posterior distribution (Dongen, 2006).

In particular, the distribution for $\boldsymbol{\beta}$ is again based on the assumption of prior independence of its components, the usual choice being either Gaussian distributions or non-informative improper distributions. As the resulting posterior is in both cases proper, we use the improper one, that is,

$$p(\boldsymbol{\beta}) = p(\beta_0,\beta_1) = p(\beta_0)p(\beta_1) \propto 1\,.$$

With respect to the proportion κ, the natural choice is a uniform distribution between 0 and 1, $\kappa \sim \mathcal{U}(0,1)$.

For the Matérn function parameters, $\boldsymbol{\theta} = (\phi,\nu)$, and taking into account that we are using the parameterisation proposed by Handcock and Wallis (1994) in which the parameter ϕ is largely independent of ν, we propose using a product of two independent distributions. In particular, our choice for the prior distribution of ϕ is

$$p(\phi) = \mathcal{U}\left(\frac{1}{d_1},\frac{1}{d_2}\right), \tag{10}$$

where d_1 is the furthest distance between two locations, and d_2 is the minimum distance between the two nearest locations. Following recommendations by Stein (1999) and Finley, Banerjee and Gelfand (2015), our choice of smoothing parameter ν is $\nu \sim \mathcal{U}(0.05,1.95)$.

The last parameter to be elicited is the total variability ξ^2 of the bioclimatic index. In this case, note that information is available which can be included in the prior. Indeed, as explained in the previous section, indices depend on temperature and precipitation by definition, and therefore, they only take values within a defined range (the highest and lowest value of the index in the region of study, according the Rivas-Martínez classification), denoted by $(Y_{\min},Y_{\max})$, with $Y_{\min} > 0$.

This information about ξ^2 can be incorporated in the scale parameters of different distributions. The underlying idea is to consider that the observed values of the index on a set of locations is a priori uniformly distributed between $(Y_{\min},Y_{\max})$. Note that this uniform distribution is the most disadvantageous option as this would imply that all the regions have the same orographic features. The corresponding variability of this uniform distribution is

$$\mathrm{Var}(Y) = \frac{(Y_{\max} - Y_{\min})^2}{12}, \tag{11}$$

the maximum value of which (denoted as $V_{\max}$) would be an upper bound of the variability index. A prior distribution could then be constructed by matching the range of variability $(a, V_{\max})$ with the quantile 0.95 of any chosen distribution (Chambers and Dunstan, 1986; Strupczewski et al., 2007). In other words,

$$0.95 = \int_{a}^{V_{\max}} f(y|\boldsymbol{\alpha})dy, \tag{12}$$

where f is the chosen prior distribution and $\boldsymbol{\alpha}$ its corresponding parameters. Since variability is always positive, a can be chosen to be as small as possible (e.g. $a = 0.001$). Table 1 shows the resulting scale parameters for the usual priors: uniform over the variance, uniform over the standard deviation, inverse gamma or half-Cauchy.

Table 1: *Upper bound for the variability index and prior distribution for a specific bioclimatic index range.*

$p(\xi^2) \sim \mathcal{U}(0.001, b)$	$b = \dfrac{V_{\max} - 0.00005}{0.95}$
$p(\xi) \sim \mathcal{U}(0.001, \sqrt{b})$	$b = \dfrac{V_{\max} - 0.00005}{0.95}$
$p(\xi^2) \sim \mathcal{IG}(2, \beta)$	$0.95 = \int_{0.001}^{V_{\max}} \dfrac{\beta^2}{\Gamma(2)} x^{-3} e^{-\beta x} dx$
$p(\xi^2) \sim \mathcal{HC}(\delta)$	$\delta = \dfrac{V_{\max}}{tan\left(\frac{1}{2} \cdot \pi \cdot 0.95\right)}$

To summarise, the final model for any bioclimatic index $\boldsymbol{Y}$ using the second option of Table 1 (uniform over the standard deviation) is

$$\begin{aligned} &(I) \quad \boldsymbol{Y} \sim \mathcal{N}\left(X\boldsymbol{\beta}, \xi^2 \left[(1-\kappa)\boldsymbol{H}(\boldsymbol{\theta}) + \kappa \boldsymbol{I}\right]\right); \ \boldsymbol{\theta} = (\phi, \nu) \\ &(II) \quad p(\boldsymbol{\beta}, \xi, \kappa, \phi, \nu) \propto 1 \times \mathcal{U}(0.001, \sqrt{b}) \times \mathcal{U}(0,1) \times \mathcal{U}(1/d_1, 1/d_2) \times \mathcal{U}(0.05, 1.95) \end{aligned} \tag{13}$$

Note that the advantage of this final model is that we only have to assign a prior distribution on ξ, since the remaining parameters are obtained as $\sigma^2 = (1-\kappa)\xi^2$ and $\tau^2 = \kappa\xi^2$.

4. Inference and prediction

The model in (13) contains all our knowledge about the index, but it does not yield closed analytic expressions for the posterior distribution of the parameters, $p(\boldsymbol{\beta},\xi,\kappa,\phi,\nu|Y,X)$. Therefore, numerical approximations are needed in order to make inference about them. Among others, one feasible (indeed one of the most popular) possibility is to use Markov chain Monte Carlo (MCMC) methods (Gamerman and Lopes, 2006) that draw samples from any intractable posterior by running a cleverly constructed Markov chain over a long period, the stationary distribution of which is the one we want to simulate from. Among the different ways of building these chains, the most popular are Gibbs sampling and the Metropolis-Hastings algorithm (Gilks, Richardson and Spiegelhalter, 1996).

