Università degli Studi di Padova

Department of Mathematics "Tullio Levi-Civita"

Master Degree in Data Science

Predicting Human Activities Patterns Based on Climate and Related Data

Supervisor: Prof. Fabio Aiolli

Department of Mathematics, Università degli studi di Padova

Co-Supervisor: Prof. Indrė Žliobaitė

Department of Computer Science, University of Helsinki

Student: Dario Zanelli

N. 1205199

Academic Year 2020-2021

Abstract

Human activities as agriculture and urban construction have altered a large portion of natural ecosystems and vegetation cover, defining three new land cover classes: croplands, urban and built-up areas and mosaics of croplands and natural vegetation. In a period in which climate is rapidly changing, understanding how climatic conditions are linked to the global distribution of human-modified land covers has fundamental importance. Our goal is to identify the aforementioned relationships and build models that predict realistic fractions of land covers associated with human activities only based on climate data.

Decision trees and random forests have been employed to solve the three multiple regression problems, one for each human-modified land cover class. Experiments for optimal model selection have been conducted. Out of three land cover classes, only croplands responded well to modelisation.

Decision trees exhibited sensible predictive accuracy and good potential for climatic patterns description, yet little robustness. Whereas, Random forests guaranteed higher accuracy and more stability, proving to be reasonably informative models. They provided valuable insights into the nature of the connections between climate and the distribution of croplands.

Contents

Introduction

Climate change is a modern reality that is having an increasing impact on life on Earth, causing disorders among ecosystems. Human beings face the changes adapting, strong of their capabilities and technology, which allow them to maintain the same, or even improve, living standards. Their actions have implications for the environment and, the precarious equilibrium between human beings and nature deteriorates, affecting the climate in turn.

This terrible vicious cycle is that in which the human race is trapped nowadays. To break the circle, humans have to step back, analyse the situation, identify the problems and find solutions. It might seem hard all at once, indeed there are so many interconnected topics to consider, yet, facing one at a time might be a good strategy.

With this view, the present study came to life. To restore the balance between human beings and the ecosystem, we need to understand the relationships that connect them. Then, the following research pointed at the comprehension of the links between climate and the global distribution of land cover classes associated with human activities, that is, croplands and urban areas.

Experience suggests that there are areas whose early vegetation cover has been completely replaced by cities or plantations while other lands have maintained their original aspect. Besides, many halfway shades exist. We deduce that human impact has not been the same everywhere and, climate may have been a determinant factor. Indeed, many questions arise on this theme, for instance: *How has the climate affected the distribution of croplands and urban areas? Has farming been strongly dependent on climatic conditions? Human knowledge in science and technology have been able to overcome climatic limitations when necessary?*

Moreover, aside from preventing further climate changes, we are very interested in their possible consequences, that is: *How will climate change affect the choice of areas for agriculture? Will the distribution of urban areas be influenced?* Good knowledge of how these activities are related to climatic conditions might help in understanding and predicting the effects of future climate scenarios. Hence, this has been another guide line of the research project.

This thesis is the report of the study that has been conducted to find answers to the previous questions. In practice, this has occurred through the research of patterns that relates climatic conditions to the distribution of human activities on the Earth's surface in a cause-effect relationship. In more specific terms, our purpose has been the global prediction of realistic percentages of human-modified land covers only based on climate data.

The research has inspected data organised onto a hexagonal geodesic discrete global grid. Each tile of the grid is an observation described by climatic features and fractions of land covers associated with urban areas and agriculture. The latter land cover class is divided in turn into areas with a considerable presence of croplands (at least 60%) and mosaics of small-scale cultivation (between 40% and 60%) and natural vegetation.

The prediction task is consequently structured as three distinct multiple regression problems, one for each land cover class, in a supervised learning setting. Considerable attention has been dedicated to the interpretability of the resulting models. To this end, tree-based methods have been employed (decision trees and random forests) because they are non-parametric methods that can often model complex and highly non-linear relationships while maintaining good interpretability.

The main challenge of this study is that over time humans have been able to develop strategies to adapt to hostile environments and that historical and social events have influenced the development of inhabited areas. Therefore, it is reasonable to suppose that the presence of urban areas or croplands is not only determined by climatic conditions, but many other factors might be relevant. The identification of clear and meaningful patterns could be misled by the absence of such additional information.

The content of the thesis is organised as follows:

Chapter 1 An overview of the theoretical background of decision trees and random forests is given. Moreover, their strengths and weaknesses are explained. The purpose of this chapter is to give the reader the fundamental tools for understanding the regression analysis and the reasons that motivated the employment of these models.

Chapter 2 The mathematical formulation of the regression problem is provided. The dataset is analysed, described and prepared for model learning.

Chapter 3 The framework of the solution of the regression problem via decision trees is explained. All the stages of the decisional process for finding the optimal model are described and motivated. The performances of the so-obtained decision tree are evaluated and, the patterns that it has learnt are interpreted.

Chapter 4 Similarly to Chapter 3, the regression analysis performed through random forests is described. The results are illustrated and interpreted.

Conclusion This final chapter contains an overview of the results of the research. Ideas for future works are briefly discussed.

Chapter 1

Tree-based methods

Tree-based methods are grounded on the idea of segmenting the feature space $\mathcal{X}$ into a set of simple regions and then fitting a constant value in each one in order to make predictions. The primary methods of this family of models are known as *decision trees*, since the set of splitting rules used to partition the feature space can be summarised in a tree.

The key advantages of tree-based methods are their conceptual simplicity and interpretability. Moreover, they can be used for both regression and classification problems, reason why they are generally referred to as *classification and regression trees*, in short, *CARTs* (Breiman et al., 1984).

In what follows we will focus only on the regression task, so that, terms like *decision tree* and *regression tree* will be considered interchangeable. However, many on the considerations will hold also for a classification scenario.

1.1 Regression trees

Let us consider a regression problem with a continuous response Y and a p-dimensional predictor space $\mathcal{X}$. The purpose of a *regression tree* is the description of the variation of the single response variable Y in the predictor space $\mathcal{X}$, by repeatedly partitioning the data into more homogeneous groups using combinations of predictor variables.

1.1.1 Tree growing

The process of building a *regression tree* can be outlined in two steps:

1. Partition the predictor space into M disjoint homogeneous regions $R_1, \dots, R_M$;

2. For every observation that falls into the region R_m, the same prediction is done: a constant value c_m.

These two apparently easy tasks hide more elaborate operations that we are going to investigate. In order to do that, let us formalise the context in which the regression tree should grow.

Let our data consist of N observations, each one composed of p predictors and one response, that is

$$(\mathbf{x}_i, y_i)\,, \quad i = 1, \dots, N$$

with $\mathbf{x}_i = (x_{i1}, \dots, x_{ip})$ a p-dimensional vector. Assuming that the partitioning operation of the predictor space into M regions returns $R_1, \dots, R_M$ and that the response is modelled as a constant c_m in each region R_m, $m = 1, \dots, M$, for a new observation $\mathbf{x}$ the model prediction can be computed as

$$\hat{y} := f(\mathbf{x}) = \sum_{m=1}^{M} c_m \mathbb{1}_{R_m}(\mathbf{x}).$$

The main issue of the tree-growing process is the way of creating the regions $R_1, \dots, R_M$. In principle, they could have any shape, but the most spread choice is to divide the predictor space into hyperrectangles. This choice is motivated by the ease of the building process and of the interpretation of the resulting predictive model.

The ideal goal of the partitioning algorithm is to find the regions $R_1, \dots, R_M$ that minimise the overall RSS [1] among all possible constant values c_m used for prediction in each region:

$$\min_{R_1,\dots,R_M} \sum_{m=1}^{M} \left(\min_{c_m \in \mathbb{R}} \sum_{i\,:\,\mathbf{x}_i \in R_m} (y_i - c_m)^2 \right).$$

It is straightforward to prove that the inner minimisation problem can be solved by choosing c_m as the mean of the response values for the training observations falling

[1] *Residual Sum of Squares*, i.e., $\text{RSS}(f) = \sum_{i=1}^{n} (y_i - f(\mathbf{x}_i))^2$.

within R_m (commonly referred to as *mean response*) [2] , that is

$$\hat{y}_m := \frac{1}{N_m} \sum_{i\,:\,\mathbf{x}_i \in R_m} y_i\,, \quad \text{with } N_m = |\{\mathbf{x}_i \in R_m\}|.$$

This result reduces the minimisation problem to be solved by the tree-building algorithm to

$$\min_{R_1,\dots,R_M} \sum_{m=1}^{M} \sum_{i\,:\,\mathbf{x}_i \in R_m} (y_i - \hat{y}_m)^2. \tag{1.1}$$

The argument of (1.1) is called *training error* and it is denoted by $\mathcal{R}(T)$, where T is the decision tree that segments the predictor space into $R_1, \dots, R_M$. Moreover, if we denote the within RSS of a predictor region R_m by

$$\mathcal{R}(m) = \sum_{i\,:\,\mathbf{x}_i \in R_m} (y_i - \hat{y}_m)^2$$

then the overall RSS, i.e., the training error, can be written as

$$\mathcal{R}(T) = \sum_{m=1}^{M} \mathcal{R}(m).$$

$\mathcal{R}(m)$ is the total squared deviation of each observation in R_m from the mean response $\hat{y}_m$. Clearly summing over all regions, we get the total squared deviation of each observation from the corresponding model prediction. In this sense $\mathcal{R}(T)$ is also referred to as the *impurity* of the tree T, and analogously $\mathcal{R}(m)$ is the impurity of region R_m.

When building a decision tree, considering every possible partition of the feature space into M boxes is computationally unfeasible. Therefore, the idea behind the building process is to repeatedly split the data on the basis of a simple rule applied to a single feature. At each split the data is partitioned into two disjoint groups. The splitting procedure is then recursively applied to each group separately.

[2] Let us suppose that N_m observations fall within the region R_m and that they are indexed from 1 to N_m. So, let us define the function $f(c_m) = \sum_{j=1}^{N_m} (y_i - c_m)^2$. Differentiating f once in c_m, we have

$$f'(c_m) = -2 \sum_{j=1}^{N_m} (y_j - c_m) = -2 \sum_{j=1}^{N_m} y_j + 2N_m c_m.$$

The only root of this function is $\bar{c} = \frac{1}{N_m} \sum_{j=1}^{N_m} y_j$. Looking at the sign of the second derivative, $f''(c_m) = 2N_m > 0$, we can conclude that $\bar{c}$ is the unique point of minimum for f.

A top-down, greedy algorithm, called *recursive binary splitting* comes in handy. It is top-down because it starts considering all points belonging to a single region, the entire predictor space $\mathcal{X}$ (which happens in the root of the tree), then successively splits it into subregions; each split corresponds to two new branches further down on the tree and each subregion to a new node. It is greedy because, at each step of the tree-building process, the chosen split is the best at that specific step, instead of a split that would lead to a better tree in some future step. The definition of *best split* will be clarified in a moment.

Starting with all the data points, in order to choose the *best split*, we select the splitting predictor X_j and the cutpoint s such that the regions

$$R_1(j,s) = \{\mathbf{X} \mid X_j < s\} \qquad\qquad R_2(j,s) = \{\mathbf{X} \mid X_j \geq s\}$$

produce the greatest reduction in the RSS of the resulting tree. In other words, the recursive binary splitting algorithm solves the minimisation problem

$$\min_{s,\,1\leq j\leq p} \left\{ \sum_{i\,:\,x_i\in R_1} (y_i - \hat{y}_1)^2 \; + \sum_{i\,:\,x_i\in R_2} (y_i - \hat{y}_2)^2 \right\}. \tag{1.2}$$

When the number of features p is not too large, this problem can be solved quickly, since the determination of the cutting point s for each feature is of fast execution (Hastie, Tibshirani, and Friedman, 2009).

Once the best split have been found, the data points are partitioned into two disjoint regions and the same procedure is repeated within both of them. The process continues until a given stopping criterion is reached or no further splitting is possible.

As mentioned above, as soon as the M regions have been created, every prediction is done by assigning to each observation the mean response of the region it belongs to.

1.1.2 Tree pruning

The *recursive binary splitting* algorithm can produce good predictions on the training observations, but it is likely to produce an overlarge tree that overfits[3] the training data. Tree size, i.e., the number of leaves of the tree, is a tuning parameter

[3] With the term *overfitting* we identify the situation in which the model memorises excessively the training data instead of learning the fundamental structure, hence, it fails to generalise on unseen data reliably.

governing the model complexity. This means that an excessively large tree might result to be too complex while an overly small tree might not capture the important structure of the data.

Slightly reducing the number of splits in the tree, then the number of predictor regions and the tree size, might lead to a lower variance and a better interpretation, at the cost of a little bias. This technique is known as *pruning*. Accordingly to the way it is performed, we can distinguish two main approaches:

Pre-pruning It is also known as *forward pruning* or *online pruning*. It prevents the generation of non-significant branches, by means of some stopping rule that decides when it is desirable to terminate some of the branches prematurely as the tree is generated.

When growing the tree, some significant measures can be used to assess the goodness of a split and, so, to either prevent or allow the creation of two new branches. If the split of a node results in falling below a prespecified threshold, then further partitioning of the given subset is halted, otherwise, it is expanded. High thresholds result in oversimplified and small trees, whereas low thresholds result in very little simplification and large trees. There are various techniques for pre-pruning, but in general, it is not recommended.

For instance, a possible strategy, proposed by Breiman et al. (1984), is to accept a split only when the resulting impurity decrease exceeds some threshold β. However, as Breiman et al. (1984) itself pointed out, this approach has two weaknesses. First, when the improvement threshold is too low, the resulting tree is extremely large, and the overfitting problem has not been solved yet. Second, when β is large, it is too short-sighted, indeed, there may be nodes whose profitless splits according to such a threshold value would be followed by splits with a significant impurity decrease.

Post-pruning It is also known as *backward pruning*. In this case, first a large decision tree is generated and then non-significant branches are removed.

A commonly used approach aims to retain the decision tree but to replace some of its subtrees by leaf nodes, thus converting a complete tree to one of its possible subtrees [4]. More in detail, we generate a (complete or very large)

[4] A *subtree* $T \subseteq T_0$ is defined as any tree that can be obtained by collapsing any number of

tree T_0 and we selectively prune it upward. Clearly to examine all the possible subtrees can be computationally unfeasible, then we need a way to select a small set of subtrees, among which the one with the lowest test error estimate will be chosen.

Cost-complexity pruning

Focusing on the post-pruning approach, the generally preferred strategy is the *cost-complexity pruning*, also known as *minimal error-complexity pruning* or *weakest-link pruning*, introduced by Breiman et al. (1984). The purpose of the algorithm is to find a sequence of nested trees, each of which is the best of all trees of its size.

As already mentioned in more general terms, in this technique we build a large tree T_0 by letting the partitioning procedure continue until terminal nodes are either small enough or contain only identical observations. We then define, for each subtree T, the cost-complexity criterion

$$C_\alpha(T) = \sum_{m=1}^{|T|} \sum_{i\,:\,x_i \in R_m} (y_i - \hat{c}_m)^2 + \alpha|T| =: \mathcal{R}(T) + \alpha|T|.$$

Here R_m is the predictor region corresponding to the mth terminal node and $|T|$ is the subtree size. The complexity parameter α can be interpreted as the complexity cost for each terminal node, so that, the subtree cost $C_\alpha(T)$ is a linear combination of the training error and a complexity penalty.

The idea of cost-complexity pruning is to find, for each value of $\alpha \geq 0$, the subtree $T_\alpha \subseteq T_0$ that minimises $C_\alpha(T)$, i.e., the solution of

$$\underset{T \subseteq T_0}{\arg\min}\ C_\alpha(T). \tag{1.3}$$

The tuning parameter α controls the trade-off between the subtree complexity and its adaptability to the training data. Notice that when $\alpha = 0$, then the solution of (1.3) is T_0 itself, while, as α increases, a price for having many terminal nodes is payed and the dimension of the resulting subtree decreases.

The existence of a unique solution of (1.3) for each value of α is proved by Breiman et al. (1984). Moreover, it is noticed that, although α runs through $\mathbb{R}^+$, only a finite number of subtrees of T_0 exist and only a subset of them is examined by

internal non-terminal nodes of T_0.

the cost-complexity pruning algorithm. Then, only the corresponding finite number of effective α values is identified:

$$A = \{\alpha_0 = 0,\, \alpha_1,\, \alpha_2,\, \dots\}.$$

In order to better explain this idea, let us assume T_{α_k} be the solution of (1.3) for $\alpha = \alpha_k$; thus, it keeps being solution of (1.3) for $\alpha > \alpha_k$ until a larger value α_{k+1} ($\alpha_{k+1} > \alpha_k$) is identified, such that a smaller subtree $T_{\alpha_{k+1}}$ is solution of (1.3) for $\alpha = \alpha_{k+1}$. The same reasoning can be repeated until the pruned tree collapse in the root of T_0, a single-node tree.

In conclusion Breiman et al. (1984) proved that cost-complexity pruning returns a finite increasing sequence of complexity parameters

$$\alpha_0 < \alpha_1 < \alpha_2 < \dots$$

corresponding to a finite sequence of subtrees of T_0, with progressively fewer terminal nodes. In addition, the sequence of subtrees results to be decreasing, that is,

$$\{T_\alpha\}_{\alpha \in A} \subseteq T_0 \;\; : \quad T_{\alpha'} \supset T_{\alpha''} \text{ if } \alpha' < \alpha''. \tag{1.4}$$

Furthermore, as α grows, branches of T_0 get pruned in a nested and predictable way and the pruning operations are computationally rapid, requiring only a small fraction of the total tree construction time. So, obtaining the whole sequence of subtrees as a function of α is an easy process (Breiman et al., 1984; James et al., 2013).

Selection of the best-pruned tree

The cost-complexity pruning results in a nested sequence of subtrees,

$$T_0 \supset T_1 \supset \cdots \supset \{t_1\}$$

where $T_k := T_{\alpha_k}$ and $\{t_1\}$ is the root of T_0. Since each subtree is the best of its size according to (1.3), choosing the *best* tree among them, means choosing the *best* size. The meaning of *best* depends on the criterion used to evaluate a tree.

It is evident that, if the criterion is the training error $\mathcal{R}(T)$, then the best subtree is T_0 itself. However, we are not interested in the model that is able to describe the training data the best way is possible, but we would prefer a model that is able to generalise its predictive power to new unseen data. Assuming that $\hat{\mathcal{R}}(T)$ is an unbiased estimate of the generalisation error of a regression tree, then we would choose the subtree minimising this error measure, i.e., solving

$$\arg\min_k \hat{\mathcal{R}}(T_k). \tag{1.5}$$

Breiman et al. (1984) proposed two possible choices for $\hat{\mathcal{R}}(T)$: the validation-set error estimate and the v-fold cross-validation error estimate. The former method is computationally more efficient than the latter, but the cross-validation approach makes a more effective use of the of the training data and is able to highlight information regarding the stability of the model.

Another reason in favour of the cross-validation approach is the possibility to quantify the uncertainty of each generalisation error estimate $\hat{\mathcal{R}}(T_k)$ computing its standard error in every problem setting. Indeed, v-fold cross-validation inherently returns an estimate $\hat{\mathcal{R}}(T_k)$ which is the mean of v values $\hat{\mathcal{R}}_1(T_k), \dots, \hat{\mathcal{R}}_v(T_k)$. In short terms, each of them is the generalisation error of the same decision tree (with fixed hyperparameters), computed on a different validation set that has been independently drawn from the same distribution of the training data. Therefore, the standard error on the sample mean $\hat{\mathcal{R}}(T_k)$ can be estimated as:

$$\textsc{se}\big(\hat{\mathcal{R}}(T_k)\big) = \frac{\hat{\sigma}\big(\hat{\mathcal{R}}(T_k)\big)}{\sqrt{v}}$$

where $\hat{\sigma}\big(\hat{\mathcal{R}}(T_k)\big)$ is the sample standard deviation.

Breiman et al. (1984) explains how the standard error comes in handy: it is necessary to apply the *one-standard-error rule* (1SE *rule*), that is,

> *Defined* $k_0 = \arg\min_k \hat{\mathcal{R}}(T_k)$, *then the tree selected is* T_{k_1}, *with* $k_1 > k_0$, *such that* k_1 *is the maximum* k *satisfying*
>
> $$\hat{\mathcal{R}}(T_{k_1}) \leq \hat{\mathcal{R}}(T_{k_0}) + \textsc{se}\big(\hat{\mathcal{R}}(T_{k_0})\big).$$

In other words, the 1SE rule picks the simplest tree whose generalisation error estimate is at most one standard error away from the best-performing tree's generalisation error estimate.

The 1SE rule shows its power especially when the error curve, as function of the tree size, presents the following behaviour: an initial steep decrease is followed by a flat valley in which the minimum is achieved, and a gradual increase comes after. In this case, the position of the minimum is unstable and might be affected by many factors (data distribution, model hyperparameters, random seeds, etc.). The 1SE tree size might reduce this instability still returning a model whose generalisation error is comparable with the best-performing tree.

1.1.3 Reasons behind the use of regression trees

There are many reasons to motivate the use of regression trees, and more in general, decision trees, as it is emphasised by Breiman et al. (1984), De'ath and Fabricius (2000), Hastie, Tibshirani, and Friedman (2009), James et al. (2013), Prasad, Iverson, and Liaw (2006), and Piramuthu (2008).

First of all, regression trees are easy to explain and interpret, thanks to the fact that they are graphically displayable. As already mentioned and as the name suggest, the sequence of splitting rules applied to the feature space $\mathcal{X}$ can be represented with a binary tree: the root of the tree contains the entire $\mathcal{X}$, the branches describe all the sequential divisions of $\mathcal{X}$ and the terminal nodes, or leaves, stand for the final predictor regions. Furthermore, the regression trees growing process follows closer than other methods the human decision-making. A plus point in favour of their interpretability.

Moreover, in a data analysis, regression trees can be useful for both exploring and modelling the data. When considering modelling the data, it can be done with two possible purposes: describing the data, that is, identifying a systematic structure that characterise the data or, predicting specific responses on new unobserved data. Again, trees are useful for both tasks. They perform local, greedy learning that allows to find a sensible, even if not necessarily optimal, models in a reasonable time.

Moving more into technical details, regression trees are non-parametric models, which means that they make no distributional assumptions about the predictor or the response variables and their accuracy is not affected by correlation between predictors. They can handle both qualitative and quantitative predictors without the need to create dummy variables. They are invariant to monotonic transformation of numeric predictors and responses, hence they avoid the thorny task of identifying the form of relationship between the predictors and the response variable.

When relationships between predictors and response and between predictors themselves, are complex and highly non-linear, regression trees might be very accurate. They clearly operate at their best with rectangular true models. On the other hand, they are outperformed by linear models when problems have a good linear structure, since they do not take into account linear relationships. Anyway, it can happen that, even with this problems, they are able to highlight some insights on the data structure that are not evident from a linear model.

In the event that data present some missing predictor in some observation, the commonly chosen strategies applied are the drastic removal of any incomplete observation or the filling of missing values with a specific statistic computed on the non-missing observations. However regression trees offer valid alternatives that allow to avoid the reduction of the training data and the consequent loss of information or the possible introduction of bias that the referred-to strategies might cause.

All the above mentioned advantages of regression trees explain the reason why they are perfectly suited for ecological data analysis. Indeed, ecological data are often complex, unbalanced and contain missing values. The relationship between variables often involve high-order interactions, therefore they are unlikely well described by linear models and other commonly used statistical modelling techniques (De'ath and Fabricius, 2000).

On the other hand, the major problem of regression trees is their non-robustness: a small variation in the data can result in a totally different estimated tree. Therefore, they suffer from high variance due to the hierarchical nature of the constructing process: a small variation in the top of the tree propagates down with effect on all the future splits. A diverse and more stable splitting criterion might attenuate this propagation effect, yet it is not possible to fully prevent it when dealing with a single tree.

A fact that should be highlighted is the lack of smoothness of the prediction surface generated by regression trees. As we have already mentioned, regression trees performed very well with true rectangular models since they split the predictor space into rectangles. If the size of the tree had no limit, the tree could approximate every decision boundary with arbitrary precision, but in practice, when the underlying model is expected to be smooth, performances of regression trees can degrade.

Regretfully, compared with other supervised learning approaches, regression trees are not very competitive in terms of predictive accuracy, although it always depends on the specific problem. Anyway, it is worth mentioning that the combination of multiple regression trees to yield to a single prediction results in a dramatic improvement of performances, at the expense of some loss in interpretability (Caruana and Niculescu-Mizil, 2006; García-Gutiérrez et al., 2015).

1.2 Random forests

As already mentioned, decision trees suffer from high variance resulting to be non perfectly robust models. This fact motivated the introduction of techniques that aggregate many trees together in order to obtain more powerful models with lower variance than a single tree. Let us remember that in the following the focus is on the regression scenario, even though many of the observations would hold also for a classification problem.

1.2.1 From bagging to random forests

Bagging

Bagging, or *bootstrap aggregation*, is a general-purpose procedure for reducing the variance of an estimated prediction function. It is often used in the context of high-variance and low-bias models. Regression trees are a perfect candidate, indeed, if they are sufficiently deep, they can be considered unbiased and it is a fact that they are strongly affected from variance.

The reliability of bagging is given by some fundamental properties of expectation and variance of random variables: let us consider B identically distributed random variables $W_1, \dots, W_B$ with expected value μ and variance σ^2. Their mean has expected value

$$\mathbb{E}\left[\frac{1}{B}\sum_{i=1}^{B} W_i\right] = \mu \tag{1.6}$$

that is, the same as each individual variable. If independence of the variables holds, their mean has variance

$$\mathrm{Var}\left(\frac{1}{B}\sum_{i=1}^{B} W_i\right) = \frac{\sigma^2}{B} \tag{1.7}$$

which is strictly smaller than the variance of each individual variable.

Bagging exploits these two properties. The idea, applied to regression trees, is to average the outcome of many different trees, trained on data drawn from the same distribution (hence, on identically distributed datasets), in order to obtain a single prediction for a given unseen observation. The resulting model would have low bias (due to the fact that the original trees are unbiased and property (1.6) holds) and lower variance than each single tree taken individually. Thus, the improvement due to the average, is only related to the reduction of the model variance.

So, theoretically, if one had B different training sets derived from the same population, a separate regression tree could be fitted on each of them, obtaining B predictive functions $\hat{f}^1(\mathbf{x}), \dots, \hat{f}^B(\mathbf{x})$. Then, averaging, the prediction would be computed as

$$\hat{f}_{\text{avg}}(\mathbf{x}) = \frac{1}{B}\sum_{i=1}^{B} \hat{f}^i(\mathbf{x}). \tag{1.8}$$

This argument presents two flaws. First, it is really unlikely to have at our disposal B different training sets from the same population. A possible solution to this problem is to perform *bootstrap* on the available training set, generating B different datasets sampled from the original one with replacement. The procedure so obtained is known as *bagging* (Breiman, 1996a).

Summing up, bagging applied to regression trees, involves the creation of B regression trees, using B bootstrapped training sets from the original dataset, and averaging of the resulting predictions.

Notice that trees have not been pruned. Indeed overfitting and variance related problems are solved via averaging. Adding pruning would help in reducing overfitting, but would add bias in each single tree and so, in the resulting model.

There is still the second issue to talk about. Property (1.7) holds only if the independence of the B variables is guaranteed. For instance, if the same B variables are only identically distributed with positive pairwise correlation ρ, the variance of the average will be

$$\operatorname{Var}\left(\frac{1}{B}\sum_{i=1}^{B} W_i\right) = \rho\sigma^2 + \frac{1-\rho}{B}\sigma^2. \tag{1.9}$$

Hence, as the number of trees B grows, the variance is dominated by the first term $\rho\sigma^2$, so that, the correlation between variables would diminish the advantages of averaging. In order to prevent this drawback, a trick can be used, that leads to the definition of random forests.

Random forests definition

A *Random forest* (*regressor*) (Breiman, 2001a) provides an improvement to the bagging procedure with a slight modification that decorrelates bagged trees enhancing the variance reduction of the average model. This tweak regards the choice of predictors when splits of the feature space $\mathcal{X}$ are performed.

More in detail, when a regression tree is grown on a bootstrapped training set, at each split, a random sample of m predictors, with $m < p$, is drawn from the full set of p predictors. These predictors are split candidates, which means that the best split must be searched only among them, excluding the other predictors. Once the B regression trees are grown as described above, the prediction is computed as in (1.8).

Therefore, random forests overcome the problem of correlation between bootstrapped trees by forcing each split to consider only a subset of the predictors. The decorrelation of trees makes the average model less variable and so, more reliable.

A typical value for m in regression tasks is $\lfloor p/3 \rfloor$, but it might vary depending on the specific case and it should be considered as a tuning parameter. Notice that if $m = p$, classic bagging is performed. A small value of m is helpful when there is a large number of correlated predictors. Indeed, the more m decreases, the more the pairwise correlation of trees ρ in the ensemble is reduced, hence, by (1.9), the variance of the average prediction decreases.

1.2.2 Deeper into random forests

Out-of-bag error estimation

Random forests have an important feature, the *out-of-bag (*OOB*) samples*: it can be proved that on average, each bagged tree is trained on 2/3 of data points in the original training set. The remaining 1/3 of the original training set is referred to as the OOB *sample* of the tree.

Thus, in order to obtain a honest estimate of the test error of a random forest, one could exploit the OOB samples of each tree in the forest, without the need of using cross-validation or a validation-set.

The idea of OOB error estimation is the following: for each observation $(\mathbf{x}_i, y_i)$, one can compute a prediction growing a new random forest that aggregates only the bagged trees trained on bootstrapped samples in which the ith observation did not appear. Once the OOB prediction has been computed for the entire original training set, the OOB error can be obtained and used as generalisation error estimate for the whole random forest.

The test error estimate so obtained is valid because the response for each observation is predicted only using trees that have been fitted without that observation. Moreover, the OOB error estimate is very close to the k-fold cross-validation error estimate and it is dramatically more convenient in computational terms, especially when the training set is very large (Hastie, Tibshirani, and Friedman, 2009; Breiman, 1996b).

Variable importance measures

Bagging and random forests typically result in improved accuracy over prediction using a single decision tree. However, the aggregation of many trees affects the interpretability of the resulting model, which is a relevant merit of decision trees. It is no longer possible to represent the model as a single tree.

Anyway, it is still possible to extract from a random forest an overall summary of the importance of each variable, that is, its contribution in predicting the response.

Let us consider for the moment a single regression tree. Each internal node n of the tree corresponds to the split of a region of the feature space $\mathcal{X}$ into two subregions. The improvement of the split criterion of an internal node n is the local importance attributed to the predictor used in the split corresponding to the considered node n. In the regression setting, the improvements is the total amount of which the RSS is decreased by the concerned split. The total importance of a predictor in the tree is the accumulation over all internal nodes of the local importance values of the predictor.

More formally, Breiman et al. (1984) proposed the following calculation of the importance of the jth predictor:

$$\mathcal{I}_j^2(T) = \sum_{\substack{n \text{ internal} \\ \text{node of } T}} i_n^2 \, \delta_{j,v(n)} \tag{1.10}$$

Here, i_n^2 represents the squared local importance of the predictor used in the split of node n; $v(n)$ returns the predictor index used in node n for the split and $\delta_{j,v(n)}$ is the Kronecker delta controlling whether the jth predictor and the $v(n)$th predictor correspond. Therefore, the squared importance of variable X_j is the sum of the squared local improvements over all internal nodes for which it was chosen as the splitting variable.

Measure (1.10) can be generalised to a random forest by averaging the importance of predictor X_j over all trees in the forest, i.e.,

$$\mathcal{I}_j^2 = \frac{1}{B}\sum_{i=1}^{B}\mathcal{I}_j^2(T_i)$$

The stabilisation introduced by averaging makes this measure of variable importance more reliable than that computed in a single tree. Moreover, one should be able to compute this measure for every predictor since the split-candidates selection operated by random forests makes possible that every single predictor is included in the model (Hastie, Tibshirani, and Friedman, 2009).

Variable importance values can be used to obtain ranking of variables by their importance within the model. Furthermore, what is so interesting about variable importance returned by tree-based methods is that it allows to measure the impact of each predictor even in a multiple regression problem, when predictors may be involved in high-order interactions. This is fundamental in biological and ecological settings. For instance, in an ecological classification scenario, Cutler et al. (2007) noticed that, even though the most important predictors in the variable importance rankings returned by random forests cannot be said to be *right or wrong*, they generally coincided closely with the expectations based on ecological understanding of the problem.

1.2.3 Reasons behind the use of random forests

The primary motivation for the introduction of random forests is to fix the instability problem that affect single decision trees. As already mentioned before, aggregating many trees smooths the hard cut decision boundary created by a single decision tree, which in turn reduces the model variance and maintains a low bias. The resulting model benefits of robustness and, in many cases, is able to outclass a single decision tree in describing the structure of the data and in generalising its predictive power.

Furthermore, as a ensemble of decision trees, random forests inherit many of the characteristics of decision trees, for instance: they are able to model complex interactions among predictors and between predictor and response variables; they are able to treat with missing data with alternatives to their complete removal; they have no distributional assumptions on the input data and they are invariant to monotonic transformations of predictor and response variables; their accuracy is not affected by multicollinearity problem.

Therefore, many applications of random forests in ecological settings have evidenced their competitive performances with respect to the best available statistical models, for both classification (Cutler et al., 2007; Gislason, Benediktsson, and Sveinsson, 2006) and regression (Prasad, Iverson, and Liaw, 2006) scenarios.

On the other side, random forests lose in interpretability with respect to decision trees. Obtaining a pictorial representation of the model is no longer possible, and understanding how exactly the model has returned a specific outcome is harder because of the contribute of many simpler decision trees. However alternative tools for the understanding of the model structure and the relationships between the predictors and the response are available, e.g., variable importance measures (Breiman, 2001a), permutation importance measure, partial dependence plots (Breiman, 2001b), etc.

Lastly it is necessary to notice that for its own nature, training a random forest is a longer process than training a single decision tree, therefore the computational effort and the execution time for building the model grows with the depth of the trees, their number inside the forest and the size of split candidates set used. Luckily reducing the number of slip candidates, each split is faster because only a subset of data is analysed. Moreover, random forests can be parallelised on multiple machines, which results in a faster computation time.

Chapter 2

Data analysis

In this chapter we inspect the data and we formally define the leading regression problems of this work. Once the formal setting is determined, we move to the description of the preprocessing and the analysis phases of the dataset, which purpose is to lay the best the foundations for the machine learning modelling phase.

2.1 Dataset description and problem statement

2.1.1 Dataset description

The dataset used in this study consists of bioclimatic predictors and land cover fraction variables for the entire Earth's surface.

Bioclimatic predictors are a set of 19 standard measures of temperature and precipitation, used frequently in ecological applications. They describe annual trends, seasonal mean climate conditions, and extreme or limiting environmental factors. A brief description, obtained by O'Donnel and Ignizio (2012), is given in Table A.1. All the predictors are numerical variables and it can be noticed that they may be considered divided into two main groups: *Temperature-Related (* TR*) predictors* (BIO01, ..., BIO11) and *Precipitation-Related (* PR*) predictors* (BIO12, ..., BIO19).

Bioclimatic data have been derived from the open access data source *WorldClim v2.0* (Fick and Hijmans, 2017), containing high spatial resolution global weather and climate data. This dataset has been built with observations averaged across a temporal range of thirty years, 1970-2000.

Code	Name	Description
IGBP12	*Croplands*	At least 60% of area is cultivated cropland.
IGBP13	*Urban and built-up lands*	At least 30% impervious surface area including building materials, asphalt and vehicles.
IGBP14	*Croplands/Natural vegetation mosaics*	Mosaics of small-scale cultivation 40-60%, with natural tree, shrub or herbaceous vegetation.

Table 2.1: Description of human-modified land cover classes (Friedl and Sulla-Menashe, 2019).

Land cover fraction variables have been obtained from the *MODIS Land Cover Type Product* (*MCD12Q1*) (Friedl and Sulla-Menashe, 2019), year 2018. Each land cover fraction variable gives the fraction, i.e. the percentage, of the associated land cover type, classified according to the *International Geosphere-Biosphere Programme* (*IGBP*) legend. The description of all 17 land cover classes is in Table A.4.

For the purpose of this work the interest falls on the land cover classes associated with human activity: *croplands* (IGBP12), *urban and built-up lands* (IGBP13) and *croplands/natural vegetation mosaics* (IGBP14). They are briefly explained in Table 2.1 and the difference among them can be immediately understood by looking at the aerial photographs in Figure 2.1.

Both datasets are global mappings of the variables over regular grids with square cells. However, because *WorldClim* and *MCD12Q1* are provided in differing coordinate reference systems, at differing spatial resolutions, these datasets have been spatially resampled onto an Icosahedral Snyder Equal Area (ISEA), aperture 3, hexagonal geodesic discrete global grid, having equal-area (2600 km^2) hexagonal cells (Sahr, White, and Kimerling, 2003).

Briefly, each observation of the dataset resulting from the spatial resampling operation is an hexagonal tile, labelled by the geographical longitude and latitude coordinates of its centroid, described by the bioclimatic predictors and the land cover fraction variables corresponding to its specific geographic region.

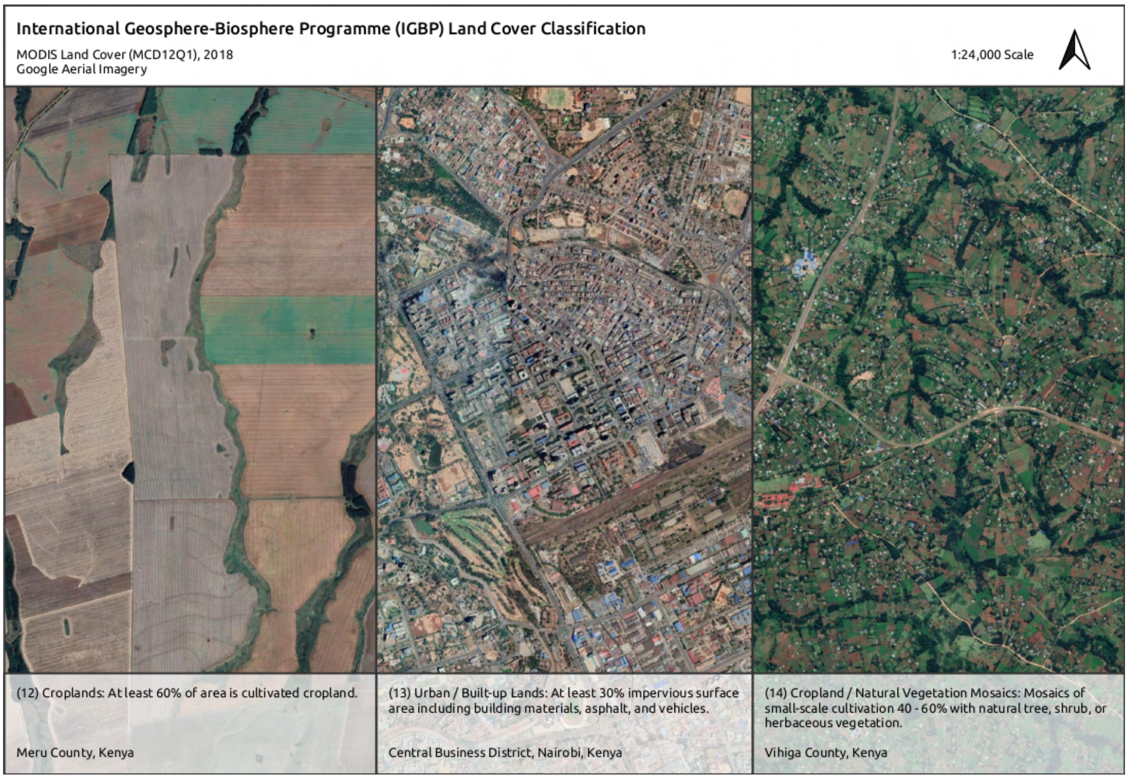

International Geosphere-Biosphere Programme (IGBP) Land Cover Classification
MODIS Land Cover (MCD12Q1), 2018
Google Aerial Imagery
1:24,000 Scale
(12) Croplands: At least 60% of area is cultivated cropland.
Meru County, Kenya
(13) Urban / Built-up Lands: At least 30% impervious surface area including building materials, asphalt, and vehicles.
Central Business District, Nairobi, Kenya
(14) Cropland / Natural Vegetation Mosaics: Mosaics of small-scale cultivation 40 - 60% with natural tree, shrub, or herbaceous vegetation.
Vihiga County, Kenya

Figure 2.1: Aerial photographs illustrating the three human-modified land cover classes in the IGBP's classification system. Respectively, *croplands* (IGBP12) on the left, *urban and built-up lands* (IGBP13) in the middle and *croplands/natural vegetation mosaics* (IGBP14) on the right. These photographs are all from Kenya, and they are all mapped at the same spatial scale.

2.1.2 Problem statement

The dataset just described can be regarded as 4 distinct elements. Bioclimatic predictors form the largest part of the dataset and they constitute the feature matrix in the regression task, that is, they can be thought as a matrix

$$\mathbf{X} \in \mathbb{R}^{n\times 19}$$

where the jth predictor corresponds to the jth column and the ith grid tile to the ith row of $\mathbf{X}$. Here n is the number of observations in the dataset[1]. Denoting the jth bioclimatic predictor with BIOj or with X_j is equivalent.

[1] In Section 2.2 a number for n will be properly defined.

The remaining three elements of the dataset are the three land cover fraction variables IGBP12, IGBP13 and IGBP14 that can be mathematically represented as three n-dimensional vectors, respectively

$$Y_1,\, Y_2,\, Y_3 \in [0,1]^n.$$

Clearly, there is a correspondence between the ith row of $\mathbf{X}$ and the ith component of Y_k, $k=1,2,3$. These variables represent three possible responses of a regression analysis. Hence, three different regression problems are defined, one for each human-modified land cover fraction variable.

So, the aim of this work is the identification, by means of tree-based methods introduced in Chapter 1, of three possible functions $\hat{f}_1, \hat{f}_2, \hat{f}_3 : \mathbb{R}^{19} \to [0,1]^n$ such that

$$\hat{y}_k = \hat{f}_k(\mathbf{x})\,, \quad k = 1,2,3 \tag{2.1}$$

i.e., three models for predicting human-modified land cover fraction $\hat{y}_k$ from bioclimatic conditions $\mathbf{x}$.

Simultaneously, given the strong descriptive power of tree-based methods, if such models exist, they may be used in order to extrapolate, from the data, eventual patterns that relates climatic conditions with human activities.

2.2 Data preprocessing and analysis

The dataset is composed of $n=196\,832$ data points, each one corresponding to a tile of the hexagonal geodesic discrete global grid on which the Earth's surface is represented. However, since not all the observations might be relevant, essential and reliable, preprocessing is required. The operations involved in this phase ensure a higher data quality, simplify and accelerate the training process, prevents the learning methods from integrating useless or misleading information and from being affected by missing or noisy values.

Moreover, even though we have seen that tree-based models are non-parametric models, that is, there are no specific distributional assumptions on the input data, it is important to analyse the dataset in order to put the learning methods in the best position for modelling the information that data carry along.

Thus, all the data points have been object of preprocessing and analysis regarding grid tiles composition, missing values, outliers and correlation.

2.2.1 Identification of land tiles

As the Earth's surface is divided between lands and seas, it is necessary to consider that many of the grid tiles describe only water bodies (IGBP00, including oceans, seas, lakes). So, selecting only grid tiles corresponding to lands was required.

It is obvious that grid tiles are not clearly classifiable into *only-land tiles* and *only water tiles*, because many of them describe areas in which water bodies and lands meet. Hence a threshold of 50% land covering percentage has been set in order to select the grid tiles that would have been subject of the analysis. We generically refer to the so selected grid tiles as *land tiles*. They form the 28.59% of the original dataset.

2.2.2 Dealing with missing values

When processing real-world data it is often possible to find missing information, whose absence might be motivated by many factors, such as impossibility to collect or compute variables, errors in the collection phase or in the data entry process, etc. Therefore, a fundamental step in the preprocessing phase has been seeking for missing values in the dataset and treating them in the best way.

A first action done about missing values has been keeping only data points having at least one bioclimatic predictor available. Luckily only the 0.71% of the land tiles did not satisfy this assumption. The result was a new dataset composed of $n = 55\,867$ observations (28.38% of the original dataset, 99.29% of the land tiles).

Then we have checked for the presence of eventual missing values in the feature matrix $\mathbf{X}$ and in the response variables Y_1, Y_2 and Y_3. Responses did not present any missing value, while only seven rows of the feature matrix presented one missing value each, all of them in the predictor BIO03, *isothermality*. Additional investigation proved that such data points are geographically located in the frigid climatic zones, whose very low temperatures do not always have enough variability to allow the computation of the isothermality measure. Considering the fact that the above mentioned areas are poorly connected with human activities, the easiest and most convenient solution to treat the missing values has been the removal of such data points from the dataset.

In conclusion, the resulting dataset is composed of $n = 55\,860$ observations (still 28.38% of the original dataset, 99.29% of the land tiles), each one described by 19

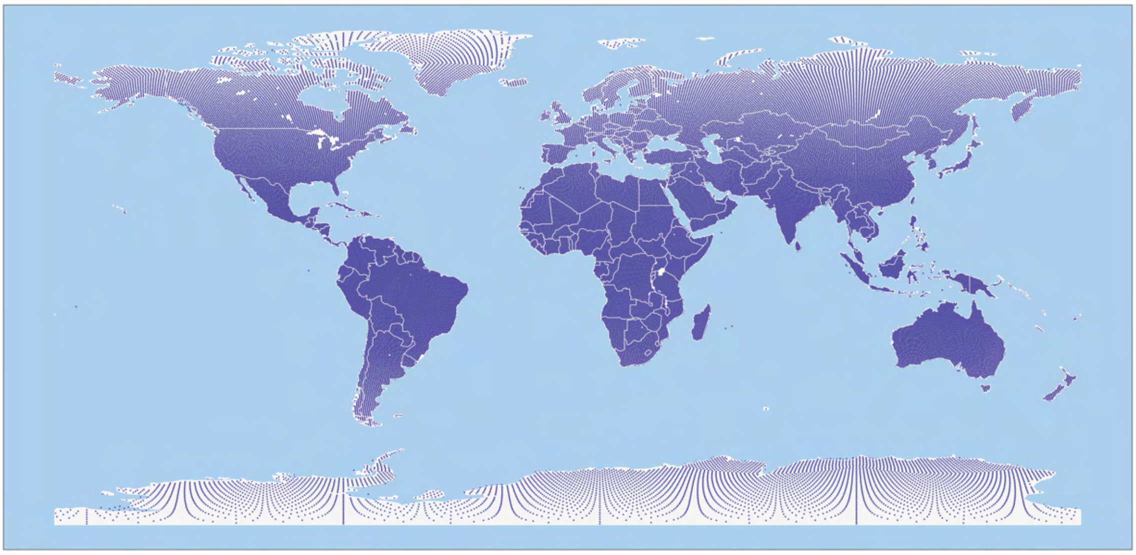


Figure 2.2: Dataset geographic map. Each violet point represents an observation in the dataset after preprocessing. All the points are equidistant from the neighbouring observations, but the projection of the *spherical* surface of the Earth on a flat surface causes the expansion on the upper and lower sides of the map (*equidistant cylindrical map projection*).

bioclimatic predictors and associated to 3 possible land cover fraction variables without any missing information. The geographic map of the data points is in Figure 2.2.

2.2.3 Dealing with outliers

It is always a good practice to look for unusual values in the dataset, which might affect the learning phase. Indeed, if the trained models have *learnt* outliers, the resulting patterns could be significantly different from the real patterns without the outliers. Even decision trees, that are generally considered to be robust to outliers, might be misled by the presence of outliers (Piramuthu, 2008).

This becomes a dangerous problem when the unusual value of an observation is caused by an error somewhere in the data-gathering phase, because the model will learn an incorrect information. However, when outliers derive from natural variations of the data generation process, they must be taken into account because they are an intrinsic constituent of the data distribution that one wants to study, learn and describe. Therefore, in the latter case, even if the removal of outliers could result in the reinforcement of the model statistical power, their presence is essential for the creation of a realistic and suitable model that guarantees the inherent characteristics of the training data.

In this study outliers have been identified for each bioclimatic predictor with the interquartile method, that quantifies the mildness/extremeness of each value in relation to the overall empirical distribution of the predictor itself. Then three operations have been performed for each bioclimatic predictor:

1. count of the outliers;
2. count of the outliers for each climatic zone[2];
3. graphic analysis of the empirical distribution by means of histograms, kernel densities and boxplots.

In Table A.2 it is possible to find the number of outliers for each bioclimatic predictor. Looking at the table it is evident that only four predictors do not have outliers, while the rest has unusual values in percentages that vary between 0.09% and 12.48%.

An apparent aspect for each bioclimatic predictor is the presence of only either *upper outliers* (values above the maximum whisker) or *lower outliers* (values below the minimum whisker). In particular, among the predictors with more outliers, TR predictors present lower outliers, while PR predictors present upper outliers. In other terms, outliers are related to low temperature and heavy precipitation.

The same conclusion can be drawn by looking at the empirical distributions in Section A.1.3. Notice that, due to the huge values of outliers in the PR predictors,

[2] On the basis of latitudinal extent, the Globe is divided into three *climatic zones*, identified by the major circles of latitude and by uniform inner climatic characteristics:

Frigid Zones Respectively North, and South. The North Frigid Zone is lower bounded by the Arctic Circle at 66°33' N. The South Frigid Zone is upper bounded by the Antarctic Circle at 66°33' S.

Temperate Zones Respectively North, and South. The North Temperate Zone is between the Arctic Circle at 66°33' N and the Tropic of Cancer at 23°27' N. The South Temperate Zone is between the Tropic of Capricorn at 23°27' S and the Antarctic Circle at 66°33' S.

Torrid Zone It is between the Tropic of Cancer at 23°27' N and the Tropic of Capricorn at 23°27' S.

they have been excluded from the colorbar scale in the geographic maps, to allow differences among the inner values of the distributions to be perceived. The same colour of the maximum value represented in the colorbar is used for the outliers.

Lastly, the count of outliers falling in each climatic zone has been performed and the result can be seen in Table A.3. Here, it is possible to see that almost all the outliers of the TR predictors, with a significant number of outliers, are geographically located in the Frigid Zones (North and South); conversely, almost all the outliers in the PR predictors, with a significant number of outliers, are geographically located in the Torrid Zone. Matching observations can be done in the geographical representation of outliers in Figure A.1.