In our case, we use WinBUGS (Lunn et al., 2000), a flexible software for performing the Bayesian analysis of complex statistical models (see Banerjee, Carlin and Gelfand, 2014 for examples of how to implement spatial hierarchical Bayesian models with WinBUGS). The reason for this choice is that it gives us more flexibility when specifying the matrix variance-covariance of the first hierarchy level. Moreover, it allows us to easily set prior distributions over the standard deviation.

As usual in MCMC, we run three chains for a long period discarding the first hundreds or thousands (depending on the convergence, the burn-in period can be extended) and then take samples from the three chains. Regarding convergence (to the correct stationary distribution) assessment, the Brooks-Gelman-Rubin statistic and the effective sample size (see Gelman et al., 2013 for more information about these statistics) can be calculated for every parameter in the model. The Brooks-Gelman-Rubin statistic must have a value under 1.1, while the effective number of iterations must be above 100 for every mentioned parameter.

Once the inference has been carried out, the next step is to predict the values of the bioclimatic indices in the rest of the area of interest, especially in unsampled locations. In our case, as we are using the Bayesian approach, prediction is reduced to obtain the posterior predictive distribution of the indices in a set of new locations.

In particular, if $\boldsymbol{Y}_p$ represents the values of a bioclimatic index in a new set of locations with observed covariates X_p, then the posterior predictive distribution of the new values $\boldsymbol{Y}_p$ (conditional to the observed ones, henceforth, $\boldsymbol{Y}_o$) is

$$p(\boldsymbol{Y}_p|\boldsymbol{Y}_o,X_o,X_p) = \int p(\boldsymbol{Y}_p|\boldsymbol{Y}_o,X_p,\boldsymbol{\beta},\xi,\kappa,\phi,\nu)p(\boldsymbol{\beta},\xi,\kappa,\phi,\nu|Y,X)d(\boldsymbol{\beta},\xi,\kappa,\phi,\nu)\,, \tag{14}$$

where the extended data vector $p(\boldsymbol{Y}_p|\boldsymbol{Y}_o,X_p,\boldsymbol{\beta},\xi,\kappa,\phi,\nu)$ has a conditional multivariate normal distribution arising from the joint multivariate distribution of $\boldsymbol{Y}_p$ and $\boldsymbol{Y}_o$ in (7).

As with the posterior distribution of the parameters, expression (14) has no closed form, and again numerical approximations are needed. One way to obtain a simulated sample from this posterior predictive distribution is via the composition method. In particular, if $\{\boldsymbol{\beta}_i,\xi_i,\kappa_i,\phi_i,\nu_i\}_{i=1}^{M}$, represents a simulated sample from the posterior distribution of the parameters, then a simulated sample from the posterior predictive distribution is obtained by simulating from the conditional multivariate distribution of the observed values $\boldsymbol{Y}_p$, that is, $\{p(\boldsymbol{Y}_p|\boldsymbol{Y}_o, X_p, \boldsymbol{\beta}_i, \xi_i, \kappa_i, \phi_i, \nu_i)\}_{i=1}^{M}$.

Note that the conditional multivariate distribution $p(\boldsymbol{Y}_p|\boldsymbol{Y}_o, X_p, \boldsymbol{\beta}, \xi, \kappa, \phi, \nu)$ is a multivariate normal distribution with mean

$$E\left[\boldsymbol{Y}_p|\boldsymbol{Y}_o\right] = \boldsymbol{\mu}_p + \boldsymbol{\Sigma}_{po}\boldsymbol{\Sigma}_{oo}^{-1}(\boldsymbol{Y}_o - \boldsymbol{\mu}_o) \tag{15}$$

and variance-covariance matrix

$$V\left[\boldsymbol{Y}_p|\boldsymbol{Y}_o\right] = \boldsymbol{\Sigma}_{pp} - \boldsymbol{\Sigma}_{po}\boldsymbol{\Sigma}_{oo}^{-1}\boldsymbol{\Sigma}_{op}\,, \tag{16}$$

where

$$\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{pp} & \boldsymbol{\Sigma}_{po} \\ \boldsymbol{\Sigma}_{op} & \boldsymbol{\Sigma}_{oo} \end{pmatrix}$$

is the covariance matrix of the joint multivariate normal distribution of the extended data vector $(\boldsymbol{Y}_p, \boldsymbol{Y}_o)$.

As we are following the reparametrisation by Yan et al. (2007) in (9), the conditional multivariate distribution $p(\boldsymbol{Y}_p|\boldsymbol{Y}_o, X_p, \boldsymbol{\beta}, \xi, \kappa, \phi, \nu)$ is a multivariate normal distribution but with mean

$$E(\boldsymbol{Y}_p|\boldsymbol{Y}_o) = X_p\boldsymbol{\beta} + ((1-\kappa)\boldsymbol{H}_{po}(\boldsymbol{\theta}) + \kappa\boldsymbol{I})\left((1-\kappa)\boldsymbol{H}_{oo}(\boldsymbol{\theta}) + \kappa\boldsymbol{I}\right)^{-1}(\boldsymbol{Y}_o - X_o\boldsymbol{\beta}) \tag{17}$$

and variance-covariance matrix

$$V(\boldsymbol{Y}_p|\boldsymbol{Y}_o) = \tag{18}$$
$$\xi^2\left[((1-\kappa)\boldsymbol{H}_{pp}(\boldsymbol{\theta}) + \kappa\boldsymbol{I}) - ((1-\kappa)\boldsymbol{H}_{po}(\boldsymbol{\theta}) + \kappa\boldsymbol{I})\left((1-\kappa)\boldsymbol{H}_{oo}(\boldsymbol{\theta}) + \kappa\boldsymbol{I}\right)^{-1}((1-\kappa)\boldsymbol{H}_{op}(\boldsymbol{\theta}) + \kappa\boldsymbol{I})\right]$$

where

$$\boldsymbol{H}(\boldsymbol{\theta}) = \begin{pmatrix} \boldsymbol{H}_{pp}(\boldsymbol{\theta}) & \boldsymbol{H}_{po}(\boldsymbol{\theta}) \\ \boldsymbol{H}_{op}(\boldsymbol{\theta}) & \boldsymbol{H}_{oo}(\boldsymbol{\theta}) \end{pmatrix}$$

is the Matérn correlation matrix between predicted and observed locations.