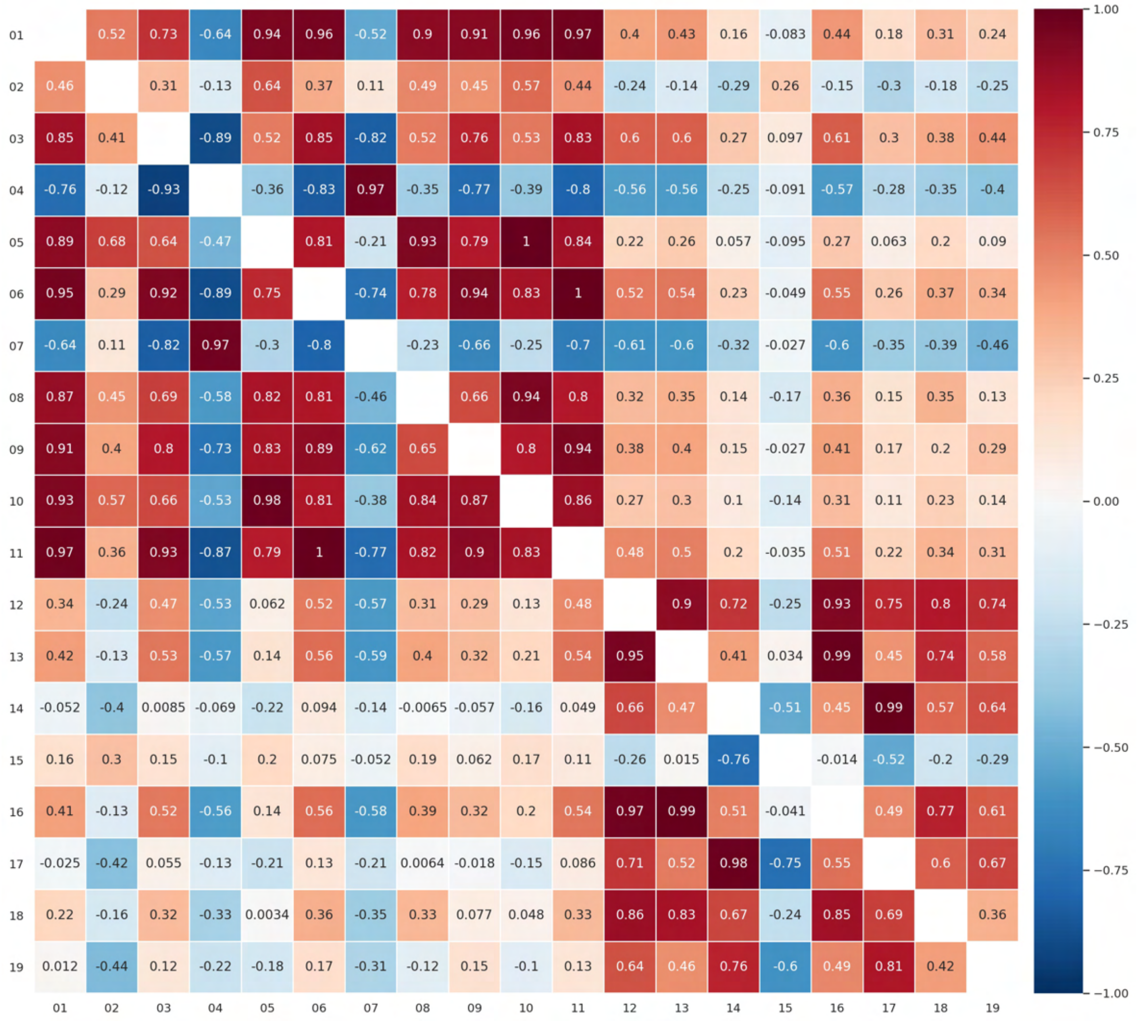
A reliable interpretation of the above considerations could be the following: observed outliers are due to natural variations of the climate over the Globe at local level. It is typical that very low temperature values are detected in the Frigid Zones, even if they are unusual in the rest of the Globe; similarly it is natural that very heavy precipitation values are detected in the Torrid Zone.

Thus, one should deduce that the high number of outliers is caused by local climatic variations on the Earth's surface that cannot be incorporated in a data distribution on a large scale. Hence, it was necessary to keep the outliers in the bioclimatic predictors in order to preserve and learn such local climatic differences.

2.2.4 Data Correlation

The correlation matrix of bioclimatic predictors can be visualised in Figure 2.3. It hides a particularity: the correlation coefficients contained in the triangular lower half have been computed with the Spearman method, while in the triangular upper half the Pearson method has been used. Thanks to the symmetry of correlation matrices there is no information loss with this compact representation.

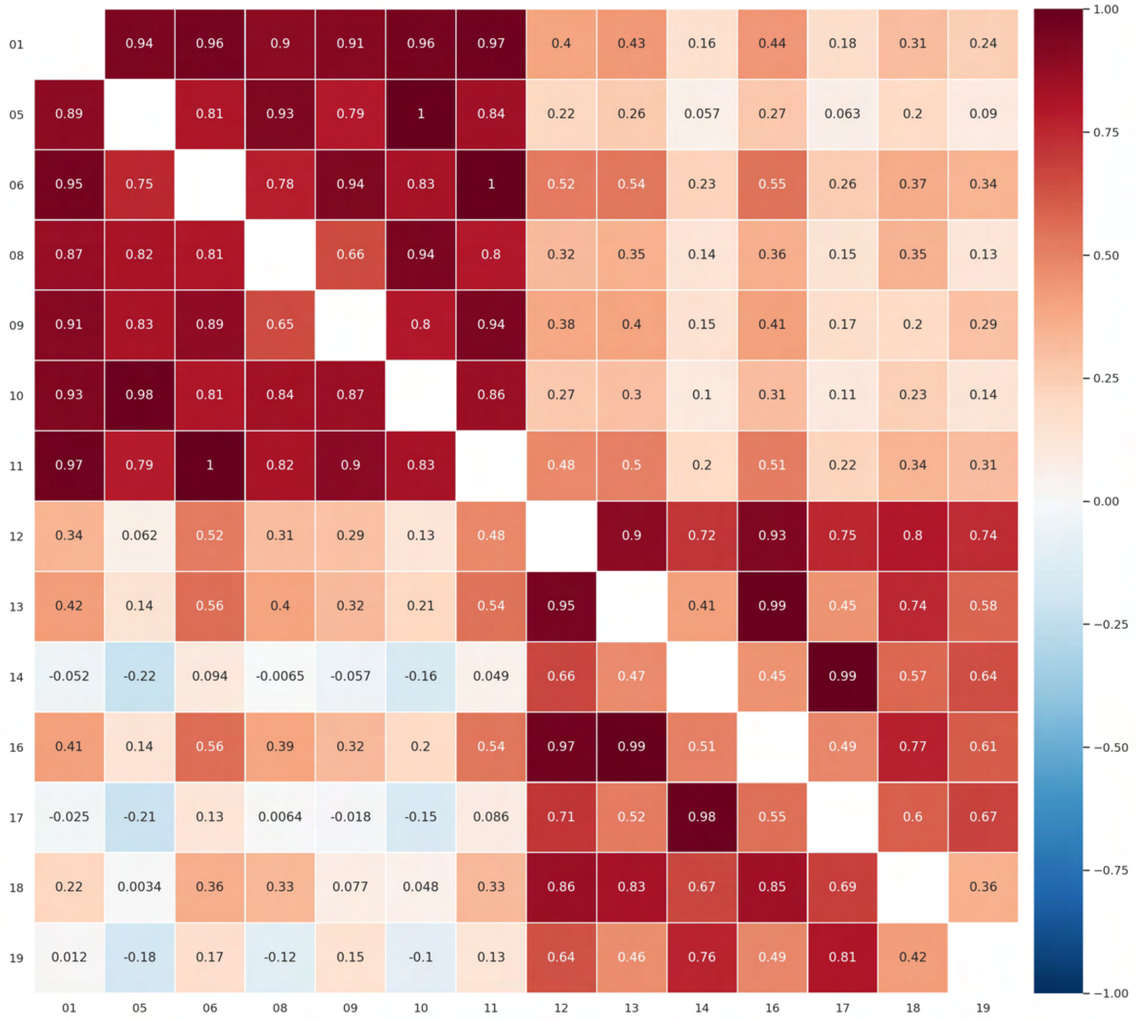
Let us point out that Pearson correlation coefficients identify only linear relationships between pairs of continuous variables, while Spearman correlation coefficients are able to evaluate monotonic relationships between them. More precisely, there are differences in the relationships that the two methods highlight: the Spearman method can identify a larger variety of relationship but it is not capable of saying whether or not they are linear, which is a prerogative of the Pearson coefficient.



	01	02	03	04	05	06	07	08	09	10	11	12	13	14	15	16	17	18	19
01		0.52	0.73	-0.64	0.94	0.96	-0.52	0.9	0.91	0.96	0.97	0.4	0.43	0.16	-0.083	0.44	0.18	0.31	0.24
02	0.46		0.31	-0.13	0.64	0.37	0.11	0.49	0.45	0.57	0.44	-0.24	-0.14	-0.29	0.26	-0.15	-0.3	-0.18	-0.25
03	0.85	0.41		-0.89	0.52	0.85	-0.82	0.52	0.76	0.53	0.83	0.6	0.6	0.27	0.097	0.61	0.3	0.38	0.44
04	-0.76	-0.12	-0.93		-0.36	-0.83	0.97	-0.35	-0.77	-0.39	-0.8	-0.56	-0.56	-0.25	-0.091	-0.57	-0.28	-0.35	-0.4
05	0.89	0.68	0.64	-0.47		0.81	-0.21	0.93	0.79	1	0.84	0.22	0.26	0.057	-0.095	0.27	0.063	0.2	0.09
06	0.95	0.29	0.92	-0.89	0.75		-0.74	0.78	0.94	0.83	1	0.52	0.54	0.23	-0.049	0.55	0.26	0.37	0.34
07	-0.64	0.11	-0.82	0.97	-0.3	-0.8		-0.23	-0.66	-0.25	-0.7	-0.61	-0.6	-0.32	-0.027	-0.6	-0.35	-0.39	-0.46
08	0.87	0.45	0.69	-0.58	0.82	0.81	-0.46		0.66	0.94	0.8	0.32	0.35	0.14	-0.17	0.36	0.15	0.35	0.13
09	0.91	0.4	0.8	-0.73	0.83	0.89	-0.62	0.65		0.8	0.94	0.38	0.4	0.15	-0.027	0.41	0.17	0.2	0.29
10	0.93	0.57	0.66	-0.53	0.98	0.81	-0.38	0.84	0.87		0.86	0.27	0.3	0.1	-0.14	0.31	0.11	0.23	0.14
11	0.97	0.36	0.93	-0.87	0.79	1	-0.77	0.82	0.9	0.83		0.48	0.5	0.2	-0.035	0.51	0.22	0.34	0.31
12	0.34	-0.24	0.47	-0.53	0.062	0.52	-0.57	0.31	0.29	0.13	0.48		0.9	0.72	-0.25	0.93	0.75	0.8	0.74
13	0.42	-0.13	0.53	-0.57	0.14	0.56	-0.59	0.4	0.32	0.21	0.54	0.95		0.41	0.034	0.99	0.45	0.74	0.58
14	-0.052	-0.4	0.0085	-0.069	-0.22	0.094	-0.14	-0.0065	-0.057	-0.16	0.049	0.66	0.47		-0.51	0.45	0.99	0.57	0.64
15	0.16	0.3	0.15	-0.1	0.2	0.075	-0.052	0.19	0.062	0.17	0.11	-0.26	0.015	-0.76		-0.014	-0.52	-0.2	-0.29
16	0.41	-0.13	0.52	-0.56	0.14	0.56	-0.58	0.39	0.32	0.2	0.54	0.97	0.99	0.51	-0.041		0.49	0.77	0.61
17	-0.025	-0.42	0.055	-0.13	-0.21	0.13	-0.21	0.0064	-0.018	-0.15	0.086	0.71	0.52	0.98	-0.75	0.55		0.6	0.67
18	0.22	-0.16	0.32	-0.33	0.0034	0.36	-0.35	0.33	0.077	0.048	0.33	0.86	0.83	0.67	-0.24	0.85	0.69		0.36
19	0.012	-0.44	0.12	-0.22	-0.18	0.17	-0.31	-0.12	0.15	-0.1	0.13	0.64	0.46	0.76	-0.6	0.49	0.81	0.42	

Figure 2.3: Correlation matrix of the bioclimatic predictors. The lower triangular half of the matrix contains the Spearman correlation coefficients, while the upper triangular half contains the Pearson correlation coefficients. Each predictor BIO01,. . . , BIO19 is denoted only with its numerical id (e.g., BIO01 → 01). Correlation values always belong to the interval $[-1, 1]$, where the value 1 identify a perfect positive correlation and the value -1 a perfect negative correlation. The value 0 denotes the absence of any correlation.

Anyway, it is evident that both methods emphasise the same correlation patterns among the bioclimatic predictors with really small differences. Let us exclude for the moment the bioclimatic predictors that describe temperature or precipitation variability, that are BIO02, BIO03, BIO04, BIO07 and BIO15 (Figure 2.4). It is



	01	05	06	08	09	10	11	12	13	14	16	17	18	19
01		0.94	0.96	0.9	0.91	0.96	0.97	0.4	0.43	0.16	0.44	0.18	0.31	0.24
05	0.89		0.81	0.93	0.79	1	0.84	0.22	0.26	0.057	0.27	0.063	0.2	0.09
06	0.95	0.75		0.78	0.94	0.83	1	0.52	0.54	0.23	0.55	0.26	0.37	0.34
08	0.87	0.82	0.81		0.66	0.94	0.8	0.32	0.35	0.14	0.36	0.15	0.35	0.13
09	0.91	0.83	0.89	0.65		0.8	0.94	0.38	0.4	0.15	0.41	0.17	0.2	0.29
10	0.93	0.98	0.81	0.84	0.87		0.86	0.27	0.3	0.1	0.31	0.11	0.23	0.14
11	0.97	0.79	1	0.82	0.9	0.83		0.48	0.5	0.2	0.51	0.22	0.34	0.31
12	0.34	0.062	0.52	0.31	0.29	0.13	0.48		0.9	0.72	0.93	0.75	0.8	0.74
13	0.42	0.14	0.56	0.4	0.32	0.21	0.54	0.95		0.41	0.99	0.45	0.74	0.58
14	-0.052	-0.22	0.094	-0.0065	-0.057	-0.16	0.049	0.66	0.47		0.45	0.99	0.57	0.64
16	0.41	0.14	0.56	0.39	0.32	0.2	0.54	0.97	0.99	0.51		0.49	0.77	0.61
17	-0.025	-0.21	0.13	0.0064	-0.018	-0.15	0.086	0.71	0.52	0.98	0.55		0.6	0.67
18	0.22	0.0034	0.36	0.33	0.077	0.048	0.33	0.86	0.83	0.67	0.85	0.69		0.36
19	0.012	-0.18	0.17	-0.12	0.15	-0.1	0.13	0.64	0.46	0.76	0.49	0.81	0.42	

Figure 2.4: Correlation matrix of the bioclimatic predictors excluding that quantifying variability (BIO02, BIO03, BIO04, BIO07 and BIO15). The matrix is structured as that in Figure 2.3.

possible to notice that the remaining TR predictors (now quantifying absolute temperatures) are strongly positively correlated among them, and the same holds for the remaining PR predictors (that now quantify absolute precipitation). When looking at correlations between the two groups, they appear weakly, and sometimes not at all, correlated. These results might suggest that, at global level, temperature and precipitation do not vary together; this fact might be false, instead, on a local scale.

When including the predictors that describe variability in the temperature and precipitation measures, the correlation values computed among the entire set of predictors appear to be more complicated to interpret. One could deduce that variability measures might be more relevant on a local scale while they lose power when considered on a global scale.

If predictors are highly correlated, we talk about *multicollinearity*. This is the case of our bioclimatic predictors indeed. Fortunately for us, large pairwise correlation among predictors is not of concern for tree-based models (Piramuthu, 2008), whose performances are immune to multicollinearity by nature; indeed each split of the predictor space is performed by means of one, and only one, predictor, and it cannot be altered by the effect of the others.

However even with tree-based models multicollinearity has a side effect: imagine that a group of very influential predictors is highly correlated with other predictors that are weakly, or not at all, associated with the response variable. In the tree-growing process, the latter predictors might appear as well suited as the truly important predictors for splitting the predictor space (Strobl et al., 2008). It follows that interpreting the importance of predictors in the model is harder and requires more care than when treating uncorrelated, or slightly correlated, predictors.

2.2.5 Analysis of the response variables

The final step of the dataset analysis has been dedicated to the response variables IGBP12, IGBP13 and IGBP14, and their interactions with the bioclimatic predictors.

In each tile of the grid that approximates the Earth's surface, the three response variables can take values in the interval $[0, 1]$, that is,

$$Y_k^{(i)} \in [0,1] \quad \forall\, i = 1, \dots, n, \;\; k = 1, 2, 3.$$

The value 0 identifies the absence and the value 1 identifies the absolute dominance, of the land cover class in the surface area described by the grid tile. Each intermediate value corresponds to the percentage of the area occupied by the land cover class.

Empirical distributions

The fact that tree-based models do not have specific assumptions on the distribution of the input data is a real relief in ecological analysis. Indeed real-world ecological data have often uncomfortable distributions that do not satisfy the assumptions of many statistical methods. The three possible response variables of this study, IGBP12, IGBP13 and IGBP14, are no exception: their distributions are unimodal and strongly positive skewed.

Table 2.2 presents a numerical summary of their empirical distributions, while graphical representations (histograms, kernel densities and boxplots) are in Figures A.21a, A.23a, A.24a. It is evident that the largest percentage of the values in the response variables is equal to 0: approximately 60% in IGBP12 and IGBP13, even 80% in IGBP14. Looking individually at each variable we can notice what follows:

IGBP12 has the distribution with largest variability among the others. More than half of the population has value 0, but the remaining 38% of the observations spreads the entire interval $(0, 1]$, even though, it maintains a positive skewed distribution toward 0 (Figure A.21c).

IGBP13 has a distribution in which more than half of the values is equal to 0, similarly to IGBP12, but less variability of the positive fractions is present. Indeed the positive values do not spread all over the interval $(0, 1]$ but their density drops to zero rapidly the farther we are from 0 (Figure A.23c) and the maximum value is only 0.8695.

IGBP14 has a distribution that collapses on 0. Each positive value is considered to be an outlier. Indeed, only less than the 20% of the observations has a strictly positive value. Variability of the positive fractions is better than that of IGBP13, with a slightly heavier tail toward the larger values (Figure A.24c), but again value 1 is never achieved.

It is a fact that the strongly positive skewed distribution of the three response variables might be an obstacle in the regression problem, introducing a bias into the models in favour of the prediction of null values, or at least, values close to zero. Therefore, this problem needs to be taken into consideration when analysing the performances of the models.

Pairwise relationships between responses and bioclimatic predictors

The last step of the data analysis has focused on the research of possible pairwise relationships between the human-modified land cover fraction variables and the

	IGBP12	IGBP13	IGBP14
Count	55860	55860	55860
Count of zero fractions	34645	35262	44808
Percentage of zero fractions	62.02%	63.13%	80.21%
Minimum	0.0	0.0	0.0
Minimum whisker	0.0	0.0	0.0
1st quartile	0.0	0.0	0.0
Median	0.0	0.0	0.0
3rd quartile	0.0088	0.0012	0.0
Maximum whisker	0.0219	0.0031	0.0
Maximum	1.0	0.8695	0.9246
Mean	0.0791	0.0051	0.0096
Standard deviation	0.2101	0.0248	0.0491
Sample skewness*	3.014	13.01	8.75
Sample skewness of positive fractions*	1.41	8.16	3.803

* Computed by means of the *adjusted Fisher-Pearson standardised moment coefficient.*

Table 2.2: Numerical summary of the empirical distribution of the human-modified land cover fraction variables.

bioclimatic predictors. Indeed, this work aims exactly at finding patterns inside the bioclimatic predictors that could relate to the land covers. As it has already been remarked, in an ecological analysis it is really difficult to find simple connections between the predictors and the response, but it is worth making an attempt, even if only for having an idea of how powerful the predictors might be.

Correlation coefficients among the response variables and the bioclimatic predictors have been computed and they are displayed in Figures A.25, A.28 and A.31. Both, Pearson and Spearman, correlation coefficients have been used, but it is evident that there are neither linear nor monotonic relationships among the response variables and the bioclimatic predictors. Generally speaking the Spearman coefficients show a larger monotonic correlation with the PR predictors, but all the coefficients do not even pass the value 0.4, therefore any further interpretation might be misleading.

An additional investigation has been performed by means of pairwise scatter plots (Figures A.26, A.29, A.32) and 2-dimensional, 50 × 50 equal-area bins histograms (Figures A.27, A.30, A.33). For each pair predictor-response (X_j, Y_k), $j=1,\ldots,19$ and $k=1,2,3$, both the plots have been created, in which the bioclimatic predictor is the independent variable and the response is the dependent variable.

Similar considerations might be done referring either to the scatter plots or to the 2D histograms. Then, let us take into consideration the 2D histograms because they better allow perceiving the concentration of points in each point of the plane. It is immediately evident that the largest concentration of data points is in the lowest row of the grid describing the plane, which means close to the null value for the response variable and, apparently, in a sufficiently independent way of the value of the bioclimatic predictor on duty. This fact is clearly due to the strongly positive skewed distributions of the response variables but also means that there is no specific range of values for each bioclimatic predictor for which the response variables are more likely to take the zero value.

If we look only at the 2D histograms for the response IGBP12 (Figure A.27), it happens that the data points corresponding to positive values of the response variable are distributed like vertical slots of the plane (the purple vertical stripes in the plots). This means that, even if there is no apparent specific pattern that relates the response variable with each bioclimatic predictor taken individually, it is possible to identify ranges of values for each predictor in which it is very unlikely that there will be a land cover composed of croplands (the black vertical stripes). This kind of patterns might be easily identified by tree-based models.

As for the responses IGBP13 and IGBP14, similar ranges of predictor values might be identified, even though the plots do not present the vertical slots that characterise IGBP12. Indeed no data points are present in the upper part of each plot because of the more positive skewed distributions of the non-null fractions of IGBP13 and IGBP14 and the smaller number of positive values for these responses.

It must be remarked that this is a general impression and that on singular pairs predictor-response it might be possible to describe a slightly more detailed pattern, but none of them appears to be so relevant to be worth mentioning.

Chapter 3

Regression analysis with decision trees

In the course of this chapter, the solution of the regression problems (2.1) by means of decision trees is described. Therein, the theoretical background, introduced in Chapter 1, meets our real-world dataset, described in Chapter 2, and the technical obstacles of the implementation.

Retain that, three distinct regression problems have been developed in parallel. The outline of the regression analysis on each response variable has been mostly the same, but every one has been treated individually. The results are completely contrasting among the three response variables, indeed, while regression trees seem to be a good modelling choice for cropland fractions (IGBP12), the other response variables related to urban areas (IGBP13) and mosaics of croplands and natural vegetation (IGBP14) do not react just as well. For these reasons and for a better understanding of the way the regression analysis has been developed, the explanation of each step will be focused only on the response IGBP12. At the end of the chapter an overview of the outcomes for the other two responses is available.

3.1 Implementation

The code produced for the results described in this chapter is in `Python 3`. The machine learning tools employed are supplied by the `scikit-learn` library (Pedregosa et al., 2011). Many other Python libraries have provided the fundamental tools for processing, organisation and visualisation steps, in particular: `pandas` (McKinney, 2010; Reback et al., 2020), `geopandas` (Jordahl et al., 2020), `NumPy` (Harris et al., 2020), `seaborn` (Waskom et al., 2020), `matplotlib` (Hunter, 2007; Caswell et al., 2020).

Let us look at a few technical matters, before moving to the true regression analysis.

Scoring metrics and estimation strategy of generalisation performances

The performance of a regression tree can be evaluated in many ways, indeed, there exist various scoring metrics in the literature that can be applied to a regression model in order to estimate the goodness of the prediction on some input data. Regression trees, as described in Section 1.1.1, are built in order to solve the minimisation problem (1.1) on the tree impurity, then, the natural scoring metric that one can use is the RSS, or equivalently, the MSE[1], computed among the true and the predicted response.

RSS and MSE are absolute measures of the lack of fit of the model to the training data. However, since their unit is the same as the response values, understanding which is a good value is not always easy. The unique rule that holds is *the lower, the better*, but *how low is low enough?* As a matter of fact, in this work, where response values are in $[0, 1]$, MSE appeared to be misleading because its values seemed to be always small, even when the model performances were not excellent.

In view of the above, we decided to support the analysis with a second scoring metric, whose value is independent of the scale of the response: the *coefficient of determination* R^2, i.e.

$$R^2 = 1 - \frac{\text{RSS}}{\sum\limits_{i=1}^{n}(y_i - \overline{y})^2}.$$

This statistic measures the percentage of variance in the response variable that is explained by the predictors, or, in other words, it quantifies the goodness of fit of the model to the training data. Its value is generally inside the interval $[0, 1]$ (indeed, it represents a percentage). A R^2 of 1 means that the regression predictions perfectly fit the true response values, while a value 0 identifies the model whose predictions are always equal to the expected response, independently of the predictors. Anyway,

[1] *Mean Squared Error*, i.e., $\text{MSE}(\hat{f}) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}(\mathbf{x}_i))^2$.

it happens that R^2 assumes negative values when the model fits the data worse than the mean response.

It is important to notice that MSE measures the lack of fit, while R^2 measures the goodness of fit of the model prediction with respect to the true response variable, then, the first metric is to be minimised, while the second is to be maximised. Throughout this work, in order to have consistency in the scoring metric objective (minimisation or maximisation), we decided to apply a simple linear transformation to the R^2 statistic and use $1 - R^2$ in its place. The so obtained scoring metric is equivalent to R^2, in the sense that there is no information loss by using one or the other, but, in order to identify the best-performing model, this new measure must be minimised, as well as the MSE.

In conclusion, the scoring metrics used in this work to evaluate the performances of the trained models are the MSE and $1-R^2$. In particular, cooperating with 10-fold cross-validation, they have been used to estimate the performances over unseen data.

Division of the dataset into training and test sets

Before starting the regression analysis, we have randomly split the whole dataset into training (90%) and test (10%) subsets to save part of the data points from being learnt by the models and exploit them for model evaluation. Once the two subsets were created, we ensured that both of them kept approximately intact the distribution of the original response variable. Apparently, random splitting seemed to guarantee this property, then no further deed on the two subsets has been required.

3.2 Tree growing and pruning

In Section 1.1.1 and 1.1.2, the processes of growing and pruning a regression tree have been extensively described. Here, we explain how they have been put into practice when dealing with our real-world data.

The main tool of this phase has been the function `DecisionTreeRegressor`, provided by `scikit-learn`. It is able to build a full unpruned regression tree, but also, it allows large customisation of the model growing and pruning by means of a set of hyperparameters: it permits to set the split quality measure and the strategy used to select the best split when growing the tree, to apply forms of pre-pruning and cost-complexity pruning.

Throughout the entire investigation, the approach used to grow the regression trees has been the recursive binary splitting, as it is described in Section 1.1.1. The criterion employed to measure the impurity of the trees and the subregions of the predictor space has been the MSE, which is nothing but an unbiased version of the RSS used in the theoretical description of the greedy algorithm.

Tree growing and pruning proceeded according to the following outline:

1. the effect of the maximum depth of the tree on the scoring metrics has been inspected;
2. reliable optimal values of the complexity parameter α have been researched;
3. combinations of pre-pruning and cost-complexity pruning have been attempted in order to identify possible relevant models for this work. During this step, an alternative random splitting approach has been employed too.

Let us look into more in detail at each of these stages.

3.2.1 Controlling the maximum depth of the tree

This first step of the regression analysis had the simple purpose of familiarising with the regression trees and understanding how they were going to be able to model the distribution of cropland fractions around the globe.

The results can be summarised with Figure B.1, which describes the evolution of the scoring metrics as the maximum depth of the regression tree increases, transforming the tree root into a full unpruned tree. The scoring metrics have been evaluated on the training set (blue line) and estimated for unseen data (orange line). Three important observations should be noticed by looking at Figure B.1:

- The two scoring metrics equally describe the model performances, even though using two different ranges of values. If we should consider only the tree MSE, its values when the tree is only a few levels deep, are close to 0.03, which is a small error per se. However, when we observe the correspondent $1-R^2$ values, they are close to 0.6, which means that, at such shallow depth, the models are not yet well-performing, even if their MSE might suggest so. This is the confirmation that including the $1-R^2$ scoring metric has been a wise choice.

- It is evident how, the more the tree grows deeper into the predictor space, the more it is overfitting the training data. Indeed, the training MSE and $1-R^2$ drop approximately to zero around depth 25.

- Although the training scores might be considered almost zero after depth 25, their generalisation estimates reach the minimum around depth 13 (MSE: 0.016, $1-R^2$: 0.368), just after a steep decrease, and slowly start increasing again thereafter, levelling off once crossed depth 25. This fact suggests that, after a given depth, the trees do not benefit too much from further growing, which should be kept under control instead.

No matter how simple these plots might seem, they have already highlighted very crucial aspects of the regression analysis, first of all, that pruning is definitely required in order to train models able to generalise their predictive power.

3.2.2 Researching complexity parameters

The second stage of the analysis has focused on cost-complexity pruning and, specifically, on the identification of a reliable optimal complexity parameter α.

When we have applied cost-complexity pruning to the full regression tree that can be fitted on our data (denoted by T_0), the set A of all the possible complexity parameters α has been returned. Unfortunately, the cardinality of this set was really large, i.e., $|A| = 18\,168$. Since each complexity parameter corresponds to a specific subtree of T_0, in order to find the best-pruned tree over unseen data, it seemed unreasonable to solve the minimisation problem (1.5) on the entire set A, because it would have been extremely computationally demanding. Therefore we had to deal with this inconvenient more cautiously, by following two steps:

1. First, solving the minimisation problem (1.5) on a reasonably small random sample of A with the purpose of identifying a temporary optimal complexity parameter $\tilde{\alpha}_{min}$ and, if present, the 1SE complexity parameter $\tilde{\alpha}_{1\text{SE}}$ (that is, obtained with the 1SE rule).

2. Then, solving (1.5) in the subset of A obtained by including all the complexity parameters contained in a reasonably small neighbourhood of $\tilde{\alpha}_{min}$ and $\tilde{\alpha}_{1\text{SE}}$, in order to identify the optimal complexity parameter α_{min} and, if present, its 1SE complexity parameter $\alpha_{1\text{SE}}$.

By means of this approach it has been possible to reduce the number of complexity parameters to test, that is, the number of regression trees to validate through 10-fold cross-validation, from 18 168 to a few hundred in total. Let us take a closer look at the previous steps.

It is important to remark that, since the distribution of the complexity parameters is strongly positively skewed in the interval $[0, 0.002]$, a simple random sampling of restrained size[2], would have had a hard time in representing correctly the distribution of the data points. Thus, the following strategy has been used: I) Orders of magnitude[3] of the complexity parameters in A have been computed. II) In each subset of A of the form $A \cap [10^k, 10^{k+1}]$, the $\lceil 1\% \rceil$ of the elements has been randomly sampled, so that, every order of magnitude has been represented by at least one parameter, and more populated subsets have been represented by larger samples. III) The union of all the complexity parameters so identified has composed the set on which the minimisation problem (1.5) has been solved in step 1.

The evolution of the generalisation scoring metrics estimates as function of the sampled complexity parameters at step 1 are shown in Figure B.2. Both scoring metrics agreed on the temporary optimal and 1SE complexity parameters values, contained in Table 3.1.

At this point, a reasonable assumption was that the magnitude of the optimal complexity parameter α_{min} was the same, or at least very close, to that of the just found $\tilde{\alpha}_{min}$. Therefore, during step 2, we have researched for α_{min} in a neighbourhood of $\tilde{\alpha}_{min}$ and $\tilde{\alpha}_{se}$ of reasonable size, i.e., $A \cap [9\cdot10^{-6},\, 3\cdot10^{-5}]$.

Similarly as before, Figure B.3 and Table 3.2 summarise the results of the solution of (1.5) over the selected complexity parameters at step 2. In this case, when the 1SE rule has been applied, the two scoring metrics have identified two distinct values for $\tilde{\alpha}_{se}$.

[2] Approximately 200 parameters.

[3] Given a real number x, its *order of magnitude* is the smallest power of 10 used to represent that number. So, if $x = a \cdot 10^k$, then k is its order of magnitude.

	$\boldsymbol{\alpha}$ (10^{-5})	MSE	$\mathbf{1-}\boldsymbol{R}^{\mathbf{2}}$
$\tilde{\alpha}_{min}$	1.08083	0.01562	0.35176
$\tilde{\alpha}_{1\text{SE}}$	1.60994	0.01571	0.35361

Table 3.1: Complexity parameters selected during step 1.

	$\boldsymbol{\alpha}$ (10^{-5})	MSE	$\mathbf{1-}\boldsymbol{R}^{\mathbf{2}}$
α_{min}	1.13655	0.01552	0.34953
$\alpha_{1\text{SE, MSE}}$	1.94197	0.01580	-
$\alpha_{1\text{SE}, 1-R^2}$	1.52811	-	0.35240

Table 3.2: Complexity parameters selected during step 2.

At the end of the procedure, we had three distinct possible complexity parameters to use for pruning our full regression tree T_0: α_{min}, $\alpha_{1\text{SE, MSE}}$ and $\alpha_{1\text{SE}, 1-R^2}$.

3.2.3 Combining pre-pruning and cost-complexity pruning

During the last stage of the tree growing and pruning phase, we have let interact the facts of the theoretical background of pruning with the reality of our data. From a purely theoretical perspective, cost-complexity pruning is preferable to a pre-pruning strategy in view of the fact that it is grounded on a rigorous mathematical construction, while pre-pruning strategies are more heuristic approaches. In practice, everything depends on the specific regression problem and on the data it is based on. Therefore it would not be wise the a priori exclusion of some pre-pruning techniques that might reveal useful and effective.

So, we have decided to take into consideration not only regression trees simplified with cost-complexity pruning but, to let this approach interact with pre-pruning strategies based on limiting the maximum depth of the trees and fixing the minimum number of samples required to split an internal node of the tree. In other words, we have inspected the generalisation performance estimates of the regression trees obtained by fixing every possible combination of values of the hyperparameters in Table 3.3, that is, every 3-tuple of the form

$$(d,\, l,\, \alpha) \tag{3.1}$$

where d denotes the maximum depth of the tree, l is the minimum number of samples required to split an internal node and α is the complexity parameter. According to the values in (3.1), we obtain different combination of pruning techniques as explained by Table 3.4.

Model hyperparameter	**Values**
MAXIMUM DEPTH d	10, 11, 12, ..., 30, `None`*
MINIMUM SAMPLES FOR A SPLIT l	2, 5, 10, 15, 20
CCP α	0, $1.13655 \cdot 10^{-5}$, $1.52811 \cdot 10^{-5}$, $1.94197 \cdot 10^{-5}$

* `None` stands for no limitation to the maximum depth of the tree.

Table 3.3: Pruning hyperparameters and candidate values.

Combination of hyperparameters	**Pre-pruning**	**CCP**
$(d,\, l,\, \alpha)$, $d \neq$ `None`, $l \neq 2$, $\alpha > 0$	✓	✓
(`None`, 2, α), $\alpha > 0$	-	✓
$(d,\, l,\, 0)$, $d \neq$ `None`, $l \neq 2$	✓	-
(`None`, 2, 0)	-	-

Table 3.4: Combinations of pruning techniques.

Taking a step back to the tree growing process, so far, regression trees have been built by means of recursive binary splitting, which is based on the identification of the *best* split (the *best predictor* X_j and its *best cutpoint* s) in each division of the predictor space. However, during this stage of the regression analysis, an alternative splitting method has been taken into consideration: the *random splitting approach*. This technique simplifies the minimisation problem (1.2) solved by recursive binary splitting by randomly picking a predictor X_j, and only after that, computing the *best cutpoint* s. Hence, (1.2) reduces to

$$\min_s \left\{ \sum_{i\,:\,x_i \in R_1} (y_i - \hat{y}_1)^2 \;+ \sum_{i\,:\,x_i \in R_2} (y_i - \hat{y}_2)^2 \right\}.$$

In conclusion, we included the possibility of this alternative splitting approach in the investigation of the pruning hyperparameters described above, so that, each 3-tuple (3.1) became a 4-tuple

$$(d,\, l,\, \alpha,\, \sigma)$$

where σ denotes the splitting approach and can take two values: `best` for recursive binary splitting, and `random` for the homonym approach.

The generalisation scores estimates resulted from the above-described inspection, implemented via exhaustive grid search, are graphically displayed in Figures B.4 and B.5. Plots are arranged in both figure according to the same specific structure:

- Each column identifies a different splitting approach of the predictor space during the tree growing phase. On the left, all the blue curves are related to regression trees grown with recursive binary splitting. On the right, the orange curves are related to models built by means of the random splitting approach.

- Each row identifies trees that share the minimum number of samples required to split an internal node. This value increases from the top to the bottom of the figure, which means that moving downwards the regression trees are pruned by preventing splits of nodes of size below the pre-specified threshold.

- Each curve in a single plot connects the scores of regression trees which share the characteristics described at the previous points and that are post-pruned with the same complexity parameter α.

- Finally, each curve represents the evolution of the generalisation scoring metric estimate as a function of the maximum depth of the tree.

Further considerations about these results will be given in the next section.

3.3 Evaluation of relevant regression trees

At this point of the investigation, we have explored many regression trees, grown over the same dataset and differently pruned (or not pruned at all). Here, we draw conclusions about their performance and, we pick the best model for our goal.

In order to do so, we decided to compare the performance of some regression trees, selected as follows: for each combination of pruning techniques (Table 3.4), and for each method of splitting the predictor space (`best` or `random`), the model that minimises the generalisation score estimates (according to only one metric, or both of them) has been selected. We refer to them as *relevant models* and, for not to get confused between one model and another, we assign them a label as described in Table 3.5.

Label	CCP	Pre-pruning	Splitter	Label	CCP	Pre-pruning	Splitter
T_1	✓	✓	`best`	T_5	-	✓	`best`
T_2	✓	✓	`random`	T_6	-	✓	`random`
T_3	✓	-	`best`	T_7	-	-	`best`
T_4	✓	-	`random`	T_8	-	-	`random`

- $T_{i,\text{MSE}}$ means that the relevant model has been chosen because it minimised the generalisation MSE for that specific combination of pruning techniques. Analogously for $T_{i,1\text{–}R^2}$.
- If both metrics agree on a model for a specific combination of pruning techniques, T_i is used.

Table 3.5: Labels of the relevant models.

3.3.1 Assessment of the relevant models' generalisation performance estimates

The pruning hyperparameters and the generalisation performance estimates of the relevant models are described in Table B.1. Immediately it leaps out that almost all the relevant models identified by means of MSE correspond to the relevant models selected through $1-R^2$. The metrics disagree only on T_2, for which they pick two trees that differ for the maximum depth parameter, but whose generalisation scores estimates are not significantly different (this means that $T_i = T_{i,\text{MSE}} = T_{i,1\text{–}R^2}\ \forall\, i \neq 2$).

Secondly, one can observe that the complexity parameter which guarantees better generalisation performances is always $\alpha_{min} = 1.13655 \cdot 10^{-5}$, whether pre-pruning is carried out or not. Hence, in this regression problem, the 1SE rule for the selection of the complexity parameter has not guaranteed improvements.

Speaking of numbers, the top three models supposed to be better in generalising their predictive power, i.e., the regression trees whose generalisation scores estimates are the smallest three, are, in order, T_6, T_1 and T_2. The random splitting approach is used in two out of three of them (T_6 and T_2) so that it seems to be a better choice than recursive binary splitting. On the contrary, the regression trees that are less capable of generalising are that not pruned at all, namely, T_7 and T_8, confirming the theoretical results.

Finally, it must be noticed that the relevant models ranked far from the top three have generalisation scores estimates which do not differ too much from that of the top models, although significantly smaller than that of unpruned trees.

3.3.2 Evaluation of the relevant models on the test set

The right moment for confirming the generalisation performance estimates over the test set has come. The regression trees were evaluated on the unseen data and, the results are shown in Table B.2.

Here, the rankings based on the scores computed on the test set seem not to comply with the rankings derived from their estimates seen so far in Table B.1. The regression trees that keep their rank unchanged are T_1, which maintains 2nd place on the podium, T_7 and T_8, which confirm to be the worst models at generalisation.

Conversely to the top three relevant models according to the generalisation scores estimates, the true test scores see on the podium only models built with recursive binary splitting, in order: T_3, T_1 and T_5. Thus, even though cross-validation seemed to favour models trained with the random splitting approach, actually, recursive binary splitting appears to guarantee the construction of models that are more reliable and capable of better extend their predictive power to unseen data.

Using T_1 as a reference regression tree might be reasonable, indeed, employing the combination of recursive binary splitting, pre-pruning and cost-complexity pruning, it can perform on the test set as well as cross-validation has estimated.

Nevertheless, we must keep in mind what has been observed in Section 2.2.5: we need to make sure that the positively skewed distribution of the cropland fractions is not pushing the model in favour of the prediction of zero values. To this end, we have evaluated the scoring metrics, separately, on the positive values and the null values of the response variable. In other words, if $\mathbf{y}$ is the test response variable, we have split it into two subvectors: $\mathbf{y}_+$, containing all the positive values, and $\mathbf{y}_0$, composed of all the zeros. Then, the scoring metrics have been calculated among $\mathbf{y}_+$ and the corresponding model prediction $\hat{\mathbf{y}}_+$, and among $\mathbf{y}_0$ and $\hat{\mathbf{y}}_0$. The results are contained in Table B.2[4].

These two new metrics stress that all the models can definitely predict zero fractions better than positive fractions. Indeed, comparing the MSE values on the positive fractions with the MSE values on the null fractions, the latter are always way smaller than the former. When the relevant models are ranked according to the new metrics, we observe that T_1 keeps 2nd place if dealing with zero fractions, while it

[4] Only MSE has been computed on the zero fractions since a null vector has no variability of its components that can be detected by R^2.

drops at 4th place when considering only the positive fractions. This suggests that its ability in predicting the null values slightly compensates the errors on positive values. On the contrary, T_3, through only cost-complexity pruning, outdoes T_1 in both cases, which suggests that the model might be as much a good choice as a reference model as T_1, even if cross-validation has not placed it within the top models.

In conclusion, both T_1 and T_3 are the best trees according to the metrics computed on the whole test set and, they still perform well when null values are separated from positive values in the response variable so that strong bias toward zeros might be excluded. However, to avoid bias also toward the test set, we stick to the choice of T_1 as the optimal model since its sufficiently good performances are confirmed by the cross-validation estimates.

Therefore, the desired function $\hat{f}_1$ in (2.1) is that returned by T_1. From now on, we refer to this regression tree simply as T and we denote with $\hat{f}_{1,T}$ the correspondent regression function.

3.3.3 Final considerations on the regression tree T

Summing up, the regression tree T has been built using recursive binary splitting and perfected by combining pre and post-pruning techniques:

- the maximum depth of the tree is $d = 27$;
- a minimum of $l = 15$ observations inside an internal node of the tree is required for a split to be performed;
- cost-complexity pruning has been carried out with parameter $\alpha = 1.13655 \cdot 10^{-5}$.

According to Table B.2, the model has been able to explain the 64% of the variability of the response variable values over the unseen data that compose the test set; this percentage reduces to 56% if we consider only the positive values of the response[5]. Anyway, the employment of pruning techniques allowed to improve the percentage of explained variability by 6% with respect to a full unpruned tree built with recursive binary splitting[6]. These outcomes are not so trivial if we consider

[5] T has a test $1-R^2$ equal to 0.359, which implies a R^2 of 0.641 and a test $1-R^2$ for $y>0$ equal to 0.445343, that is, R^2 of 0.554657.

[6] The full unpruned tree T_7 has a test $1-R^2$ equal to 0.420501, and a test $1-R^2$ for $y>0$ equal to 0.503715.

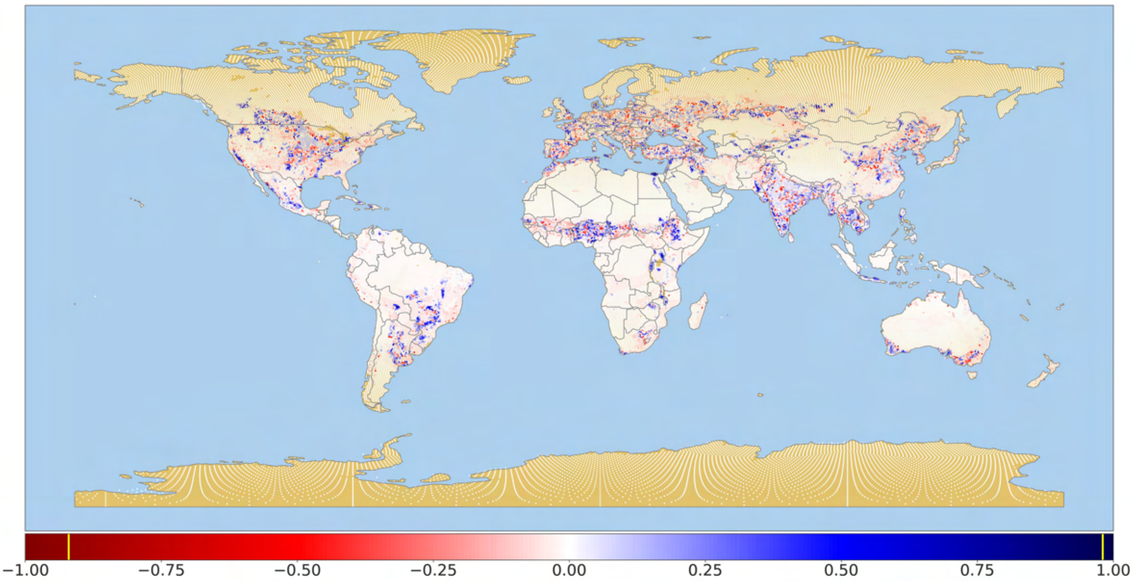


Figure 3.1: Geographic map of the residuals of the regression tree T computed on the whole dataset (training set + test set). Red shades denote under estimation, while blue shades indicate over estimation of the true response values. Every white point represents a correctly predicted response value.

the complexity level of the relationships we want to detect and the simplicity of the model that has been used.

Furthermore, we can have a practical sense of the errors that the tree T commits and check whether there is any geographical pattern concerning them. Indeed, the dataset on which this regression problem is defined hides an additional feature behind the purely numerical values of predictors and responses, that is, the geographical distribution of the data points. Then, it is possible to plot the regression model residuals on top of a geographic map.

The result of this idea can be found in Figure 3.1. Comparing this map with the distribution of the fractions of cropland in Figure A.22, we can confirm that our model is definitely more capable of predicting whether no cropland is present than the correct fraction if cropland cover is there. Indeed, the residuals with large positive absolute value are concentrated in the areas in which there are the largest cropland fractions.

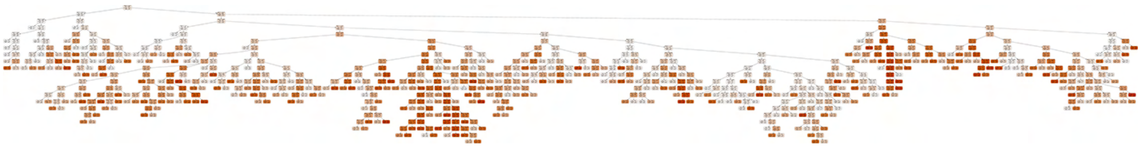


Figure 3.2: 2D graphic representation of T.

In this context, it is worth mentioning that, by repeating multiple times the regression analysis with different random states of the function that splits the dataset in training and test subsets, we have noticed slightly different performances of the relevant models. Thus, this is proof that the instability problem of regression trees is actually present and, despite pruning, it has been influencing the analysis.

Therefore, to obtain a more stable model with better generalisation performance and more reliable informative capabilities, we have decided to combine the power of many regression trees, using the random forest method, as described in Chapter 4. However, before that, further analysis on T will be done in order to extract insights about the relationships between the bioclimatic predictors and the land cover fraction variable IGBP12 that it has been able to learn.

3.4 Interpretation of the results

So far, we have inspected how the regression trees have been built and pruned, we have numerically quantified their predictive performances on unseen data and, we have identified the regression tree T to be the optimal reference model. However, we are still missing the core question of this work, that is, *are there relationships between the climatic conditions and the percentage of lands dedicated to crops in a given geographic area?*

The purpose of this section is answering to this question, or more precisely, extracting human-understandable insights from T regarding the patterns that it has learnt. It could be possible to look directly at the 2D representation of the tree (Figure 3.2), but, given its large size, deriving useful information might be arduous. Then, in order to do that, we have exploited different techniques: impurity-based variable importance measure, permutation importance measure and partial dependence plots.

Impurity-based variable importance measure

In Section 1.2.2 we have seen that the importance measure of a predictor according to a decision tree is evaluated as the decrease of the model impurity (MSE) induced by the splits in which the predictor itself has been involved. Figure B.6 shows the ranking of the importance values of each bioclimatic predictor within the regression tree T. In line with this criterion, the variables that mainly drive the model are:

1. BIO11: *mean temperature of coldest quarter*
2. BIO10: *mean temperature of warmest quarter*
3. BIO12: *annual precipitation*
4. BIO08: *mean temperature of wettest quarter*
5. BIO03: *isothermality*

The decrease of MSE induced by the previous variables is much larger than that caused by the remaining, indeed, a sudden drop arises in the barplot in Figure B.6, just after BIO03. Anyway, it is worth noticing that every bioclimatic predictor has been employed for the partition of the predictor space, with no exclusions.

Permutation importance measure

Conversely to the impurity-based variable importance measure, computed exploiting the characteristic inner structure of a decision tree, *permutation importance measure* (Breiman, 2001a) is a more general inspection technique that can be used for a large variety of learning methods, as long as the dataset has a tabular structure.

In a few words, for a given predictor, this measure is defined to be the decrease in the model prediction score when the predictor values are randomly shuffled among the observations. Rearranging the predictor values breaks the relationship existent in the model between the predictor itself and the response. Hence, the difference between the model prediction scores obtained before and after this operation quantifies the dependence of the response on the predictor.

Lastly, permutation importance can be evaluated either on the training or the test set and, the results assume different connotations: in the first case, we obtain the importance of the predictor in the description of the training data, while in the second, the contribution of the predictor in the generalisation power of the model is estimated. Predictors that reveal to be important on the training set but not on the test set may be related to overfitting.

We have measured the permutation importance of the bioclimatic predictors in the regression tree T. The scoring metric used was the prediction MSE. Besides, in order to have more stable estimates of the importance values, they have been averaged over 10 random shufflings for each predictor. Both the training (Figure B.7a) and the test set (Figure B.7b) have been used, so that we have obtained two comparable rankings. Even though there are some little variations in the position of the middle-ranked predictors, the permutation importance values for model description and model generalisation capability agree on the most important predictors in the model, namely:

1. BIO12: *annual precipitation*
2. BIO10: *mean temperature of warmest quarter*
3. BIO08: *mean temperature of wettest quarter*
4. BIO03: *isothermality*
5. BIO01: *annual mean temperature*

Notice that four out of five predictors are the same elected by the impurity-based variable importance. The relevant difference regards the predictor BIO11 (*mean temperature of the coldest quarter*), which causes the largest MSE decrease in T, yet its permutation importance is only at the 5th place for model description and the 6th place for model generalisation capability. It is worth remarking that the decrease of importance of the bioclimatic predictors in these two rankings does not present a sharp drop after the top five predictors, like in the previous, but it is more regular.

Partial dependence plots

The most effective way to understand the model $\hat{f}_{1,T}$ would be looking at its graphical representation as a function of its arguments. Unfortunately, human perception is limited to 3 dimensions, so that there is no way of looking at the plot of a 19-dimensional function as a whole. A useful alternative is *partial dependence plots* (Hastie, Tibshirani, and Friedman, 2009), that is, a collection of plots, each one showing the partial dependence of the model function $\hat{f}_{1,T}$ on a subset of at most two predictors. Despite these plots cannot provide a comprehensive representation of the model function, they might emphasise relevant interactions between the response variable and some predictors.