Implementing the above composition method implies evaluating this mean vector and variance-covariance matrix for each of the simulations. But note that this evaluation can be computationally expensive. Dealing with 15000 simulations (5000 per chain) from the posterior distribution and about 1000 new locations (to predict) would involve evaluating 15000 times expressions (17) and (18). This is the reason why we do not use WinBUGS, because although feasible, it is really slow.

An obvious (but naive) option would be to consider fewer points over the surface to predict, and a small random sample from the posterior distribution. However, this option would produce posterior predictive distributions with lower resolution and, therefore, the resulting predictive maps would have no practical interest. Other options would be to use the spatial-temporal modelling R library `spBayes` (Finley et al., 2015), or to directly implement equations (15) and (16) using programming languages such as the R matrix computation language (Bates and Maechler, 2015); C++ via the interface package `Rcpp` to connect with R (Eddelbuettel et al., 2011); or directly C++ (Sanderson, 2010).

Our approach is to use intensive computation techniques such as parallel computation (Adams et al., 1996; Blackford et al., 1997; Rosenthal, 2000; Rossini, Tierney and Li, 2007; Whiley and Wilson, 2004), that allow us to increase the performance when doing matrix calculations, and therefore, work with a large number of new locations to predict with all the samples previously obtained by simulation from the posterior distribution using WinBUGS. Nevertheless, as stated by Golub and Van Loan (1996) and Cuenca, Giménez and González (2004), the use of parallel computation is convenient only if computational times are substantially reduced.

In this study we use C language to program the prediction equations, and then the ScaLAPACK and PLAPACK libraries to perform the linear algebra calculations needed to obtain the mean vector and variance-covariance matrix. Interestingly, with this parallelisation of the algorithm for generating a multivariate normal sample, we reduce the computation time by close to 80% compared to other options such as `spBayes` and similar R packages.

Graphical representation of the posterior predictive distributions of Bioclimatic indices

Having obtained the posterior predictive distribution of the indices, our final task is to represent these distributions throughout the area of interest in order to obtain a good visualisation of their behaviour in the area. We present two different representations of these predictive distributions, the first one being the mean and the standard deviation of the posterior predictive distribution, and the second one, the probability distribution of each bioclimatic index belonging to different bioclimates.

To obtain the map of the mean (similarly the map of standard deviation), we use multilevel B-splines Approximation (Lee, 1997) to interpolate the values of the mean (the standard deviation) of the bioclimatic indices over the whole area using the obtained values of the posterior mean (standard deviation) predictive distribution on the predicted locations.

Although the mean and the standard deviation reflect most of the information about the posterior predictive distributions, the most valuable information we can get from these distributions comes from the way that they can show us the probability of each location belonging to the different bioclimates. Indeed Rivas-Martínez's bioclimatic classification system uses different ranges of the bioclimatic indices to classify the different bioclimates. For example, the Continentality Index ranks the climate in three types, namely, Hyperoceanic ($CI \in [0,11[$), Oceanic ($CI \in [11,21[$), and Continental ($CI \in [22,65]$). Note that representing the probability of the predictive distribution of belonging to each of these ranges can be very relevant for studying changes in vegetation zones, climate change advances, and many other climatic issues that could provide valuable information for the management and use of land in the area under study.

Obtaining this probability is straightforward using the simulated values of the predictive distribution. If a bioclimatic index Y is defined in l disjoint intervals $R_1, R_2, \ldots, R_l$ that describe l bioclimates, and $\{r_{ik}\}_{i=1}^{n}$ represents a sample from the posterior predictive distribution for each location in $\{s_k\}_{k=1}^{m}$, then the posterior probability that each location belongs to each interval constituting the index is given by:

$$\begin{aligned} P(Y(s_k) \in R_j) &= \int_{Rj} \int p(Y(s_k)|\mathbf{Y}_o, X_p, \boldsymbol{\beta}, \xi, \kappa, \phi, \nu) p(\boldsymbol{\beta}, \xi, \kappa, \phi, \nu | Y, X) d(\boldsymbol{\beta}, \xi, \kappa, \phi, \nu) dY(s_k) \\ &\approx \frac{\#\{r_{ik} \in R_j\}}{n}, \ \ j = 1, \cdot, l, k = 1, \cdot, m. \end{aligned} \tag{19}$$

The result is a discrete probability distribution for each location that we call the spatial bioclimatic probability distribution. Note that the best way to represent this distribution is by presenting a single figure made up of different graphs, each one showing the probability of belonging to each bioclimate (see Figure 4 for an example).

The representation of each probability can be seen as a puzzle of pieces that fit by overlapping and provide the distribution boundaries between the types of bioclimates for each index. These boundaries are highly relevant because they determine the areas that could be about to change in the near future (caused for example by a slight change in climatic parameters). This representation is therefore critical in studies about climate change and its effects on the vegetation of a region.

5. Bioclimatic classification of the island of Cyprus

We illustrate the usefulness of the approach presented here through an application to analyse two bioclimatic indices (Ombrothermic Index and Thermicity Index) on the island of Cyprus with the final aim of showing its bioclimatic classification.

Cyprus is an island country in the Eastern Mediterranean. It is the third largest and the third most populous island in all the Mediterranean. Some of its geographical characteristics are as follows: it measures 240 kilometres (149 miles) long and 100 kilometres (62 miles) wide at its widest point; it lies between latitudes 34° and 36° N, and longitudes 32° and 35° E. Cyprus is dominated by two mountain systems, the Troodos and the Kyrenia Mountains, between which lies a central plateau, the Mesaoria.

The information gathered to create the bioclimatic classification of the island consisted of the geographical location, the altitude and the values of the two bioclimatic indices from 59 meteorological stations across the island, together with the geographi-

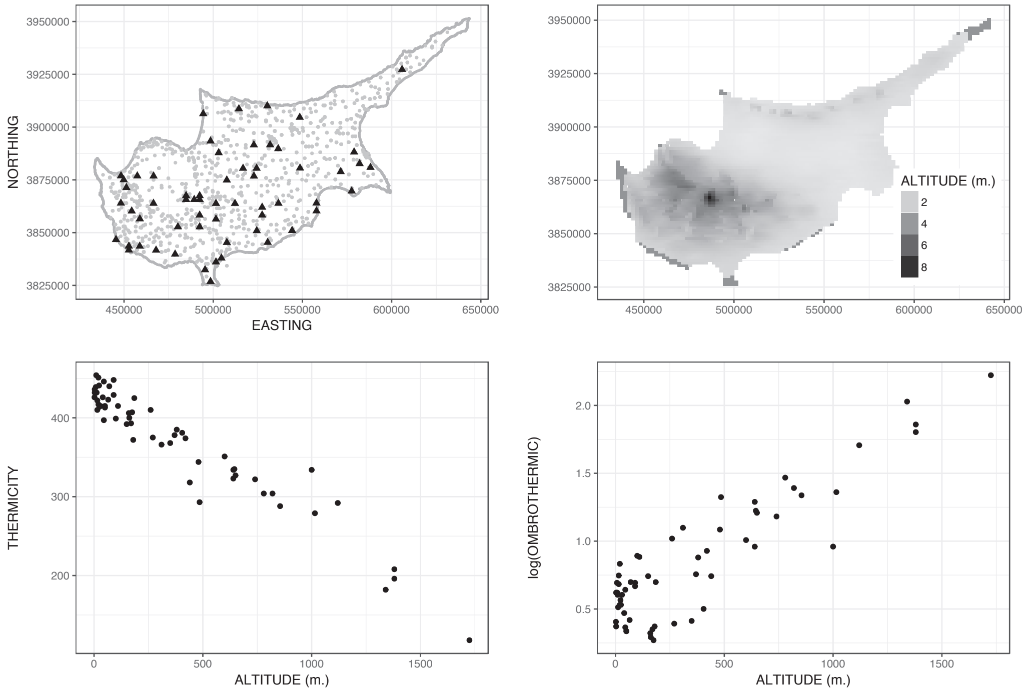

Figure 1: *Upper left: geographical location of observed and predicted sites in the Cyprus island. Black triangles represent the 59 meteorological stations (observed locations), while red points represent the 755 locations where prediction had to be performed. Upper right: contour map of the island. Lower left: thermicity and altitude relationship. Lower right: Log(Ombrothermic index) and altitude relationship.*

cal location and altitude of other 775 locations (used to predict the indices), in particular, the ones that the geographical map of the island provides. Figure 1 shows the geographical location of observed and predicted sites, jointly with the contour map of the island and the relationship between both indices and the altitude. It is worth mentioning that these two indices are not related, as can be seen in the left side of Figure 2. This allows us to analyse both indices independently. If the indices were related, a joint modelling would be necessary (see the right side of Figure 2 for an example of two related indices, namely the ombrothermic and continentality indices).

5.1. Ombrothermic Index

We first present the results obtained when analysing the Ombrothermic Index (using the logarithm transformation to improve its linear relationship with altitude). Table 2 presents the median of the posterior distribution of the parameters of the model in equation (13) along with their corresponding 95% credible intervals. These posterior distributions were obtained by simulation using WinBUGS (Lunn et al., 2000). Each posterior distribution was approximated from 15000 (5000 from each of three simulation chains)

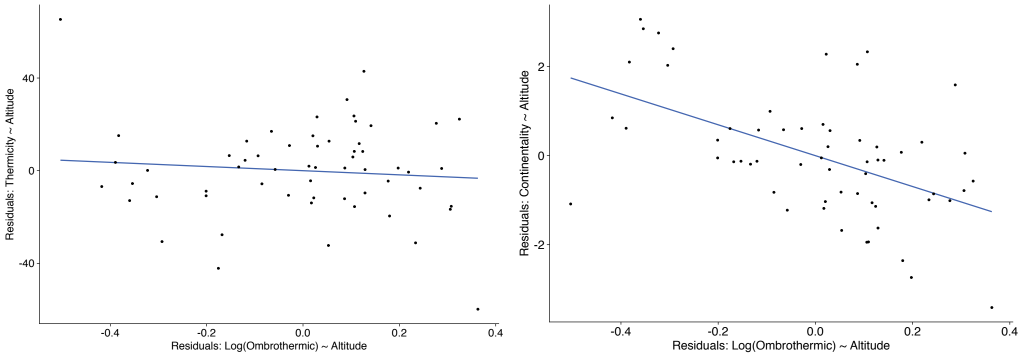

Figure 2: *Relationship between indices after adjusting linear regression of each index by altitude. Left side, relationship between residuals of ombrothermic and thermicity indices. Right side, relationship between residuals of ombrothermic and continentality indices.*

Table 2: *Median of the posterior distribution and* 95% *credible intervals of the parameters for the Ombrothermic Index model.*

Parameters	Median	$p_{2.5}$	$p_{97.5}$
β_0	5.28×10^{-1}	3.63×10^{-1}	6.84×10^{-1}
β_1	7.61×10^{-4}	6.02×10^{-4}	9.12×10^{-4}
ξ^2	4.80×10^{-2}	2.37×10^{-2}	1.70×10^{-1}
κ	1.03×10^{-1}	1.91×10^{-2}	3.83×10^{-1}
ϕ	5.54×10^{-5}	2.92×10^{-5}	9.55×10^{-5}
ν	1.48	1.02	1.93

simulated values (obtained after discarding ten thousand simulations from a burn-in period that guaranteed convergence). As commented above, these posterior distributions were obtained using the uniform distribution over the standard deviation.