In Figure B.8, one can see 19 partial dependence plots, one for each bioclimatic predictor. Every plot contains the estimated marginal average of $\hat{f}_{1,T}$ on a given predictor (blue line) and the corresponding confidence interval (light blue area). Clearly, the estimation of the partial dependence of $\hat{f}_{1,T}$ on a predictor making use of all its values would have been computationally too expensive, hence 100 equally-spaced values for each predictor have been utilised instead.

In a few words, in every partial dependence plot, we can perceive the effect of each bioclimatic predictor on the model response values after accounting for the (average) effect of the other predictors on $\hat{f}_{1,T}$. Although most of the variables have an approximately constant relationship with the response variable, few of them show more peculiar non-linear relationships, mainly characterised by drops and jumps:

BIO01 The average predicted fraction of cropland suddenly rises immediately after the annual mean temperature crosses 0 degrees and slightly keeps increasing thereafter.

BIO03 The relationship between the isothermality values and the predicted cropland fractions is approximately inverse linear, with a low slope.

BIO10 The association with the average predicted fraction of croplands is constant as long as the mean temperature of the warmest quarter is lower than 15 degrees, then, the fraction values start rising and they present a relevant jump around 29 degrees, after which they level off.

BIO11 The constant association between the response and the mean temperature of the coldest quarter is interrupted by a sudden drop, around 5 degrees; then the cropland fractions slightly increase again.

BIO12 After the annual mean precipitation reaches about 300 millilitres, the cropland fraction values suddenly rise, and after 1200 millilitres, slightly decrease again.

Notice that these are the same variables top-ranked by means of the previous importance measures, except for BIO08, whose partial dependence plot is hardly interpretable.

In conclusion, we have seen that when dealing with cropland fractions, regression trees have revealed to be sufficiently good methods to learn the patterns that connect the distribution of land cover fractions with the climatic conditions supplied by the bioclimatic predictors. Furthermore, they proved to be reasonably informative about these patterns, indeed, thanks to the regression tree T, we ended up with useful knowledge.

3.5 About IGBP13 and IGBP14

When the same stages of the regression analysis applied to IGBP12 have been repeated onto the response variables IGBP13 and IGBP14, we have obtained completely different results, in the sense that regression trees have not been able to learn any pattern connecting the distribution of land cover fractions for this two variables, to the climatic conditions summarised by the bioclimatic predictors.

No matter which pruning technique has been applied, the best-performing regression trees have been able to explain no more than, respectively, 11% (test $1-R^2$: 0.89) and 34% (test $1-R^2$: 0.66) of the variability in the response variables of unseen data, which classifies them as slightly better than the models that uniformly predict the mean response value.

The apparent impossibility in finding well-performing regression trees might be interpreted in two different ways:

a) Regression trees are not able to identify the desired relationships, but different machine learning methods might manage to do so.

b) There is no relationship between the distribution of the two responses values and the bioclimatic predictors.

In our opinion, at each response variable is connected a different motivation. If we consider the response IGBP13, that is, the fractions of land cover occupied by urban areas, the more likely motivation is the second one. During the past centuries, the human being has been able to develop and perfect much technology, with the main purpose of facilitating his life in all aspects: to supply his dietary and energetic needs, to defeat health problems, to improve his housing situation, and to get protection from hostile climatic conditions. Therefore, if it is reasonable to think that the presence of a human settlement was mainly driven by the most comfortable climatic conditions in the long-gone past, once the ability to shape the surrounding space to his liking had been developed, it is more likely that the choice of the location of a residential area during the past years has been motivated by alternative reasons (historical or social events, presence of raw materials, etc.). For these reasons, and probably many others that we have no knowledge of, we think that it is not possible to predict the percentage of urban areas basing on climatic conditions only.

On the other hand, IGBP14 describes the percentage of land cover constituted by small-scale cultivation mixed with natural vegetation, therefore, it is conceptually closer to the object of IGBP12 than that of IGBP13. So, *why there should be no association between the climatic conditions and the distribution of this kind of land cover fractions?* Now, the most appropriate motivation might be the first listed above: regression trees are not the perfect model to learn these eventual patterns. There could be various reasons, and each one of them might be treated in many ways, but one that has been stressed by this work, is that the strongly positive skewed distribution of the response values should definitely be taken into consideration, not only when evaluating the model, but also during its training. For instance, zero-inflated models might be employed in order to learn the fact that numerous response values are often zero.

Chapter 4

Regression analysis with random forests

The regression analysis performed with decision trees has been quite successful when dealing with cropland fractions. Though, as the theory of Chapter 1 suggests, the aggregation of the predicted power of many trees might enhance the performance and increase the stability of a single tree. Moreover, it could guarantee a more reliable interpretation of the patterns that connect the bioclimatic predictors to the response variable IGBP12.

For these reasons, throughout this last chapter, we describe how the regression analysis has been carried out using random forests. We will focus only on the regression problem involving the response variable IGBP12, indeed, as well as for regression trees, the random forests have failed with the other two responses.

4.1 Implementation

Once again, the code produced for the results presented in this chapter is in `Python 3` and many libraries have contributed with useful tools for the regression analysis (e.g., `pandas`, `geopandas`, `NumPy`, `seaborn`, `matplotlib`, etc.), in particular `scikit-learn`, which provides the `RandomForestRegressor` function, capable of growing a random forest.

Let us look at a few technical matters, before moving to the true regression analysis.

Scoring metrics and estimation strategies of generalisation performances

In Chapter 3, the employed strategy to estimate the generalisation performance of regression trees has been the typical 10-fold cross-validation, which remains a valid approach for random forests as well. Nonetheless, due to the increased complexity of random forests compared to decision trees, the running time of the cross-validation approach results to be dramatically longer.

As a consequence, we have opted to replace cross-validation with the out-of-bag error estimation, as described in Section 1.2.2. This strategy is very spread and recommended when dealing with random forests, and more generally when bagging is used (Probst, Wright, and Boulesteix, 2019; Probst and Boulesteix, 2017; Gislason, Benediktsson, and Sveinsson, 2006; Cutler et al., 2007), because it reduces the running time by k times compared to k-fold cross-validation, and, in the meanwhile, it guarantees an honest estimate of the generalisation performance of the model (Breiman, 1996b).

In order to compute the OOB score of a random forest, one should have full control over the growing process of each tree of the forest, and especially, on the exact samples used to build each tree. The `scikit-learn`'s function `RandomForestRegressor` does not allow the needed management of the bootstrapped samples, yet the function is capable of returning the out-of-bag R^2 value, therefore, it has been possible to use it for our purposes.

It follows that, as a result of the above-mentioned technical limitation, we have been forced to abandon the MSE as a scoring metric used for estimating the generalisation performance of random forests, in favour of the only $1-R^2$. It remains unaltered the role of MSE and $1-R^2$ as scoring metrics when directly evaluating the model performance on the test set.

Division of the dataset into training and test sets

For this regression analysis we kept the same training and test sets randomly created for the regression analysis with decision trees in order to produce comparable results.

4.2 Forest growing and tuning

As regards regression trees, we have seen that pruning strategies are required to enhance their generalisation power. On the contrary, random forests, as presented in Section 1.2, draw their strength from the averaging of many unbiased trees and from the randomness that characterises their growing. Therefore, in order to obtain a well-performing unbiased random forest, no pruning of the individual trees is required.

Consequently, it may appear that a random forest could be trained without the need of much thinking about its construction, leaving all the merit to the final averaging operation and to the randomness of bootstrapped training sets and split candidates. On one hand, this may be true, indeed, any random forest implementation works reasonably fine even without too much interaction with the user. On the other hand, this is possible because many of the hyperparameters that control the structure of each tree, the structure and the size of the whole forest, and its randomness are set by default to values that commonly work fine.

Though, not all the hyperparameters of a random forest are equally significant: some of them should be tuned, others must be selected according to external factors and, still, others might be ignored. During this section, we are going to see the way they have been treated in our work.

4.2.1 Tuning model hyperparameters

In accordance with Probst, Wright, and Boulesteix (2019) and the references therein, the impact of several hyperparameters on the random forest performance had been studied in the course of the literature and it revealed that the effect of tuning[1] is generally less evident with random forests than with other machine learning algorithms. Despite that, some hyperparameters proved to have a larger impact on the model performance than others, and that the research for their optimal values might result advantageous.

[1] The term *tuning* stands for the procedure of finding the optimal values of one or several hyperparameters for a machine learning algorithm on a specific dataset. Optimality may have a different meaning, in line with the considered problem, but typically refers to model performance onto unseen data.

In light of these findings, over the current regression analysis, we tuned the hyperparameters which have evidenced larger performance gain as per Probst, Wright, and Boulesteix (2019), namely, in order of relevance:

1. The number of split candidates m, that is, the quantity of randomly sampled predictors out of which the best split of the predictor space is determined when growing each regression tree.
2. The size s of the bootstrapped training sets on which the trees are fitted.
3. The minimum number of samples l required to split an internal node of the tree.

The former two hyperparameters, m and s, are related to the randomness of the forest and their tuning makes it possible to find a good compromise between the stability and the strength of the model. In different terms, it is the negotiation between the decrease of correlation among the trees in the forest and the guarantee that each one of them is sufficiently accurate in prediction. As per Breiman (2001a), this is precisely the key point of the random forest model.

For the note, the variety of trees in the forest grows (then, the correlation decreases), the lower is the number of split candidates m and the smaller is the bootstrap sample size s. This happens because when few split candidates are randomly selected, predictors with moderate effect on the response variable are more likely to be chosen than otherwise. Hence, the significant predictors are less used for splitting and the variability in the structure of trees increases. Similarly, reducing the bootstrap sample size allows fitting trees on more diversified training sets, rising variability among the estimators. As a consequence, the random forest benefits more stability.

The counterpart of reducing their values is that the model accuracy could dramatically lessen. Indeed, either when the training size of a regression tree diminishes or when the tree is trained with sub-optimal predictors, its performance obviously worsens. Thus, it is evident that a trade-off between the two side-effects is required and that the optimal values of these two hyperparameters are problem-dependent (Goldstein, Polley, and Briggs, 2011; Genuer, Poggi, and Tuleau, 2008; Martínez-Muñoz and Suárez, 2010).

Meanwhile, the minimum number of samples l directly affects the structure of the trees, limiting their depth. Even though its impact on the performance gain is

Model hyperparameter	**Values**
NUMBER OF SPLIT CANDIDATES m	1, 2, ..., 19
BOOTSTRAPPED TRAINING SET SIZE s^*	10%, 20%, ..., 100%
MINIMUM SAMPLES FOR A SPLIT l	2, 5, 10, 15, 20

* Expressed as a percentage of the whole training set size n. For instance, if $s = 30\%$, then, the bootstrap sample size is $n_s = 0.3n$.

Table 4.1: Tuning hyperparameters and candidate values.

usually less relevant than the other two, it could potentially be improved by tuning (Lin and Jeon, 2006) and it has shown relevance when noisy predictors are present. (Segal, 2004).

In our specific regression problem, tuning has been performed via exhaustive grid search to minimise the out-of-bag $1-R^2$, i.e., to optimise the generalisation performance of the random forest. The grid of values among which the optimal combination has been researched is that obtained by considering every 3-tuple of the form

$$(m,\ s,\ l)$$

where the candidate values of m, s, and l are in Table 4.1. It is necessary to specify that the tuning has been conducted for a random forest composed of 200 regression trees, but we will return on this aspect later on.

The influence of these hyperparameters on the estimated generalisation performance of the random forest is summarised by the three plots in Figure B.9. Each plot contains four curves obtained by fixing the values of two out of three hyperparameters and plotting the evolution of the out-of-bag 1–R^2 as a function of the third.

Inspecting each plot individually, two patterns are immediately perceivable: the estimated generalisation score decreases as the bootstrap sample size s increases (Figure B.9b) and it is almost directly proportional to the minimum number of samples l (Figure B.9c). These relationships appear to be independent of the number of split candidates m.

In Figure B.9a instead, we can see convex curves whose minimum point is altered by changing the values of s and l. When l is large and s is small, the optimal value of m is approximately around 10, yet, its optimality is questionable given the flat trend of the convexity. It moves toward lower values as the sample size s increases and the number of samples l decreases. It is worth mentioning that for $m=19$ we are effectively applying bagging instead of the random forest model since all the predictors are always examined when a split has to be performed. Besides, passing from bagging to random forests definitely benefits the model performances.

The inspection so concluded has elected the following random forest configuration as the optimal one in terms of generalisation power:

1. $m=6$ candidate predictors randomly drawn at each split of the predictor space;

2. training sets composed of $s=n=55\,860$ observations sampled with replacement from the true training set;

3. a minimum of $l=2$ observations inside an internal node of a tree for a split to be performed.

As concerns s and l, the default settings in `scikit-learn` have been proved to be optimal toward this regression problem, indeed, the bootstrapped training sets maintain the size of the true training set and nodes are split until no further division is possible (i.e., until each node is composed of a single observation). As for the number of split candidates, the standard behaviour of `scikit-learn` is the employment of bagging, therefore, finding the optimal value for this hyperparameter has been the most important result of the inspection. Besides, notice that $m=6$ corresponds to the commonly suggested choice $\lfloor p/3 \rfloor$, where p is the number of predictors.

4.2.2 Picking the number of trees in the forest

In order to conclude the growing and tuning phase, we need to take into account a fourth hyperparameter that affects the size and the complexity of the random forest model, namely, the number of trees which compose the ensemble, denoted by t. We have been treating this hyperparameter separately from the previous because it must not be considered as a tuning parameter, but, it should be set sufficiently high as long as it is computationally manageable (Probst and Boulesteix, 2017). Let us argument this statement and see how it had influenced our regression analysis.

Breiman (2001a) proved that the generalisation MSE of a random forest is convergent as the number of trees t grows and, he provided an upper bound to such error.

Later on, Probst and Boulesteix (2017) theoretically and empirically proved that the generalisation MSE[2], either estimated via out-of-bag samples or evaluated on a test set, is a monotonously decreasing function of t. Furthermore, their empirical analysis showed that the largest error reduction is achieved when the first 100 trees are trained and, thereafter, the performance gain obtained by adding more trees is minimal. In other terms, the generalisation error curve exhibits a steep decrease followed by a plateau when narrowing to the limit error[3].

An additional fact that the authors remarked is the influence that the specific dataset and other model hyperparameters might have on the convergence rate. One can clearly do nothing about the data, though, as concerns the hyperparameters, an observation can be done: operating on the number of split candidates, the bootstrap sample size or other parameters affecting the structure of the trees allows to have more diversified but less powerful estimators, then, their number must be higher to guarantee precise predictions and let the ensemble reaching the error convergence.

The facts just described have had two consequences on our regression analysis. First, the tuning of hyperparameters in Section 4.2.1 has been performed onto a random forest composed of 200 trees. Even if in presence of altered values of the hyperparameters, this should be a sufficiently large size that:

- guarantees a reliable estimate of the generalisation performance,
- avoids that the simple addition of a few more trees to the ensemble might drastically change the optimal combination of values for the hyperparameters chosen.

Second, we had to make sure that the monotonously decreasing trend of the generalisation error proved for MSE was true even in our problem, where the generalisation performance is computed employing $1-R^2$.

[2] More generally, this result holds whenever the generalisation performance of the random forest is based on average loss, like the MSE, which is computed as the average of the squared errors.

[3] This observation is supported in the classification setting also by Oshiro, Perez, and Baranauskas (2012). Nonetheless, in the course of the empirical study of Probst and Boulesteix (2017), a non-monotonous behaviour of the generalisation performance of random forests in binary classification has been observed, even though only in specific circumstances and for certain error measures.

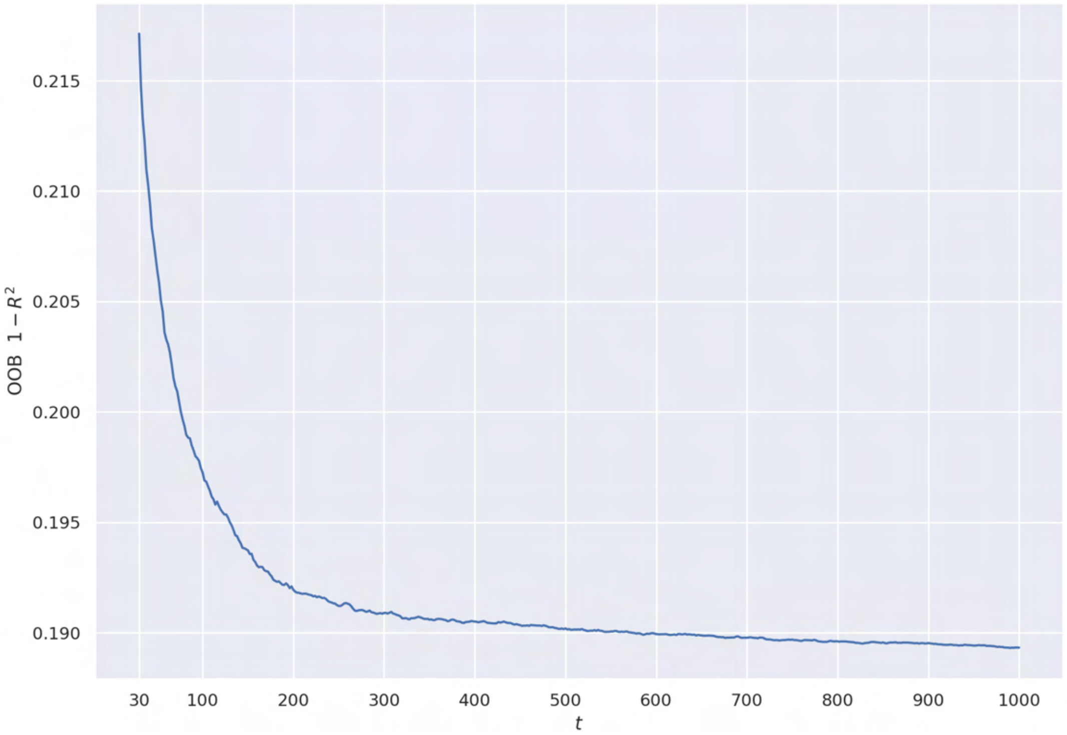


Figure 4.1: Evolution of the out-of-bag $1-R^2$ as a function of the number of trees t in the random forest with fixed hyperparameters $m=6$, $s=100\%$ and $l=2$.

To this purpose, in Figure 4.1 we can see the curve describing the model generalisation performance, estimated through out-of-bag $1-R^2$, as a function of t in $[30, 1000]$[4]. The model subject of the inspection is the random forest trained with the previously tuned hyperparameters: $m=6$, $s=100\%$ and $l=2$. The curve exhibits the same trend observed for the generalisation MSE by Probst and Boulesteix (2017), with a lower convergence rate: a large performance gain accomplished within the first 200 trees, followed by a slight descent.

Speaking of numbers, in Table 4.2 we can find a numerical summary of how the out-of-bag score improves as t grows from 30 to 1000. The performance gain for 1000 trees compared to 30 trees is 0.0278. The biggest performance gain is evidently achieved after only 100 trees, i.e., 0.019874 (71.49% of the overall gain), though, the addition of further 200 trees leads to a total gain of 0.026229 (94.35% of the overall gain). Thereafter, the gain improvement dramatically slows down.

[4] With a little abuse of notation, the interval $[t_0, t_1]$ represents the set of integers $\{t_0, t_0{+}2, \ldots, t_1\}$

$[\mathbf{t_0}, \mathbf{t_1}]$	$\hat{\mathcal{R}}(\mathbf{t_0})$	$\hat{\mathcal{R}}(\mathbf{t_1})$	**Gain** $\hat{\mathcal{R}}(\mathbf{t_0}) - \hat{\mathcal{R}}(\mathbf{t_1})$	**Gain % wrt** $[\mathbf{30}, \mathbf{1000}]$
[30, 1000]	0.217128	0.189328	0.027800	100.00%
[30, 100]	0.217128	0.197253	0.019874	71.49%
[100, 200]	0.197253	0.191948	0.005306	19.09%
[200, 300]	0.191948	0.190898	0.001049	3.77%
[300, 400]	0.190898	0.190507	0.000391	1.41%
[400, 500]	0.190507	0.190197	0.000310	1.12%
[500, 600]	0.190197	0.189942	0.000256	0.92%
[600, 700]	0.189942	0.189776	0.000166	0.60%
[700, 800]	0.189776	0.189613	0.000163	0.59%
[800, 900]	0.189613	0.189536	0.000077	0.28%
[900, 1000]	0.189536	0.189328	0.000208	0.75%

$\hat{\mathcal{R}}(t)$ here stands for the OOB $1-R^2$ of the random forest composed of t trees.

Table 4.2: Numerical summary of the generalisation performance gain estimates of the random forest ($m=6$, $s=100\%$, $l=2$) as the number of trees t grows from 30 to 1000.

Assuming that the generalisation score of the random forest composed of 1000 trees is close to the limit score, we have found that a random forest with only 300 trees should be able to have comparable performances with less computational effort.

In view of the results obtained so far, we have decided to solve the regression problem (2.1) using a random forest composed of 300 trees, with the significant hyperparameters previously tuned ($m=6$, $s=100\%$, $l=2$). From now on, we refer to this model with the label F and we denote $\hat{f}_{1,F}$ the associated regression function.

		Regression tree $\boldsymbol{T_1}$	**Random Forest** $\boldsymbol{F}$
Hyperparameters		$d{=}27$, $l{=}15$, $\alpha{=}1.13655{\cdot}10^{-5}$	$t{=}300$, $m{=}6$, $s{=}100\%$, $l{=}2$
Test scores estimates	**CV*MSE**	0.015288	-
	CV*$\boldsymbol{1-R^2}$	0.344193	-
	OOB $\boldsymbol{1-R^2}$	-	0.190898
Test scores	**MSE**	0.014919	0.008458
	$\boldsymbol{1-R^2}$	0.359000	0.203534
	MSE $(y>0)$	0.038134	0.021776
	$\boldsymbol{1-R^2}$ $(y>0)$	0.445343	0.254307
	MSE $(y=0)$	0.000966	0.000454

* Estimates obtained via 10-fold cross-validation.

Table 4.3: Comparison of the generalisation performance of the random forest F and the regression tree T.

4.3 Evaluation of the random forest

There was nothing left to do but evaluate the performance of the random forest F over our test set. In Table 4.3 we can see the generalisation scores of F compared with that of the regression tree T obtained in Chapter 3.

Evidently, there is a large advancement in the descriptive and predictive capabilities when using a random forest in place of a regression tree. According to each metric evaluated on the test set, F outdoes T, approximately halving each score. In terms of R^2, the random forest F can explain 80% of the variability in the test response and, when we isolate the positive fractions, the percentage drops to 75%. Thus, what is surprising is that not only F enhance the overall R^2 of T by 16%, but it improves, even more, the ability in predicting positive fractions, that is by 19%.

Nevertheless, despite this improvement toward positive fractions, the F, as well as T, better predicts null fractions. Indeed, the MSE on the positive fractions is still higher than that on the null fractions. The same conclusion can be drawn by the map of the model residuals in Figure 4.2.

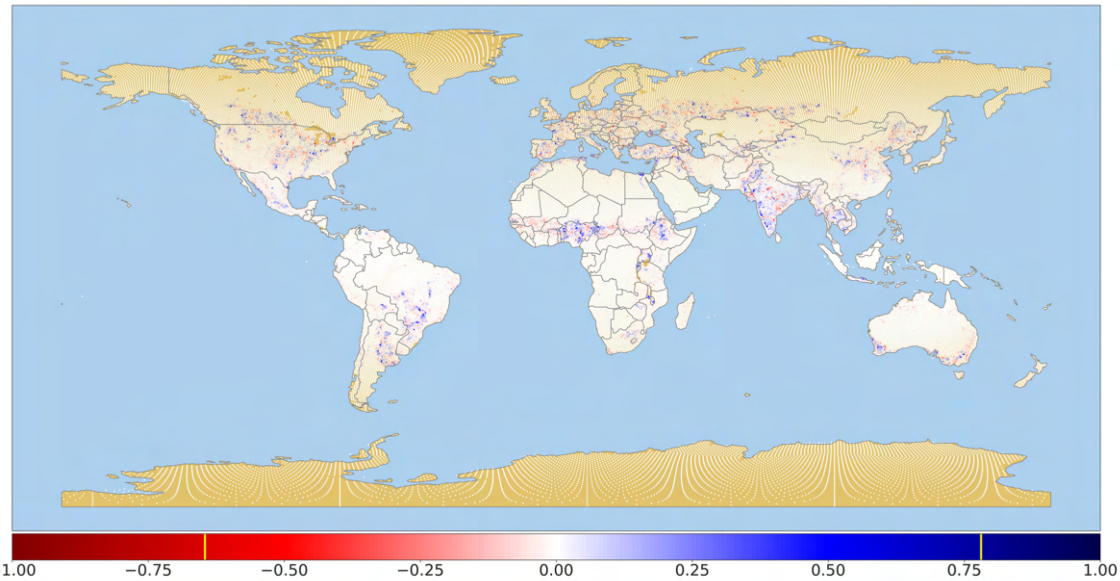


Figure 4.2: Geographic map of the residuals of the random forest F computed on the whole dataset (training set + test set). Red shades denote under estimation, while blue shades indicate over estimation of the true response values. Every white point represents a correctly predicted response value.

4.4 Interpretation of the results

Once again, in this section, we try to answer the question: *are there relationships between the climatic conditions and the percentage of lands dedicated to crops in a given geographic area?*

The regression tree T provided some responses when importance measures of predictors and partial dependence plots have been computed (Section 3.4). Therefore, we have repeated the same techniques on our random forest F to obtain comparable, and possibly more reliable, answers.