As expected, results for β_1 in Table 2 show a positive effect on the altitude. Note also that the spatial effect is necessary to describe the behaviour of the index, as expressed by the small value of κ (which indicates the small proportion of non-spatial variability with respect to the total variance). It is also worth noting that the maximum variance (used in expression (11) to obtain the prior distributions) in Mediterranean bioclimate, 0.515, does not affect our results. Indeed, this shows that our prior construction methodology can really be considered as uninformative.

It is worth noting that a sensitivity analysis about the prior selection was performed for both indices. In particular, we fitted different models using all the different priors introduced in Table 1. Results indicate that both estimations and credible intervals obtained were similar independently of the priors used.

Figure 3 shows the maps of the mean and standard deviation (as a prediction error measure) of the posterior predictive distribution of the Ombrothermic Index. As men-

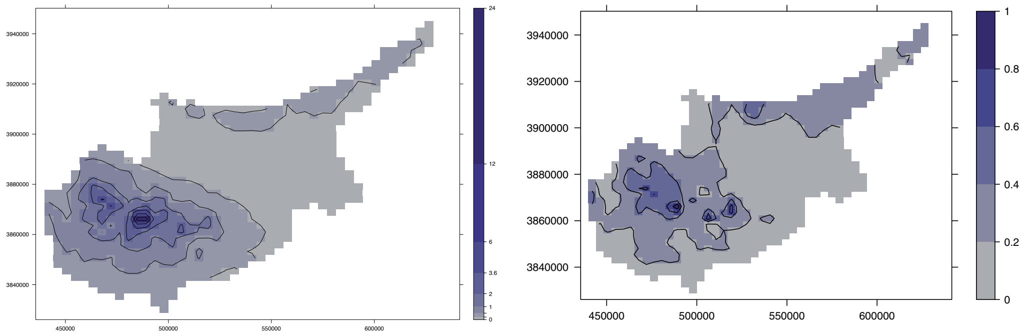

Figure 3: *Mean (left) and standard deviation (right) of the posterior predictive distribution of the Ombrothermic Index.*

tioned above, this predictive distribution was approximated by means of intensive computation techniques that allow us to predict the values of the bioclimatic indices in the 775 unsampled locations.

The mean map clearly reflects the topography of the island, while the standard deviation map shows the uncertainty in areas with no data but, more importantly, it also reflects the areas where the terrain is changing on the island. Note also that the scale of the observed prediction error is very small compared to the scale of measurement of the index considered throughout the island. The proposed method is therefore a very powerful tool for creating the bioclimatic rating of Cyprus based on the Ombrothermic Index. Note also that the map of the mean is similar that the one we could obtain using multiple linear regression followed by ordinary kriging of the regression residuals as in Garzón-Machado et al. (2014), although with our approach we can explore further the behaviour of the indices.

From a biological point of view, also note that the mean map in Figure 3 also properly reflects zones with higher altitude (corresponding to larger values of the index), and those areas with the highest rainfall. Indeed, the predicted map obtained shows the landscape changes that can be observed in any orthophoto of the island.

Once we have the posterior predictive distribution of the index we can use it to obtain the maps of the spatial bioclimatic probability distribution introduced in the previous section. As mentioned above, these maps show the posterior probability of an index belonging to each subtype.

Figure 4 shows the posterior probability of the nine possible ombrotypes (categories of the Ombrothermic Index) that can be observed in the Mediterranean bioclimate. The figure represents the probability of one location on the island belonging to each ombrotype. Note that in Cyprus the subtypes Hyperhumid, Ultrahyperhumid, Arid, Hyperarid and Ultrahyperarid are not possible, while probabilities greater than zero indicate that Humid and Subhumid are possible at the highest altitudes, and the Dry subtype is possible on the coast and Semiarid in the north and east.

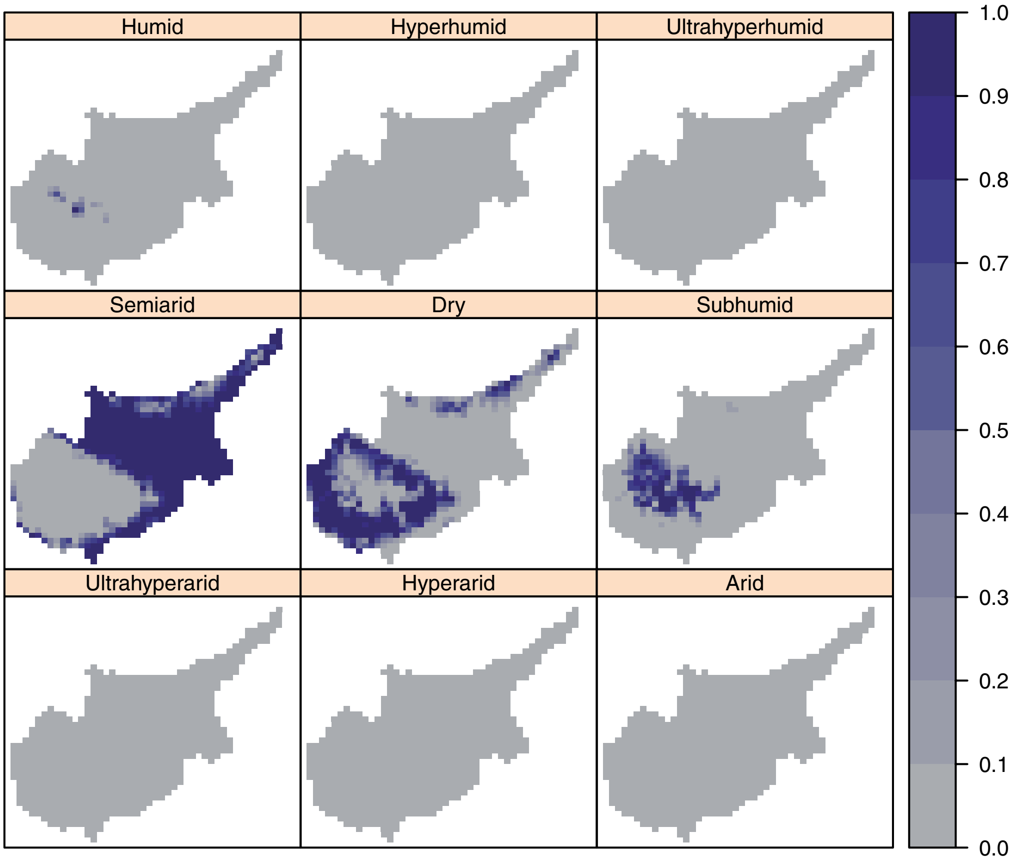

Figure 4: *Spatial bioclimatic probability distribution of the Ombrothermic Index.*

As it can be seen from the figure, there is a high probability of the Humid subtype being found in the two mountain peaks of the Central mountains and of the Subhumid subtype occurring in the mountainous area of the central mountains. The Dry subtype has a high probability of occurring on the hillsides of those peaks and the northern ridge of the island and finally the Semiarid subtype is likely to be found in the central plateau. It is worth noting how important these probability distribution maps are from a biological point of view, as they provide more accurate information on the subtype boundaries, by using a gradient map showing the border from one subtype to another.

5.2. Thermicity Index

We now show the results for the Thermicity Index in Cyprus. This index presents a peculiar relationship with the orography, and obviously with the temperature-altitude pair, i.e., higher altitude is associated with lower temperature.

Table 3: *Median of the posterior distribution and 95% credible intervals of the parameters for the Thermicity Index model.*

Parameters	Median	$p_{2.5}$	$p_{97.5}$
β_0	6.10	6.06	6.13
β_1	-5.48×10^{-4}	-6.04×10^{-4}	-4.92×10^{-4}
ξ^2	7.90×10^{-3}	5.53×10^{-3}	1.21×10^{-2}
κ	3.95×10^{-1}	1.94×10^{-2}	9.11×10^{-1}
ϕ	3.12×10^{-4}	2.52×10^{-5}	5.40×10^{-4}
ν	0.594	0.0763	1.75

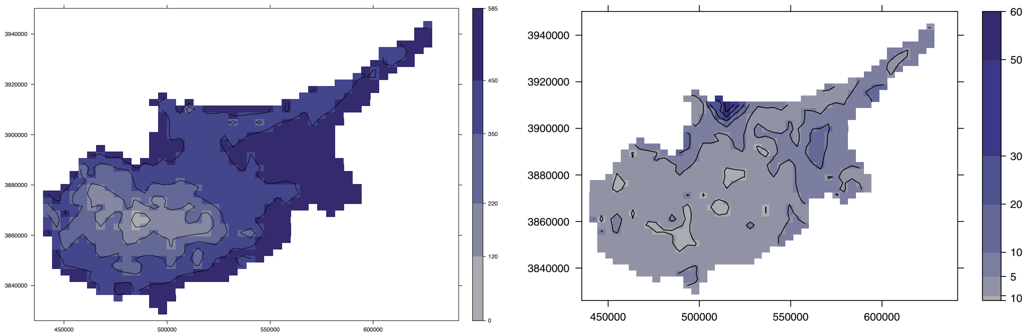

Figure 5: *Mean (left) and standard deviation (right) of the posterior predictive distribution of the Thermicity Index.*

Table 3 shows the median of the posterior distribution of the parameters along with their corresponding 95% credible intervals for this index. As above, these posterior distributions were obtained by simulation using WinBUGS, although in this case neither efficiency (in terms of computational time) nor convergence were as good as for the Ombrothermic Index (indeed the number of discard simulations needed in the burn-in was 20000 for this index).

Results for β_1 in Table 3 now show a negative effect on the altitude, which corresponds to the effect in climatology known as the mountain-valley wind effect. The value for κ is around 0.395 with a credible interval that nearly covers the whole $[0,1]$ interval. This clearly indicates that the model can not distinguish between the spatial and non-spatial variabilities. The fact that some weather stations present different values even though they are close to one another, clearly indicates that this index probably does not have a major spatial effect.

Figure 5 shows the mean and the standard deviation of the posterior predictive distribution of the parameters for the Thermicity Index model. The mean map clearly shows the island's mountain system, which is a real factor in explaining the variability for the