Impurity-based variable importance measure

Figure B.10 shows the ranking of the bioclimatic predictors according to the impurity-based importance computed on the random forest F as explained in Section 1.2.2. We can notice that the leading predictors are the same observed in the regression tree T by means of the permutation importance measure, even though in a different order:

1. BIO10: *mean temperature of warmest quarter*
2. BIO12: *annual precipitation*
3. BIO03: *isothermality*
4. BIO08: *mean temperature of wettest quarter*
5. BIO01: *annual mean temperature*

Differently from the ranking of T based on the same measure (Figure B.6), we can see that the importance is much more distributed among the 19 predictors. This fact is due to the diversification of the regression trees within F, mainly caused, in turn, by the reduction of the number of split candidates from 19 (as in T) to 6 (as in F).

Moreover, only the first three variables (BIO10, BIO12 and BIO03) stand out from the others, since there is a drop in the importance values just after BIO03, followed by a roughly linear decrease. Therefore, it seems inappropriate to talk about the top-5 predictors when we should talk about the top-3 instead. However, we maintain this structure for cohesion with the results in Chapter 3.

Finally, we must notice that the predictor BIO11, which was on the 1st place for T, has lost many positions in this new ranking, dropping to the 13th. A reasonable motivation is that the large correlation between it and other significant predictors, such as BIO01, BIO10, BIO08, combined with the small number of split candidates, diminished its effect on the MSE reduction.

Permutation importance measure

Following the same procedure described for regression trees, we have computed the permutation importance of the bioclimatic predictors for data description and generalisation capability of the random forest F. The rankings so obtained are in Figure B.11. The leading predictors for data description (Figure B.11a) are:

1. BIO12: *annual precipitation*
2. BIO10: *mean temperature of warmest quarter*
3. BIO03: *isothermality*
4. BIO01: *annual mean temperature*
5. BIO08: *mean temperature of wettest quarter*

When analysing the generalisation capability of the random forest (Figure B.11b), the set of the most important predictors remains unchanged, but BIO01 and BIO03 are inverted. Other small differences between the two rankings can be observed for middle/low-ranked variables, but none of them is worth mentioning.

Once again, we are seeing these five bioclimatic predictors on the podium, confirming the outcomes obtained with the impurity-based measure on F. Furthermore, a larger distribution of importance between predictors can be observed also with this measure, but less evidently than in the impurity-based importance ranking.

Partial dependence plot

The influence that each bioclimatic predictor has on the regression function $\hat{f}_{1,F}$ can be perceived thanks to the 19 partial dependence plots in Figure B.12. For every plot, the marginal average of $\hat{f}_{1,F}$ (blue line) has been estimated onto 100 equally-spaced values of the predictor, just like it has been done for the regression tree T.

Comparing these partial dependence plots with those of T (Figure B.8), at first sight, the smoothness that characterises the curves in those of the random forest leaps out. Inspecting more carefully the plots, one can see that most of the partial dependences that were roughly constant for T appear to be so also for F (BIO06, BIO14, BIO17, BIO18), though, others revealed small features, such as, weak depressions (BIO02 and BIO09), increases or decreases relegated toward extreme values (BIO05, BIO08, BIO15, BIO16 and BIO19) and, still, linear dependence with tiny slopes (BIO07 and BIO13). On the contrary, the harsh drop that was present in the partial dependence plot of BIO11 is now almost totally levelled off.

Anyway, the most peculiar relationships are exhibited again by BIO01, BIO03, BIO10, BIO12. Jumps and drops that characterised their partial dependence curves according to T are now replaced by smooth traits, but the substance of the plots has not change:

BIO01 The average predicted fraction of cropland starts rising when the annual mean temperature crosses 0 degrees, yet, even though the trend of the curve is increasing, it is not monotonous: there is a weak depression around 20-25 degrees.

BIO03 The relationship between isothermality values and the model predictions of cropland fraction can still be approximated by an inverse linear function with a weak slope.

BIO10 The association with the average predicted fraction of croplands is constant as long as the mean temperature of the warmest quarter is lower than 15 degrees. Then, a monotonously increasing trend begins, characterised by two consecutive smooth jumps, around 15 and 29 degrees.

BIO12 The partial dependence curve appears more flatten than that of T. Still, a smooth jump around the value of 300 millilitres breaks the constant relationship between the average predicted fractions and the annual precipitation values. Hereafter, a constant relationship reestablishes and, crossed the threshold of 1200 millilitres, the curve starts declining again.

To these significant partial dependence plots, we must add that of BIO04, which exhibits an approximately direct linear relationship with the predicted cropland fractions. A relationship that was totally absent in the partial dependence plot of T, but that the random forest F has made emerge. As a matter of fact, the increased relevance of the predictor BIO04 can also be deduced by the three previous importance rankings, in which it always figures at the 6th place.

Conclusion

Throughout the study reported in this thesis we globally analysed the relationships between climatic conditions and three human-modified land cover classes: (large-scale) croplands (IGBP12), mosaics of small-scale croplands and natural vegetation (IGBP14) and urban areas (IGBP13). We aimed at the construction of three suitable regression models for a dual goal:

1. Correctly predict the percentage of the given human-modified land cover according to specific climatic features.
2. Find information on the true connections that exist between climate and the human activities considered.

With the results of the exploratory analysis of the dataset, a complex regression setting emerged. It was characterised by multiple predictors with many outliers, response variables with strongly positive skewed distribution and, the apparent absence of significant correlations between climatic features and land cover fraction variables. In this context, tree-based methods have been able to return a regression model that, by far, outperforms the naive regressor[5] only for one of the three land cover classes.

The distribution of urban areas completely escaped from the modelisation based on climatic data. It has already been conjectured in the introduction and further commented in Section 3.5, but let us remark the following idea: it is our opinion that, if any relationship between climate and the development of urban areas did exist, then, in the modern era, it has been masked by other factors, such as the ability of humans to adapt to the environment, the progress in terms of knowledge and technology, socio-cultural transformations and historical events.

[5] With *naive regressor* we mean the regression model that always predicts the mean response, independently of the bioclimatic predictors.

Concerning the two land cover classes linked to agriculture, IGBP12 and IGBP14, in principle, one would expect them to have similar climatic requirements to guarantee optimal harvests, independently of the extent of lands destined for cultivation. Still, completely contrasting outcomes have emerged during the regression analysis. Out of the two, only the distribution of croplands has shown good responsiveness to the learning techniques. The same methods could not identify effective patterns for the distribution of small-scale croplands.

In this case, the reason behind the failure of tree-based methods is likely to be the globally reduced presence of land cover class IGBP14. Indeed, as it has been observed in Section 2.2.5, in approximately 80% of the lands, there is no sign of mosaics of small-scale cultivation and natural vegetation. In other words, 80% of the response values are zero and, there is not sufficient variability of positive fractions to allow these methods to learn any real pattern.

Anyway, when taking into consideration only the distribution of (large-scale) croplands, tree-based methods turned out to be a valid learning option. A single regression tree, after optimisation of pruning techniques, showed sensible predictive accuracy and good potential for patterns description. However, despite pruning, so simple models exhibited little robustness: changes in the training data distribution due to different random sampling in the splitting procedure of the dataset caused the obtaining of distinct optimal regression trees. As a consequence, the resulting model interpretation was not very reliable.

The employment of random forests let us obtain more accurate predictions and, it solved the instability problem of single regression trees. Hyperparameter tuning and a careful selection of the number of trees inside the ensemble allowed us to identify a reasonably informative model. Yet, it is still worth mentioning that, as long as it is computationally feasible, increasing the number of trees in the forest allows to obtain further modest performance improvement.

Our random forest model can predict the absence of croplands with very low error. On the contrary, it has more difficulty in estimating the true fraction of land cover. The geographic map of residuals and the evaluation of scoring metrics on positive and null subsets of the response variable have been fundamental tools for the identification of this behaviour.

As regards the interpretation of the random forest model, predictor importance measures and partial dependence plots have been able to capture characteristic patterns between climate and croplands distribution. Even though we cannot say whether they are right or wrong, they are in line with expectations based on a basic understanding of agriculture.

Looking at the results as a whole, what emerged is that the distribution of cropland cover is mostly driven by the annual trends of precipitation (BIO12) and temperature (BIO01), by extreme environmental factors connected with the life cycle of plants, such as the mean temperature of the warmest (BIO10) and wettest (BIO08) quarters and, by temperature-related variability measures, like isothermality (BIO03) and temperature seasonality (BIO04). In particular, partial dependence plots emphasised some aspects:

- A sufficient quantity of annual mean precipitation is required for the presence of croplands (between 400 and 1000 ml), but too copious precipitation (over 1000 ml) does not favour cultivated lands. Clearly, copious precipitation is better than too scarce (less than 300 ml).
- A high average annual temperature is very important for agriculture, but it seems that having a temperature larger than 29 degrees during the warmest quarter of the year is the factor that, by itself, has more effect on the presence of croplands.

In the end, we want to point out a couple of observations of a more general nature that emerged during the study. Firstly, decision trees are an excellent way to approach a regression problem when the relationships between predictors and response are confusing and hard to characterise. Indeed, they can often catch important aspects of such relationships, even though they are simple models, easy to understand and interpret (conceptually and practically).

Nonetheless, when the underlying data distribution is particularly elaborate, as in ecological settings (many outliers, skewed distributions, etc.), regression trees have to become very large and intricate to accurately describe the data. Despite that, they cannot learn every significant aspect of the complex distribution and, even small variations in the data, affect the entire structure of the tree. Furthermore, even if pruning is applied, their size remains excessive, so that they lose in interpretability because their 2D representation is hardly comprehensible.

In the so described case, an ensemble of trees, like a random forest, is definitely preferable, because it joins together many models, each one of them can describe a different aspect of the distribution. And, as already mentioned, the accuracy and stability of the regression model drastically improve.

Secondly, in addition to the already emphasised advantages, a random forest owns an effective strategy of estimating the generalisation performance: the out-of-bag error. This is a huge benefit when hyperparameter tuning must be performed via an exhaustive grid search. Indeed, this strategy has a reduced running time of the error estimation per hyperparameter combination. Therefore, the inspection for the optimal combination may go through a larger grid, which would be computationally prohibitive with other estimation techniques, such as cross-validation. Then, this compensates for the weak response of a random forest to tuning.

Future research

Some aspects of this research might be better treated and further explored in future work. A first example might be finding a different strategy to model the distribution of the land cover class IGBP14, mosaics of small-scale croplands and natural vegetation. One would require a model that is capable of considering that the probability of having a zero fraction is way higher than that of a positive fraction, such as a zero-inflated model.

Alternatively, some updating of the set of predictors may be profitable in modelling this land cover class and possibly, also the other two. For instance, one could consider the additional climatic variables proposed by Title and Bemmels (2018) or the environmental and topographic predictors from Amatulli et al. (2018). Indeed, as for the latter, it is reasonable to think that mosaics of small-scale croplands and natural vegetation might be more spread over territories with alternation of plains and high grounds, with the presence of gullies or steep hillsides. On the other hand, for large-scale croplands to be present, a flatter territory should be preferable. Therefore, including variables describing characteristics such as elevation, slope or terrain roughness might be helpful.

Finally, in Section 2.2.4 it has been highlighted that multicollinearity can have a side effect on the importance that each predictor has in the model, attributing more value to some of them just because they are correlated with other significant predictors. Then, it could be a good idea to employ a different variable importance measure that could be less affected by this problem. For instance, (Strobl et al., 2008) suggests an available choice for random forests.

Bibliography

Amatulli, G. et al. (2018). “A suite of global, cross-scale topographic variables for environmental and biodiversity modeling”. *Scientific data* 5.1, pp. 1–15. DOI: 10.1038/sdata.2018.40.

Breiman, L. et al. (1984). *Classification and Regression Trees.* The Wadsworth and Brooks-Cole statistics-probability series. Taylor & Francis. ISBN: 9780412048418.

Breiman, L. (1996a). “Bagging predictors”. *Machine Learning* 24, pp. 123–140. DOI: 10.1007/BF00058655.

Breiman, L. (1996b). *Out-Of-Bag etimation.* Tech. rep. UC Berkeley, Department of Statistics.

Breiman, L. (2001a). “Random Forests”. *Machine Learning* 45, pp. 5–32. DOI: 10.1023/A:1010933404324.

Breiman, L. (2001b). “Statistical Modeling: The Two Cultures”. *Statistical Science* 16.3, pp. 199–231. DOI: 10.1214/ss/1009213726.

Caruana, R. and Niculescu-Mizil, A. (2006). “An Empirical Comparison of Supervised Learning Algorithms”. In: *Proceedings of the 23rd International Conference on Machine Learning.* ICML ’06. New York, NY, USA: Association for Computing Machinery, pp. 161–168. DOI: 10.1145/1143844.1143865.

Cutler, D. R. et al. (2007). “Random Forests for Classification in Ecology”. *Ecology* 88.11, pp. 2783–2792. DOI: 10.1890/07-0539.1.

De’ath, G. and Fabricius, K. E. (2000). “Classification and regression trees: a powerful yet simple technique for ecological data analysis”. *Ecology* 81.11, pp. 3178–3192. DOI: 10.1890/0012-9658(2000)081[3178:CARTAP]2.0.CO;2.

Fick, S. E. and Hijmans, R. J. (2017). “WorldClim 2: new 1km spatial resolution climate surfaces for global land areas”. *International Journal of Climatology* 37.12, pp. 4302–4315. DOI: 10.1002/joc.5086.

Friedl, M. A. and Sulla-Menashe, D. (2019). *MCD12Q1 MODIS/Terra+Aqua Land Cover Type Yearly L3 Global 500m SIN Grid V006*. Data set. NASA EOSDIS Land Processes DAAC. DOI: 10.5067/MODIS/MCD12Q1.006.

García-Gutiérrez, J. et al. (2015). "A comparison of machine learning regression techniques for LiDAR-derived estimation of forest variables". *Neurocomputing* 167, pp. 24–31. DOI: 10.1016/j.neucom.2014.09.091.

Genuer, R., Poggi, J.-M., and Tuleau, C. (2008). *Random Forests: some methodological insights*. arXiv: 0811.3619 [stat.ML].

Gislason, P. O., Benediktsson, J. A., and Sveinsson, J. R. (2006). "Random Forests for land cover classification". *Pattern Recognition Letters* 27.4, pp. 294–300. DOI: 10.1016/j.patrec.2005.08.011.

Goldstein, B. A., Polley, E. C., and Briggs, F. B. S. (2011). "Random Forests for Genetic Association Studies". *Statistical Applications in Genetics and Molecular Biology* 10.1. DOI: 10.2202/1544-6115.1691.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Element of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition*. Springer Series in Statistics. Springer New York Inc. ISBN: 9780387848587.

James, G. et al. (2013). *An Introduction to Statistical Learning: with Applications in R*. Springer Texts in Statistics. Springer New York. ISBN: 9781461471387.

Lin, Y. and Jeon, Y. (2006). "Random Forests and Adaptive Nearest Neighbors". *Journal of the American Statistical Association* 101.474, pp. 578–590. DOI: 10.1198/016214505000001230.

Martínez-Muñoz, G. and Suárez, A. (2010). "Out-of-bag estimation of the optimal sample size in bagging". *Pattern Recognition* 43.1, pp. 143–152. DOI: 10.1016/j.patcog.2009.05.010.

O'Donnel, M. S. and Ignizio, D. A. (2012). *Bioclimatic predictors for supporting ecological applications in the conterminous United States*. Data Series 691. Reston, VA: U.S. Geological Survey, p. 17. DOI: 10.3133/ds691.

Oshiro, T. M., Perez, P. S., and Baranauskas, J. A. (2012). "How Many Trees in a Random Forest?" In: *Machine Learning and Data Mining in Pattern Recognition*. Ed. by P. Perner. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 154–168.

Piramuthu, S. (2008). "Input data for decision trees". *Expert Systems with Applications* 34.2, pp. 1220–1226. DOI: 10.1016/j.eswa.2006.12.030.

Prasad, A. M., Iverson, L. R., and Liaw, A. (2006). "Newer Classification and Regression Tree Techniques: Bagging and Random Forests for Ecological Prediction". *Ecosystems* 9, pp. 181–199. DOI: 10.1007/s10021-005-0054-1.

Probst, P. and Boulesteix, A.-L. (2017). "To Tune or Not to Tune the Number of Trees in Random Forest". *Journal of Machine Learning Research* 18.1, pp. 1–18. DOI: 10.5555/3122009.3242038.

Probst, P., Wright, M. N., and Boulesteix, A.-L. (2019). "Hyperparameters and tuning strategies for random forest". *WIREs Data Mining and Knowledge Discovery* 9.3, e1301. DOI: 10.1002/widm.1301.

Sahr, K., White, D., and Kimerling, A. J. (2003). "Geodesic Discrete Global Grid Systems". *Cartography and Geographic Information Science* 30.2, pp. 121–134. DOI: 10.1559/152304003100011090.

Segal, M. R. (2004). "Machine Learning Benchmarks and Random Forest Regression". *UCSF: Center for Bioinformatics and Molecular Biostatistics.* URL: https://escholarship.org/uc/item/35x3v9t4.

Strobl, C. et al. (2008). "Conditional variable importance for random forests". *BMC Bioinformatics* 9.307. DOI: 10.1186/1471-2105-9-307.

Title, P. O. and Bemmels, J. B. (2018). "ENVIREM: an expanded set of bioclimatic and topographic variables increases flexibility and improves performance of ecological niche modeling". *Ecography* 41.2, pp. 291–307. DOI: https://doi.org/10.1111/ecog.02880.

Python libraries

Caswell, T. A. et al. (Sept. 2020). *matplotlib/matplotlib: REL: v3.3.2*. Version v3.3.2. DOI: 10.5281/zenodo.4030140.

Harris, C. R. et al. (Sept. 2020). "Array programming with NumPy". *Nature* 585.7825, pp. 357–362. DOI: 10.1038/s41586-020-2649-2.

Hunter, J. D. (2007). "Matplotlib: A 2D graphics environment". *Computing in Science & Engineering* 9.3, pp. 90–95. DOI: 10.1109/MCSE.2007.55.

Jordahl, K. et al. (July 2020). *geopandas/geopandas: v0.8.1*. Version v0.8.1. DOI: 10.5281/zenodo.3946761.

McKinney, W. (2010). “Data Structures for Statistical Computing in Python”. In: *Proceedings of the 9th Python in Science Conference*. Ed. by S. van der Walt and J. Millman, pp. 56–61. DOI: 10.25080/Majora-92bf1922-00a.

Pedregosa, F. et al. (2011). “Scikit-learn: Machine Learning in Python”. *Journal of Machine Learning Research* 12, pp. 2825–2830. URL: http://jmlr.org/papers/v12/pedregosa11a.html.

Reback, J. et al. (Dec. 2020). *pandas-dev/pandas: Pandas 1.2.0*. Version v1.2.0. DOI: 10.5281/zenodo.4394318.

Waskom, M. et al. (Dec. 2020). *mwaskom/seaborn: v0.11.1*. Version v0.11.1. DOI: 10.5281/zenodo.4379347.

Appendix A

Data analysis - Plots and tables

A.1 Bioclimatic predictors

A.1.1 Description of the variables

Table A.1: Bioclimatic predictors description. (O'Donnel and Ignizio, 2012)

Code	Name	Calculation	Unit
BIO01	*Annual mean temperature*	$\frac{1}{12}\sum_{i=1}^{12}\overline{T}_i$	Degrees Celsius
	The annual mean temperature.		
BIO02	*Annual mean diurnal range*	$\frac{1}{12}\sum_{i=1}^{12}(T_{max,i} - T_{min,i})$	Degrees Celsius
	The mean of monthly temperature ranges (monthly maximum minus monthly minimum).		
BIO03	*Isothermality*	$\frac{\text{BIO02}}{\text{BIO07}} \cdot 100$	Percentage
	The quantification of how large the day-to-night temperatures oscillate relative to the summer-to-winter (annual) oscillations.		
BIO04	*Temperature seasonality*	$\text{sd}(\overline{T}_1, \ldots, \overline{T}_{12}) \cdot 100$	Degrees Celsius
	The amount of temperature variation over a given year, based on the standard deviation of monthly temperature averages.		
BIO05	*Max temperature of the warmest month*	$\max_{i=1,\ldots,12} T_{max,i}$	Degrees Celsius
	The maximum monthly temperature occurrence over a given year.		
BIO06	*Min temperature of the coldest month*	$\min_{i=1,\ldots,12} T_{min,i}$	Degrees Celsius
	The minimum monthly temperature occurrence over a given year.		

Continue on next page

Code	Name	Calculation	Unit
BIO07	*Temperature annual range*	$\text{BIO05} - \text{BIO06}$	Degrees Celsius
	The measure of temperature variation over a given year.		
BIO08	*Mean temperature of the wettest quarter*	$q = \underset{j=1,\ldots,12}{\arg\max} \sum_{i=j}^{j+2} P_i$ [a] $\text{BIO08} = \frac{1}{3} \sum_{i=q}^{q+2} \overline{T}_i$	Degrees Celsius
	The quarterly index that approximates mean temperatures that prevail during the wettest season.		
BIO09	*Mean temperature of the driest quarter*	$q = \underset{j=1,\ldots,12}{\arg\min} \sum_{i=j}^{j+2} P_i$ [a] $\text{BIO09} = \frac{1}{3} \sum_{i=q}^{q+2} \overline{T}_i$	Degrees Celsius
	The quarterly index that approximates mean temperatures that prevail during the driest season.		
BIO10	*Mean temperature of the warmest quarter*	$q = \underset{j=1,\ldots,12}{\arg\max} \sum_{i=j}^{j+2} \overline{T}_i$ [a] $\text{BIO10} = \frac{1}{3} \sum_{i=q}^{q+2} \overline{T}_i$	Degrees Celsius
	The quarterly index that approximates mean temperatures that prevail during the warmest season.		
BIO11	*Mean temperature of the coldest quarter*	$q = \underset{j=1,\ldots,12}{\arg\min} \sum_{i=j}^{j+2} \overline{T}_i$ [a] $\text{BIO11} = \frac{1}{3} \sum_{i=q}^{q+2} \overline{T}_i$	Degrees Celsius
	The quarterly index that approximates mean temperatures that prevail during the coldest season.		
BIO12	*Annual precipitation*	$\sum_{i=1}^{12} P_i$	Millimetres
	The sum of monthly precipitation totals.		
BIO13	*Precipitation of the wettest month*	$\max_{i=1,\ldots,12} P_i$	Millimetres
	The total precipitation during the wettest month.		
BIO14	*Precipitation of the driest month*	$\min_{i=1,\ldots,12} P_i$	Millimetres
	The total precipitation during the driest month.		
BIO15	*Precipitation seasonality*	$\frac{\text{sd}(P_1, \ldots, P_{12})}{1 + \text{BIO12}/12} 100$	Percentage
	The measure of the variation in the monthly precipitation totals over the course of the year.		

Continue on next page

Code	Name	Calculation	Unit
BIO16	*Precipitation of the wettest quarter*	$\max\limits_{j=1,\dots,12} \sum_{i=j}^{j+2} P_i$ [a]	Millimetres
	The quarterly index that approximates total precipitation that prevails during the wettest season.		
BIO17	*Precipitation of the driest quarter*	$\min\limits_{j=1,\dots,12} \sum_{i=j}^{j+2} P_i$ [a]	Millimetres
	The quarterly index that approximates total precipitation that prevails during the driest season.		
BIO18	*Precipitation of the warmest quarter*	$q = \arg\max\limits_{j=1,\dots,12} \sum_{i=j}^{j+2} \overline{T}_i$ [a] $\text{BIO}18 = \sum_{i=q}^{q+2} P_i$	Millimetres
	The quarterly index that approximates total precipitation that prevails during the warmest season.		
BIO19	*Precipitation of the coldest quarter*	$q = \arg\min\limits_{j=1,\dots,12} \sum_{i=j}^{j+2} \overline{T}_i$ [a] $\text{BIO}19 = \sum_{i=q}^{q+2} P_i$	Millimetres
	The quarterly index that approximates total precipitation that prevails during the coldest season.		

Notation :

i identifies a specific month of the year, e.g. i=1 means January, etc.

$T_{max,i}$ denotes the mean of daily maximum temperature for month i.

$T_{min,i}$ denotes the mean of daily minimum temperature for month i.

$\overline{T}_i$ denotes the average temperature for month i, i.e. $\overline{T}_i = (T_{max,i} - T_{min,i})/2$.

P_i denotes the total precipitation for month i.

Notes :

a Quarterly indices are based on 3 months intervals. So, fixed a month i, the next two months $i+1$ and $i+2$ are evaluated. In order to evaluate the last two months of the year, the quarters are defined by using the months at the beginning of the same year. Therefore, with a little abuse of notation, when $i = 12$, then $i+1$ and $i+2$ should be interpreted as 1 and 2 respectively.

Remark: Even though predictors in the table has been defined with respect to a specific year, the corresponding values in the dataset have been obtained by averaging their values over the 30 years interval 1971-2000.

A.1.2 Outliers

BP code	N. lower outliers	N. upper outliers	N. outliers	Outliers %	BP code	N. lower outliers	N. upper outliers	N. outliers	Outliers %
BIO01	1562	-	1562	2.80%	BIO11	-	-	-	-
BIO02	2	48	50	0.09%	BIO12	-	2939	2939	5.26%
BIO03	-	-	-	-	BIO13	-	1479	1479	2.65%
BIO04	-	70	70	0.13%	BIO14	-	5268	5268	9.43%
BIO05	4934	-	4934	8.83%	BIO15	-	498	498	0.89%
BIO06	-	-	-	-	BIO16	-	1689	1689	3.02%
BIO07	-	77	77	0.14%	BIO17	-	5050	5050	9.04%
BIO08	5046	-	5046	9.03%	BIO18	-	2256	2256	4.04%
BIO09	-	-	-	-	BIO19	-	6973	6973	12.48%
BIO10	4834	-	4834	8.65%					

- **N. lower/upper outliers** column contains the count of outliers lower/larger than the minimum/maximum whisker.
- **N. outliers** column contains the total count of outliers.
- **Outliers %** column contains the percentage of observations in the dataset that is an outlier for that bioclimatic predictor.