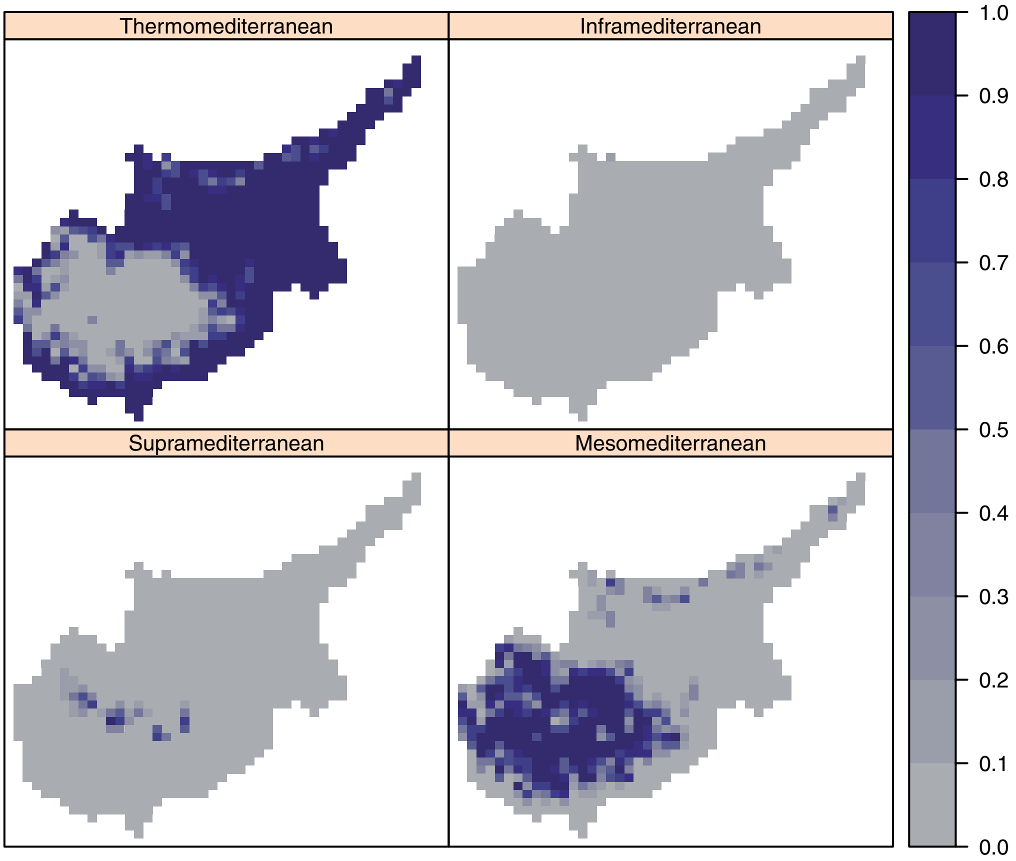

Figure 6: *Spatial bioclimatic probability distribution of the Thermicity Index.*

Thermicity Index, as mentioned previously. Figure 5 also shows the differences between the south and the north of the island and the two principal mountains.

Figure 6 shows the spatial bioclimatic probability distribution of the four possible thermotypes that can be observed in the Mediterranean bioclimate. As can be appreciated from the figure, there is a strong relationship between altitude and thermicity. The Supramediterranean subtype is very likely to be found at the highest locations, while on the hillsides there is a high probability the Mesomediterranean thermotype. Finally, there is a high probability of the ombrotype for the rest of the island beinf Thermomediteranean. Again, this map could be very helpful for landscape management, as it illustrates the vegetation frontiers, due to the close relationship between thermicity and vegetation.

6. Conclusions

In this study, we have introduced a hierarchical Bayesian model that allows us to obtain the spatial distribution of bioclimatic indices by incorporating the altitude and spatial features of each sampled location. Two of the most important advantages of the Bayesian model formulation are that it incorporates parameter uncertainty (both in the inferential and prediction processes), and also prior information can be easily handled. In this context, we have shown how to incorporate our prior knowledge about the parameters via their prior distributions taking into account the particular characteristics of bioclimatic indices. Interestingly, this approach could be easily extended in other contexts. Moreover, sensitivity analysis have shown that there is no dependence on the prior selected.

Also interest is the usefulness of the two main outcomes of the modelling. Posterior predictive distributions reflect most of the information about the bioclimates, but the most valuable information they provide comes from the fact that they inform us of the probability of each location belonging to the different bioclimates. This is done using what we have called the spatial bioclimatic probability distributions. These distributions could be a powerful tool in studies about climate change and its effects on the vegetation of a region, but also in landscape management, in particular to establish future policies or future resource management.

This study also explains how to use MCMC methods, in particular WinBUGS, for the inference in this context, and also how to perform distributed programming for the prediction, which allows us to reduce the computation time.

Another important issue to be mentioned is that in the case that the two analysed indices were related, a joint modelling should be used. In our case, as the Thermicity and Ombrothermic indices are not related there is no need for it, but with other indices the opposite applies and a joint modelling would be needed.

Finally, it should be noted that all the analytical approaches we used here to document the spatial distribution of bioclimatic indices can be applied in any other part of the world.

Acknowledgments

Xavier Barber, David Conesa and Antonio López-Quílez would like to thank the Ministerio de Economía y Competitividad (the Spanish Ministry of Economy and Finance) for its support in the form of the research grant MTM2016-77501-P (jointly financed with the European Regional Development Fund –FEDER–).

References

Adams, N., Kirby, S., Harris, P. and Clegg, D. (1996). A review of parallel processing for statistical computation. *Statistics and Computing*, 6, 37–49.

Atkinson, P. and Tate, N. (2000). Spatial scale problems and geostatistical solutions: a review. *The Professional Geographer*, 52, 607–623.

Baltensperger A. and Huettmann, F. (2015). Predicted shifts in small mammal distributions and biodiversity in the altered future environment of alaska: an open access data and machine learning perspective. *PloS one*, 10, e0132054.

Banerjee, S., Carlin, B. and Gelfand, A. (2014). *Hierarchical Modeling and Analysis for Spatial Data*, Second Edition. Chapman and Hall/CRC, Boca Raton.

Bates, D. and Maechler, M. (2015). *Matrix: Sparse and Dense Matrix Classes and Methods*. R package version 1.2-0.

Blackford, L., Choi, J., Cleary, A., D'Azevedo, E., E., Demmel, J., Dhillon, I., Dongarra, J., Hammarling, S., Henry, G., Petitet, A., Stanley, K., Walker, D. and Whaley, R. (1997). *ScaLAPACK Users' Guide*. SIAM, Society for Industrial and Applied Mathematics, Philadelphia.