Table A.2: Count of outliers for every bioclimatic predictor.

	N. outliers			Outliers %		
Climatic zone	**Frigid**	**Temperate**	**Torrid**	**Frigid**	**Temperate**	**Torrid**
BP code						
BIO01	1562	-	-	100.00%	-	-
BIO02	1	21	28	2.00%	42.00%	56.00%
BIO03	-	-	-	-	-	-
BIO04	17	53	-	24.29%	75.71%	-
BIO05	4930	4	-	99.92%	0.08%	-
BIO06	-	-	-	-	-	-
BIO07	13	64	-	16.88%	83.12%	-
BIO08	4994	52	-	98.97%	1.03%	-
BIO09	-	-	-	-	-	-
BIO10	4831	3	-	99.94%	0.06%	-
BIO11	-	-	-	-	-	-
BIO12	-	220	2719	-	7.49%	92.51%
BIO13	-	261	1218	-	17.65%	82.35%
BIO14	69	2337	2862	1.31%	44.36%	54.33%
BIO15	413	-	85	82.93%	-	17.07%
BIO16	-	262	1427	-	15.51%	84.49%
BIO17	51	2039	2960	1.01%	40.38%	58.61%
BIO18	-	512	1744	-	22.70%	77.30%
BIO19	120	1588	5265	1.72%	22.77%	75.51%

- **N. outliers** columns contain the count of outliers that falls into every specific climatic zone.
- **Outliers %** columns contain the percentage of outliers that falls into every specific climatic zone.

Table A.3: Count of outliers for every bioclimatic predictor, divided among the three climatic zones.

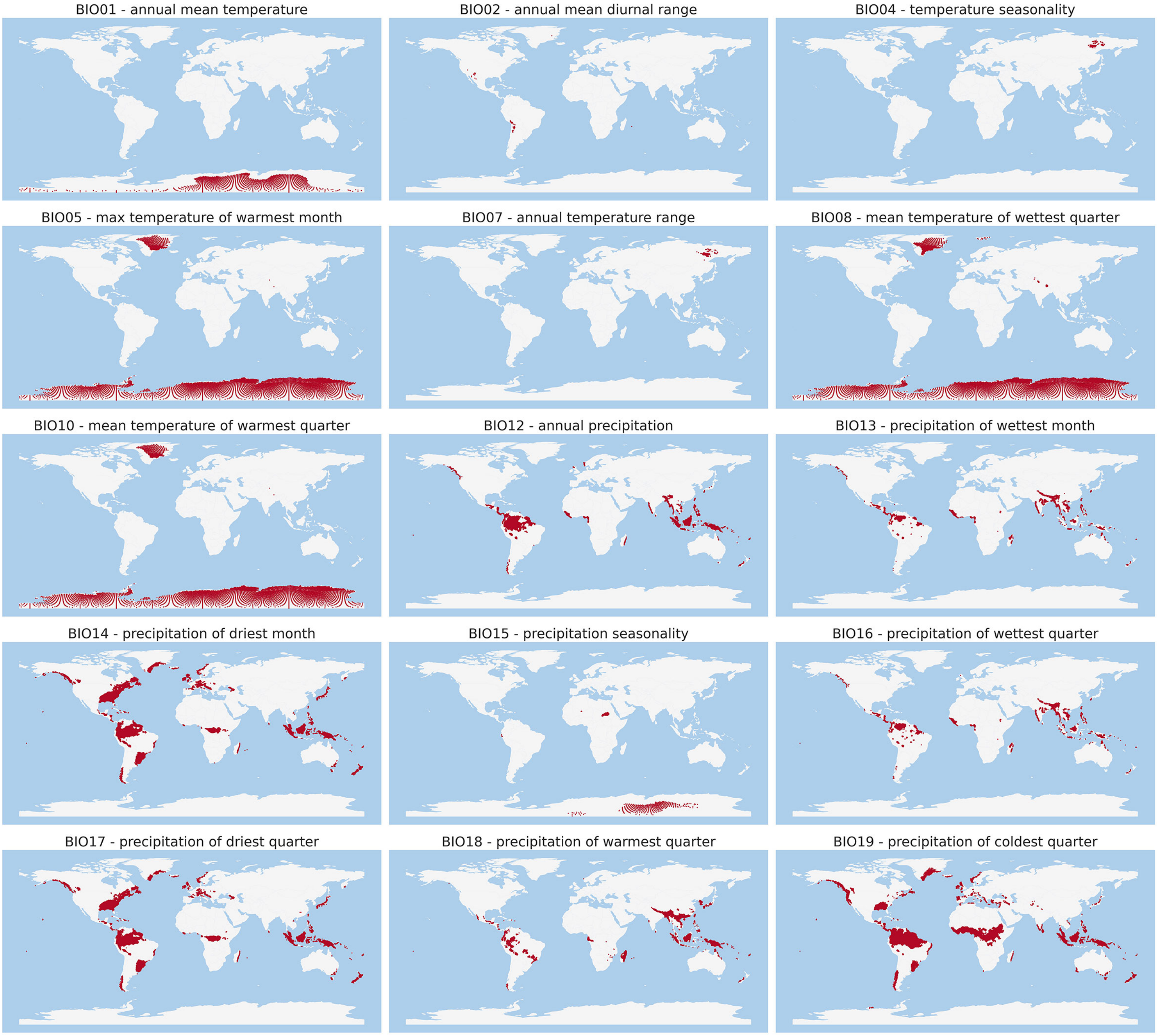
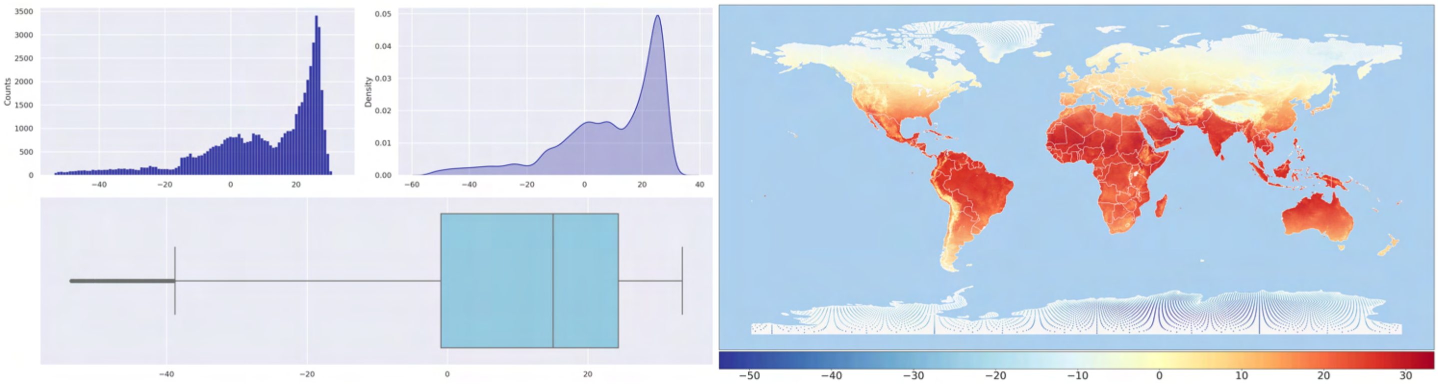

BIO01 - annual mean temperature
BIO02 - annual mean diurnal range
BIO04 - temperature seasonality
BIO05 - max temperature of warmest month
BIO07 - annual temperature range
BIO08 - mean temperature of wettest quarter
BIO10 - mean temperature of warmest quarter
BIO12 - annual precipitation
BIO13 - precipitation of wettest month
BIO14 - precipitation of driest month
BIO15 - precipitation seasonality
BIO16 - precipitation of wettest quarter
BIO17 - precipitation of driest quarter
BIO18 - precipitation of warmest quarter
BIO19 - precipitation of coldest quarter

Figure A.1: Geographic map of outliers for each bioclimatic predictors. Only predictors with a positive count of outliers are displayed.

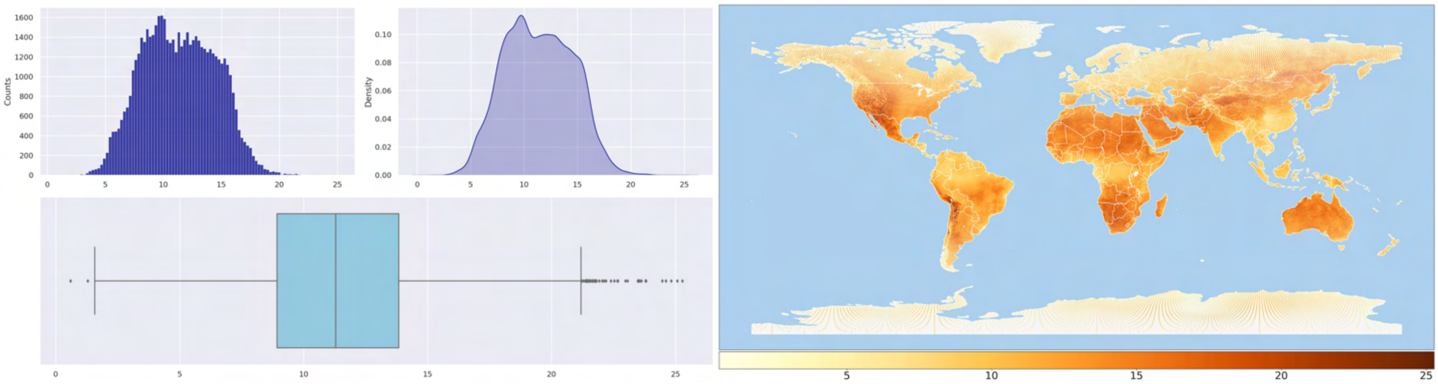
A.1.3 Empirical distributions



(a) Empirical distribution.

(b) Geographic map.

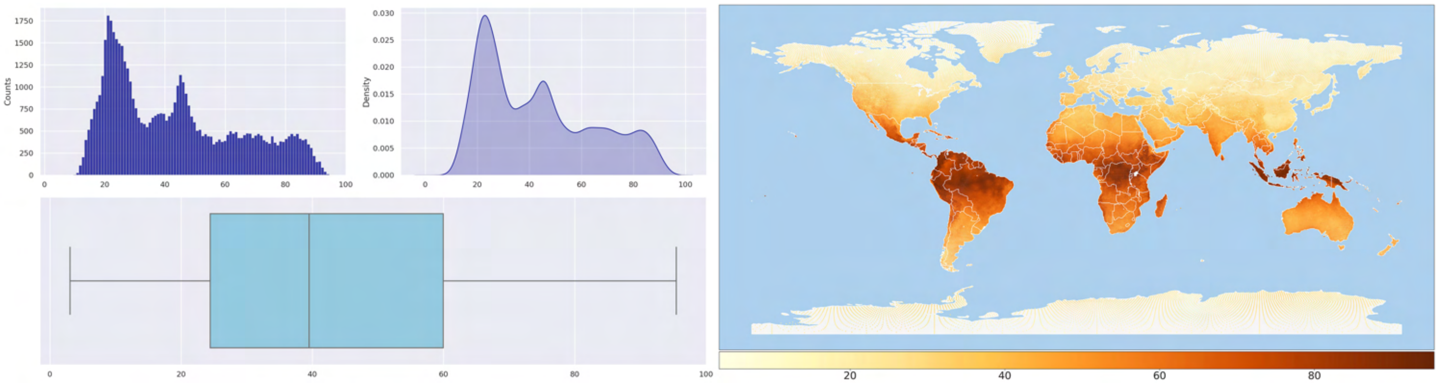
Figure A.2: BIO01 - *Annual mean temperature.*



(a) Empirical distribution.

(b) Geographic map.

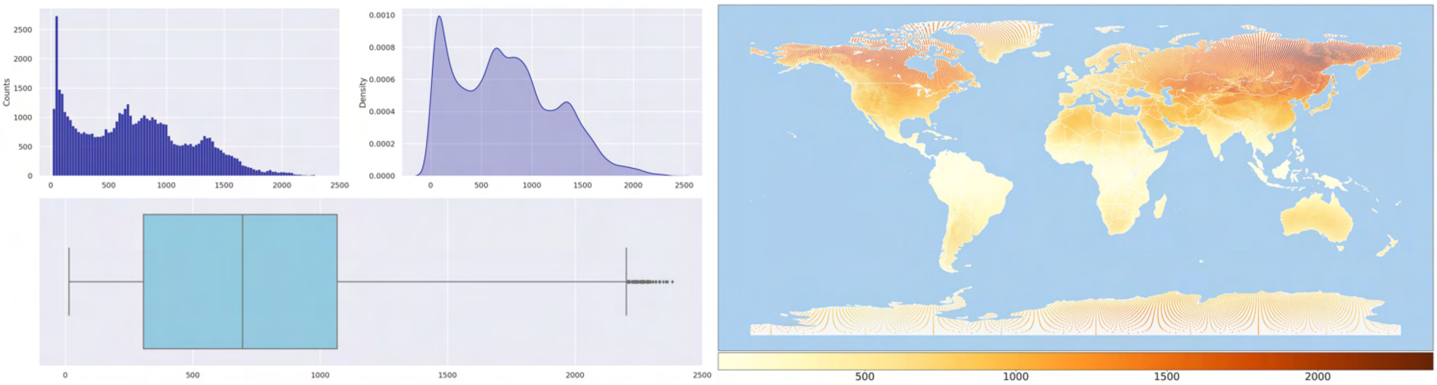
Figure A.3: BIO02 - *Annual mean diurnal range.*



(a) Empirical distribution.

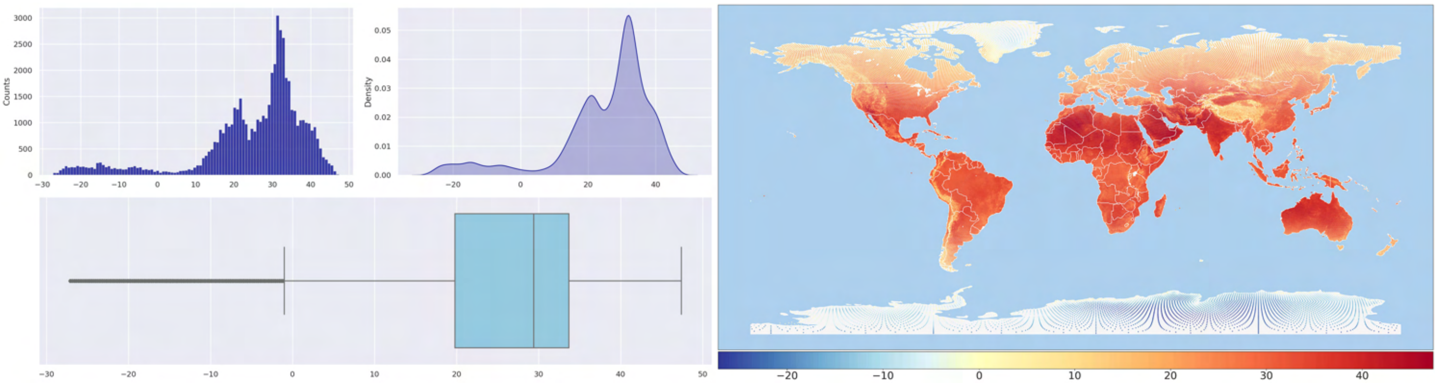
(b) Geographic map.

Figure A.4: BIO03 - *Isothermality.*



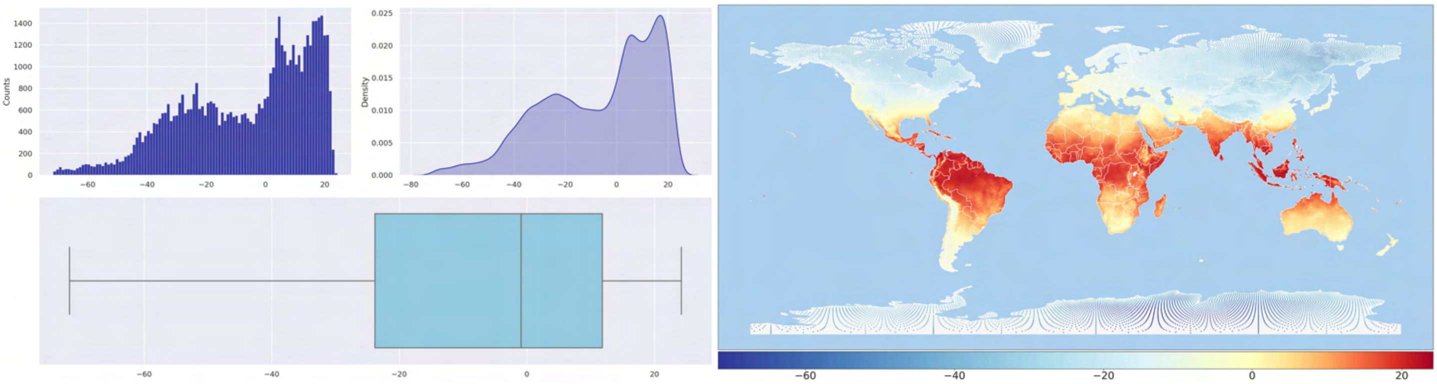
(a) Empirical distribution. (b) Geographic map.

Figure A.5: BIO04 - *Temperature seasonality.*



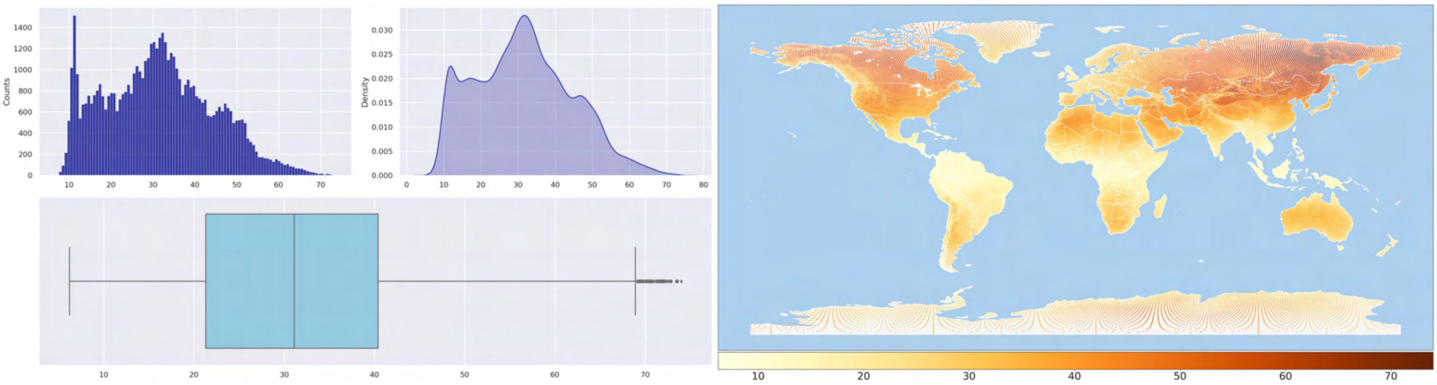
(a) Empirical distribution. (b) Geographic map.

Figure A.6: BIO05 - *Max temperature of the warmest month.*



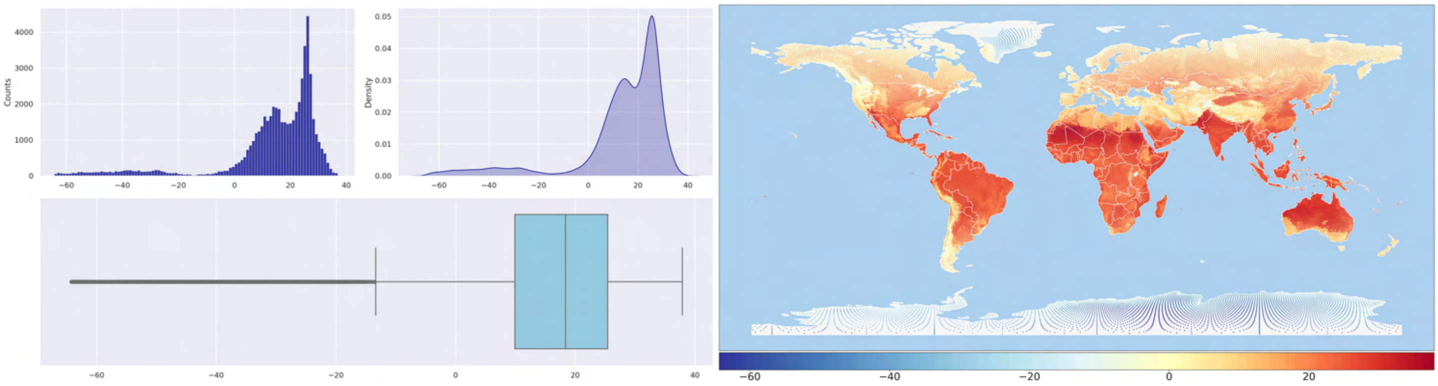
(a) Empirical distribution. (b) Geographic map.

Figure A.7: BIO06 - *Min temperature of the coldest month.*



(a) Empirical distribution. (b) Geographic map.

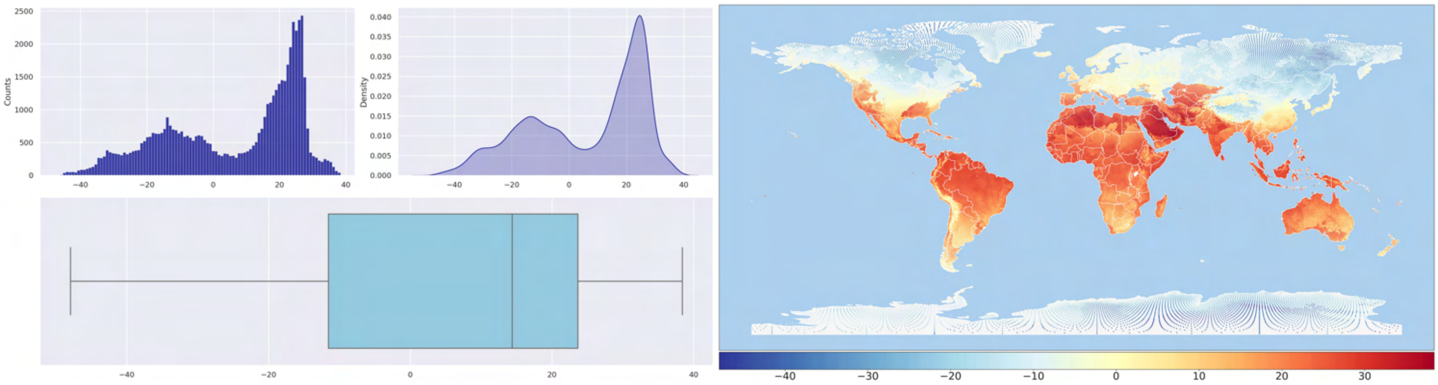
Figure A.8: BIO07 - *Temperature annual range.*



(a) Empirical distribution.

(b) Geographic map.

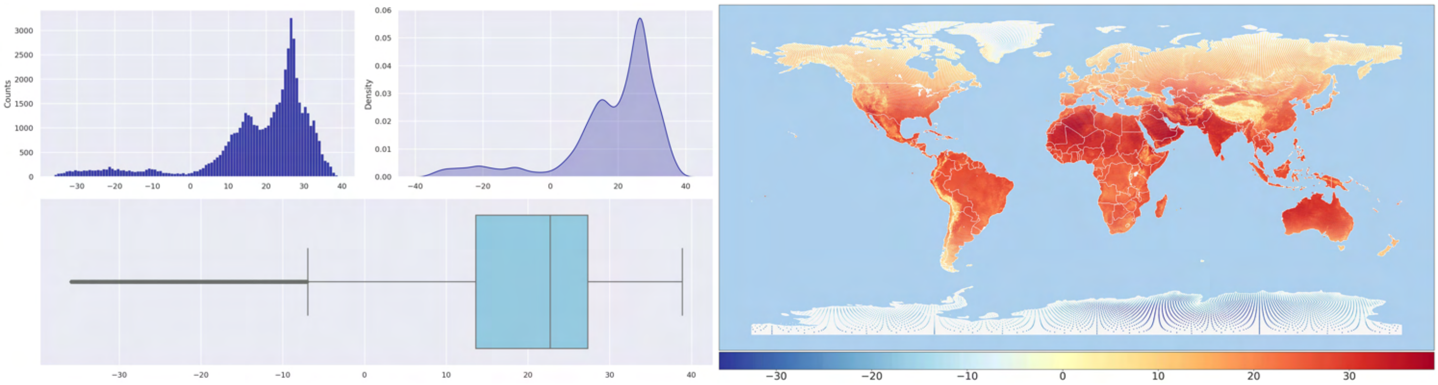
Figure A.9: BIO08 - *Mean temperature of the wettest quarter.*



(a) Empirical distribution.

(b) Geographic map.