Britton, A., Pakeman, R., Carey, P., and Marrs, R. (2001). Impacts of climate, management and nitrogen deposition on the dynamics of lowland heathland. *Journal of Vegetation Science*, 12, 797–806.

Burrough, P. (2001). GIS and geostatistics: essential partners for spatial analysis. *Environmental and Ecological Statistics*, 8, 361–377.

Chambers, R. and Dunstan, R. (1986). Estimating distribution functions from survey data. *Biometrika*, 73, 597–604.

Cheddadi, R., Guiot, J. and Jolly, D. (2001). The Mediterranean vegetation: what if the atmospheric CO_2 increased? *Landscape Ecology*, 16, 667–675.

Cuenca, J., Giménez, D. and González, J. (2004). Architecture of an automatically tuned linear algebra library. *Parallel Computing*, 30, 187–210.

De Oliveira, V. (2007). Objective Bayesian analysis of spatial data with measurement error. *The Canadian Journal of Statistics*, 35, 1–19.

Dongen, V. (2006). Prior specification in Bayesian statistics: three cautionary tales. *Journal of Theoretical Biology*, 242, 90–100.

Dostálek, J., Frantík, T. and Šilarová, V. (2014). Changes in the distribution of alien plants along roadsides in relation to adjacent land use over the course of 40 years. *Plant Biosystems*, (published online: 20 Dec 2014): 1–17.

Eddelbuettel, D., François, R., Allaire, J., Chambers, J., Bates, D. and Ushey, K. (2011). Rcpp: Seamless R and C++ integration. *Journal of Statistical Software*, 40, 1–18.

Finley, A., Banerjee, S. and Gelfand, A. (2015). spBayes for Large Univariate and Multivariate Point-Referenced Spatio-Temporal Data Models. *Journal of Statistical Software*, 63, 1–28.

Gamerman, D. and Lopes, H. (2006). *Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference*. CRC Press, Boca Raton.

Garzón-Machado, V., Otto, R. and del Arco Aguilar, M.J. (2014). Bioclimatic and vegetation mapping of a topographically complex oceanic island applying different interpolation techniques. *International Journal of Biometeorology*, 58, 887–899.

Gelman, A., Carlin, J., Stern, H., Dunson, D., Vehtari, A. and Rubin, D. B. (2013). *Bayesian Data Analysis*, Third Edition. Chapman and Hall-CRC, Boca Raton.

Gilks, W., Richardson, S. and Spiegelhalter, D. (1996). *Markov Chain Monte Carlo in Practice*. Chapman and Hall, London.

Golub, G. and Van Loan, C. (1996). *Matrix Computations*, Third Edition. Johns Hopkins University Press, Baltimore.

Handcock, M. and Wallis, J. (1994). An approach to statistical spatial-temporal modeling of meteorological fields. *Journal of the American Statistical Association*, 89, 368–390.

Lee, P. (1997). *Bayesian Statistics: an Introduction*, Second Edition. Arnold, London.

Legendre, P., Borcard, D. and Peres-Neto, P. (2005). Analyzing beta diversity: partitioning the spatial variation of community composition data. *Ecological Monographs*, 75, 435–450.

Lunn, D., Thomas, A., Best, N. and Spiegelhalter, D. (2000). WinBUGS – A Bayesian modelling framework: concepts, structure, and extensibility. *Statistics and Computing*, 10, 325–337.

Matérn, B. (1986). *Spatial Variation*, Second Edition. Springer-Verlag, Berlin.

Osborne, C., Mitchell, P., Sheehy, J. and Woodward, F. (2000). Modelling the recent historical mpacts of atmospheric CO_2 and climate change on Mediterranean vegetation. *Global Change Biology*, 6, 445–458.

Rivas-Martínez, S. (1994). Clasificación bioclimática de la tierra (bioclimatic classification system of the earth). *Folia Botanica Matritensis*, 13, 1–25.

Rivas-Martínez, S. and Rivas-Saenz, S. (2016). Worldwide Bioclimatic Classification System. Phytosociological Research Center. Spain, http://www.globalbioclimatics.org.

Rivas-Martínez, S., Rivas-Sáenz, S., Penas, A. et al. (2002). *Worldwide Bioclimatic Classification System.* Backhuys Pub.

Robertson, G. (1987). Geostatistics in ecology: interpolating with known variance. *Ecology*, 68, 744–748.

Rosenthal, J. (2000). Parallel computing and Monte Carlo algorithms. *Far East Journal of Theoretical Statistics*, 4, 207–236.

Rossi, R., Mulla, D., Journel, A. and Franz, E. (1992). Geostatistical tools for modeling and interpreting ecological spatial dependence. *Ecological Monographs*, 62, 277–314.

Rossini, A.J., Tierney, L. and Li, N. (2007). Simple parallel statistical computing in R. *Journal of Computational and Graphical Statistics*, 16, 399–20.

Sanderson, C. (2010). *Armadillo: An open source C++ linear algebra library for fast prototyping and computationally intensive experiments*. NICTA.

Stein, M. (1999). *Interpolation of Spatial Statistics Data: Some Theory for Kriging*. Springer-Verlag, New York.

Strupczewski, W., Kochanek, K., Weglarczyk, S. and Singh, V. (2007). On robustness of large quantile estimates to largest elements of the observation series. *Hydrological Processes*, 21, 1328–1344.

Tasser, E. and Tappeiner, U. (2002). Impact of land use changes on mountain vegetation. *Applied Vegetation Science*, 5, 173–84.

Whiley, M. and Wilson, S. (2004). Parallel algorithms for Markov chain Monte Carlo methods in latent spatial gaussian models. *Statistics and Computing*, 14, 171–179.

Yan, J., Cowles, M., Wang, S. and Armstrong, M. (2007). Parallelizing MCMC for Bayesian spatiotemporal geostatistical models. *Statistics and Computing*, 17, 323–335.