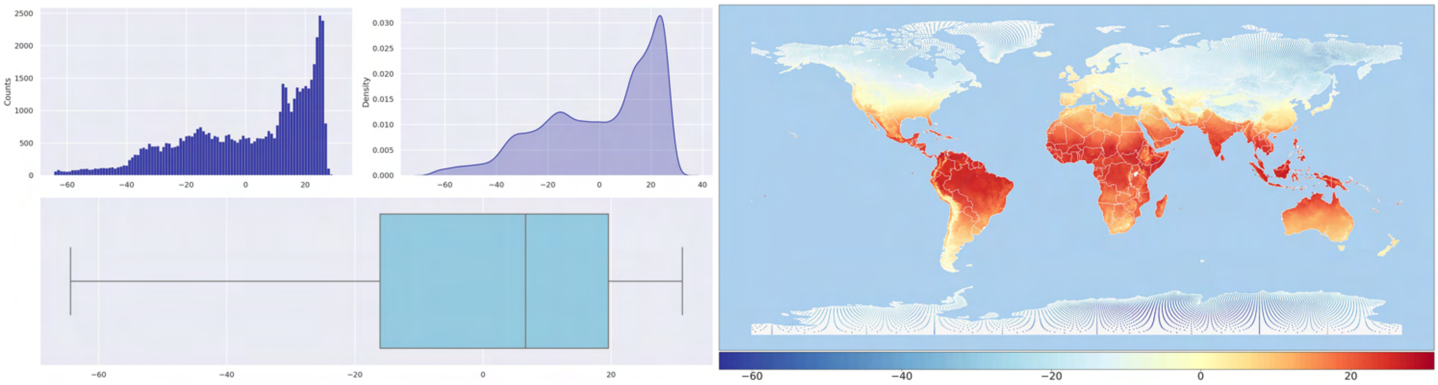
Figure A.10: BIO09 - *Mean temperature of the driest quarter.*



(a) Empirical distribution.

(b) Geographic map.

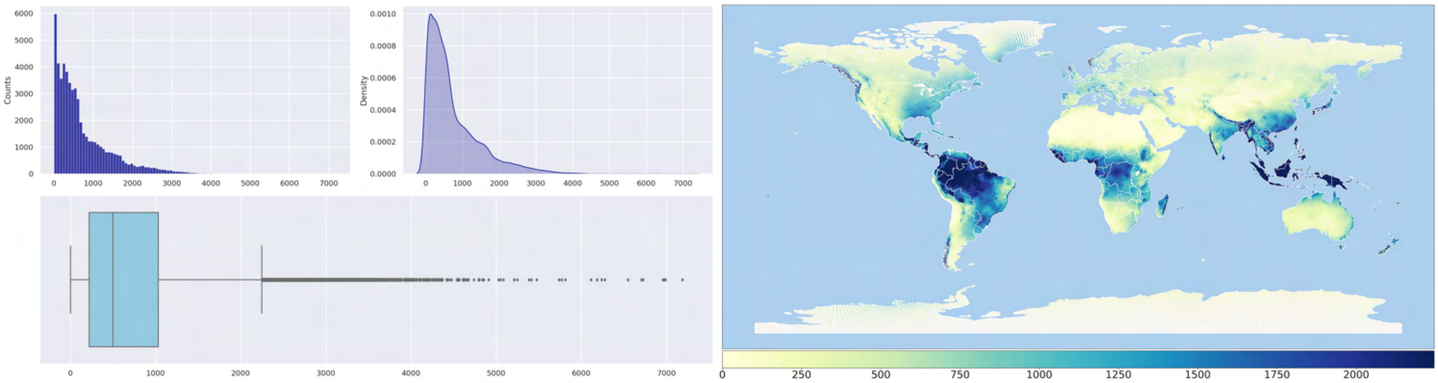
Figure A.11: BIO10 - *Mean temperature of the warmest quarter.*



(a) Empirical distribution.

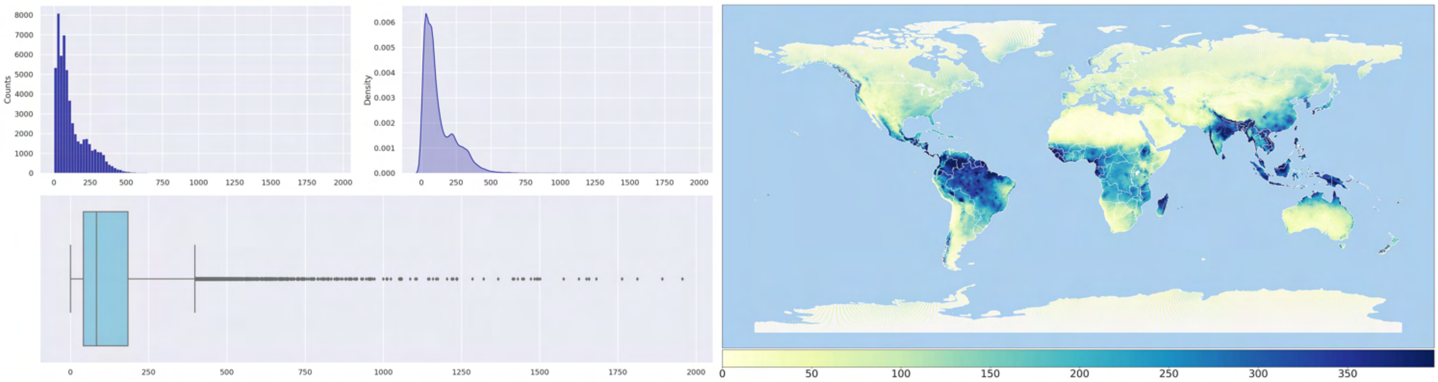
(b) Geographic map.

Figure A.12: BIO11 - *Mean temperature of the coldest quarter.*



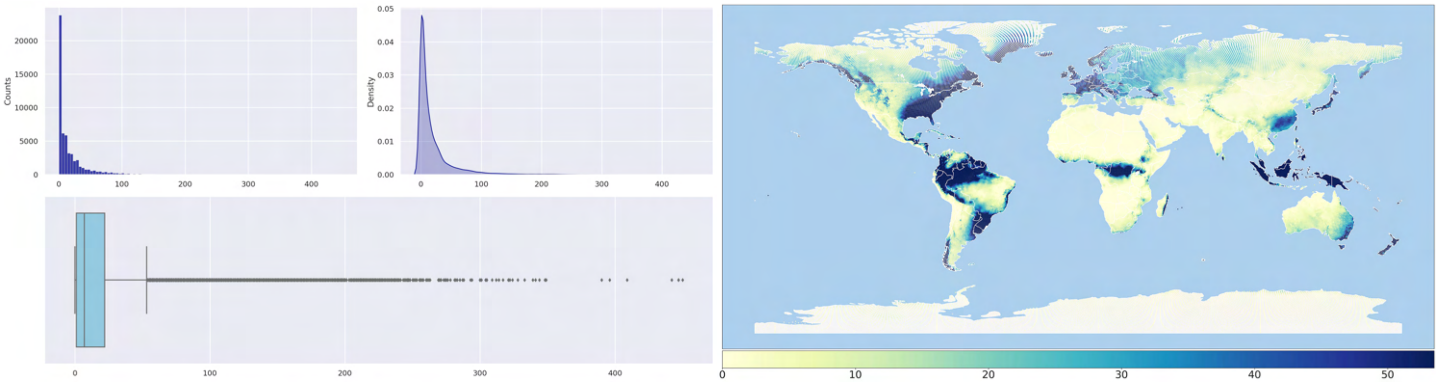
(a) Empirical distribution. (b) Geographic map.

Figure A.13: BIO12 - *Annual precipitation.*



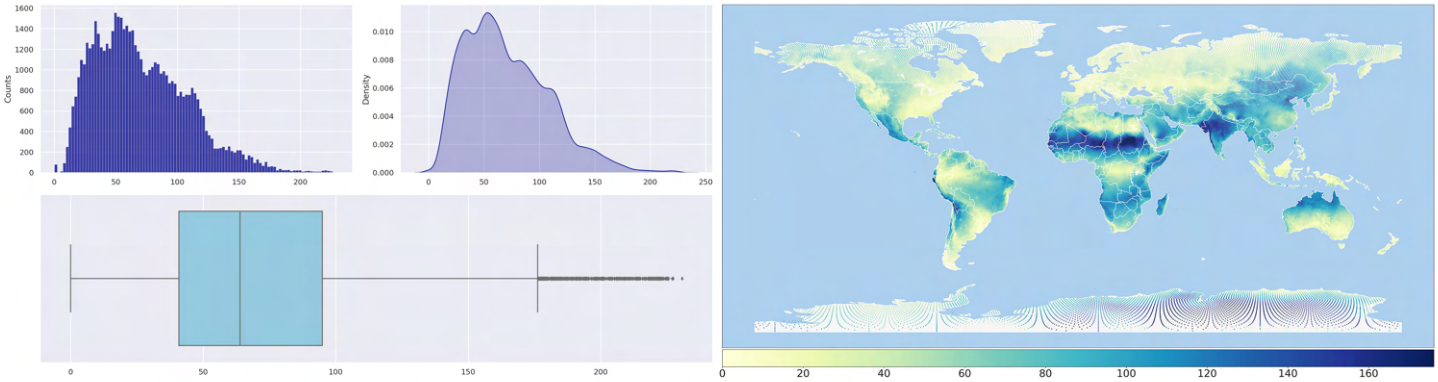
(a) Empirical distribution. (b) Geographic map.

Figure A.14: BIO13 - *Precipitation of the wettest month.*



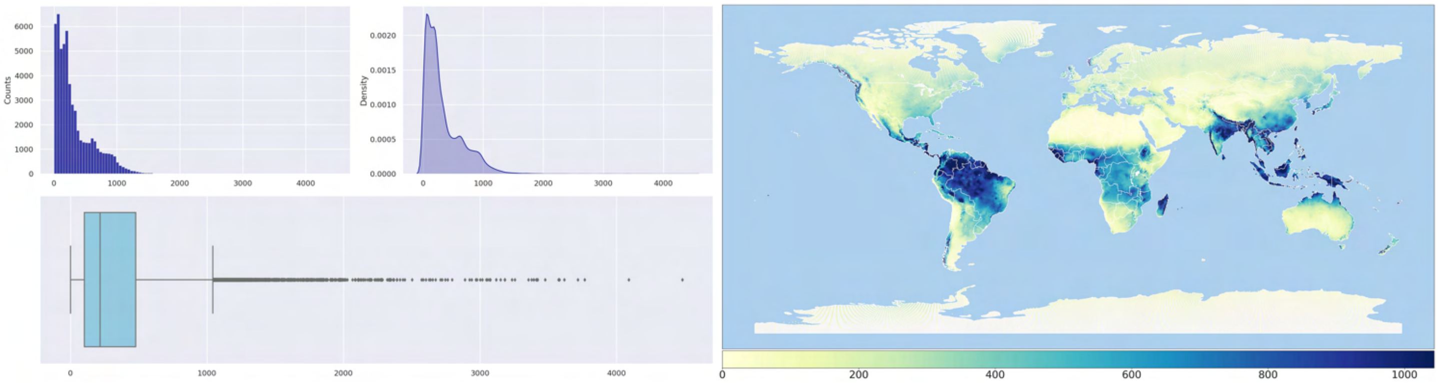
(a) Empirical distribution. (b) Geographic map.

Figure A.15: BIO14 - *Precipitation of the driest month.*



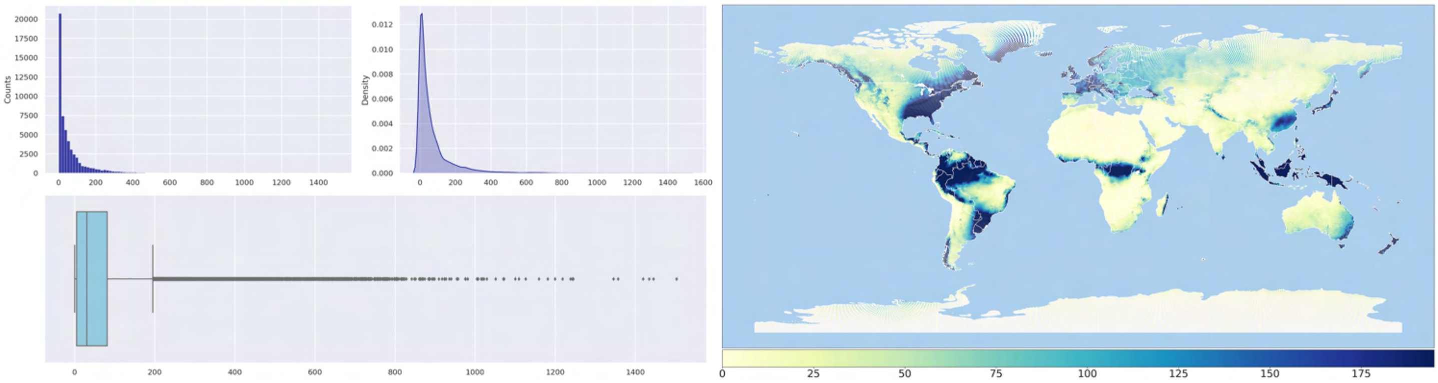
(a) Empirical distribution. (b) Geographic map.

Figure A.16: BIO15 - *Precipitation seasonality.*



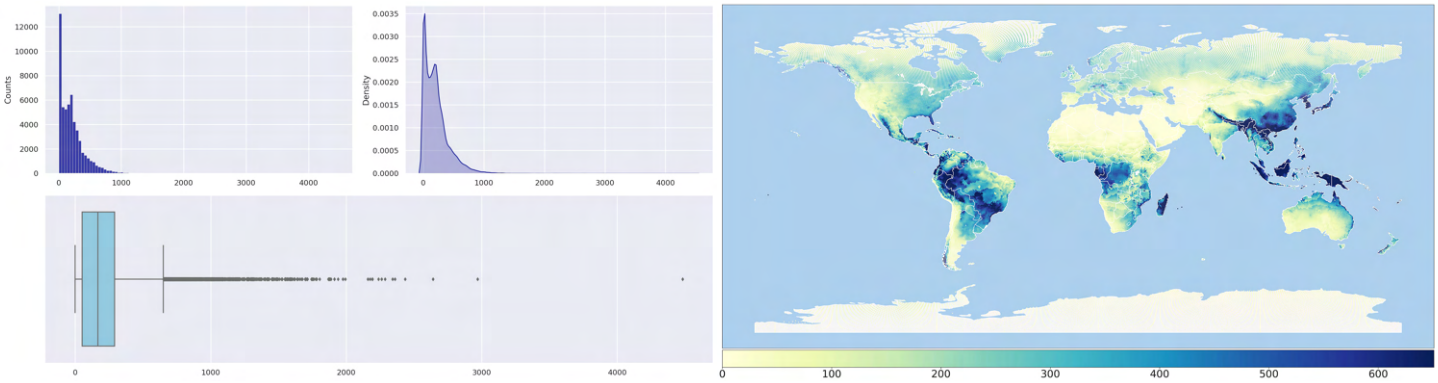
(a) Empirical distribution. (b) Geographic map.

Figure A.17: BIO16 - *Precipitation of the wettest quarter.*



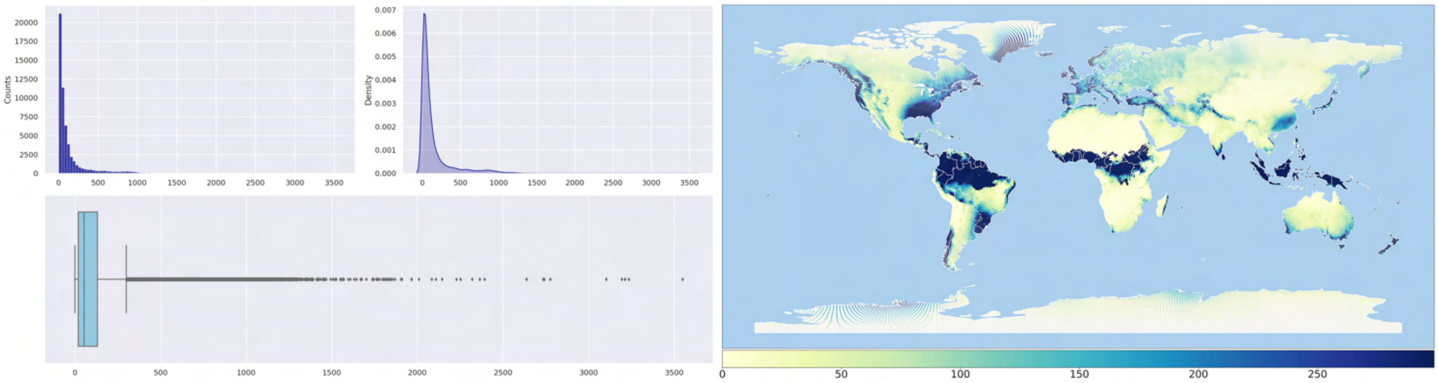
(a) Empirical distribution. (b) Geographic map.

Figure A.18: BIO17 - *Precipitation of the driest quarter.*



(a) Empirical distribution. (b) Geographic map.

Figure A.19: BIO18 - *Precipitation of the warmest quarter.*



(a) Empirical distribution. (b) Geographic map.

Figure A.20: BIO19 - *Precipitation of the coldest quarter.*

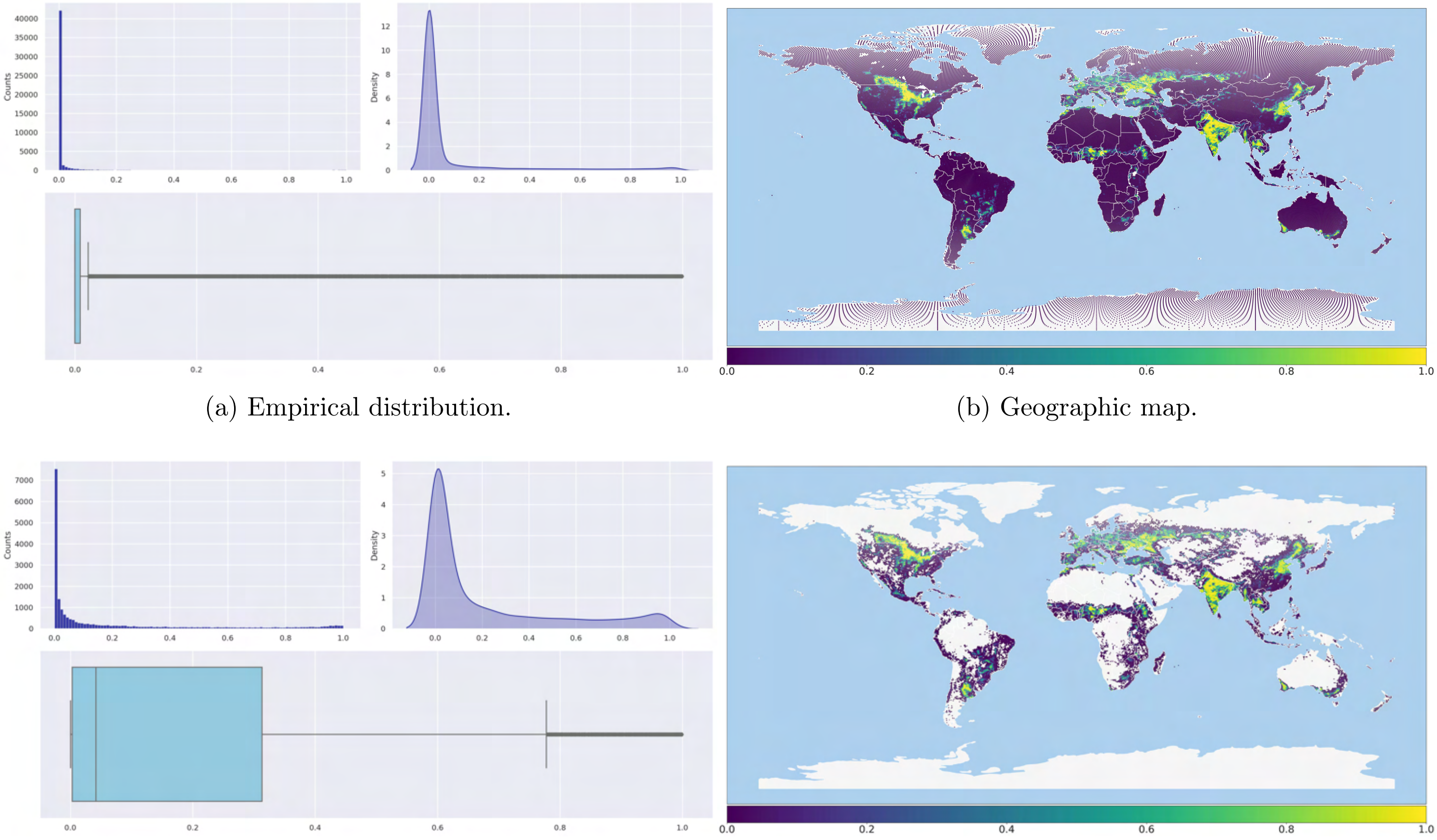
A.2 Land cover fraction variables

A.2.1 Description of the variables

Code	**Name**	**Description**
IGBP00	*Water Bodies*	At least 60% of area is covered by permanent water bodies.
IGBP01	*Evergreen Needleleaf Forests*	Dominated by evergreen conifer trees. `canopy`$> 2m$, `tree cover`$> 60\%$.
IGBP02	*Evergreen Broadleaf Forests*	Dominated by evergreen broadleaf and palmate trees. `canopy`$> 2m$, `tree cover`$> 60\%$.
IGBP03	*Deciduous Needleleaf Forests*	Dominated by deciduous needleleaf (larch) trees. `canopy`$> 2m$, `tree cover`$> 60\%$.
IGBP04	*Deciduous Broadleaf Forests*	Dominated by deciduous broadleaf trees. `canopy`$> 2m$, `tree cover`$> 60\%$.
IGBP05	*Mixed Forests*	Dominated by neither deciduous nor evergreen (40-60% of each) tree type. `canopy`$> 2m$, `tree cover`$> 60\%$.
IGBP06	*Closed Shrublands*	Dominated by woody perennials (1-2m height) $> 60\%$ cover.
IGBP07	*Open Shrublands*	Dominated by woody perennials (1-2m height) 10-60% cover.
IGBP08	*Woody Savannas*	`tree cover` 30-60%, `canopy`$> 2m$.
IGBP09	*Savannas*	`tree cover` 10-30%, `canopy`$> 2m$.
IGBP10	*Grasslands*	Dominated by herbaceous annuals ($< 2m$).
IGBP11	*Permanent Wetlands*	Permanently inundated lands with 30-60% water cover and $> 10\%$ vegetated cover.
IGBP12	*Croplands*	At least 60% of area is cultivated cropland.
IGBP13	*Urban and built-up lands*	At least 30% impervious surface area including building materials, asphalt and vehicles.
IGBP14	*Croplands/Natural vegetation mosaics*	Mosaics of small-scale cultivation 40-60%, with natural tree, shrub or herbaceous vegetation.
IGBP15	*Permanent Snow and Ice*	At least 60% of area is covered by snow and ice for at least 10 months of the year.
IGBP16	*Barren*	At least 60% of area is non-vegetated barren (sand, rock, soil) areas with less than 10% vegetation.

Table A.4: Description of land cover classes in *MCD12Q1* (Friedl and Sulla-Menashe, 2019) - IGBP legend.

A.2.2 Empirical distributions



(a) Empirical distribution.

(b) Geographic map.

(c) Empirical distribution of positive fractions.

(d) Geographic map of positive fractions.

Figure A.21: IGBP12 - *Croplands*.

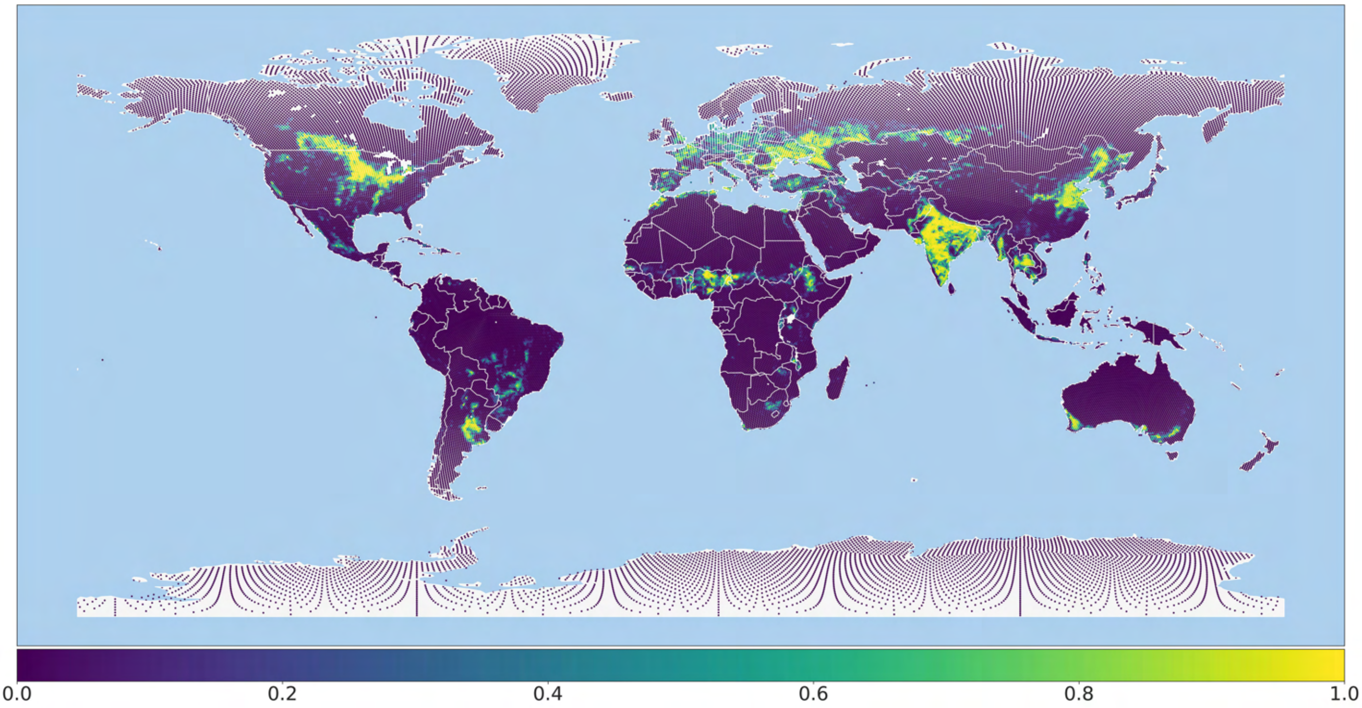
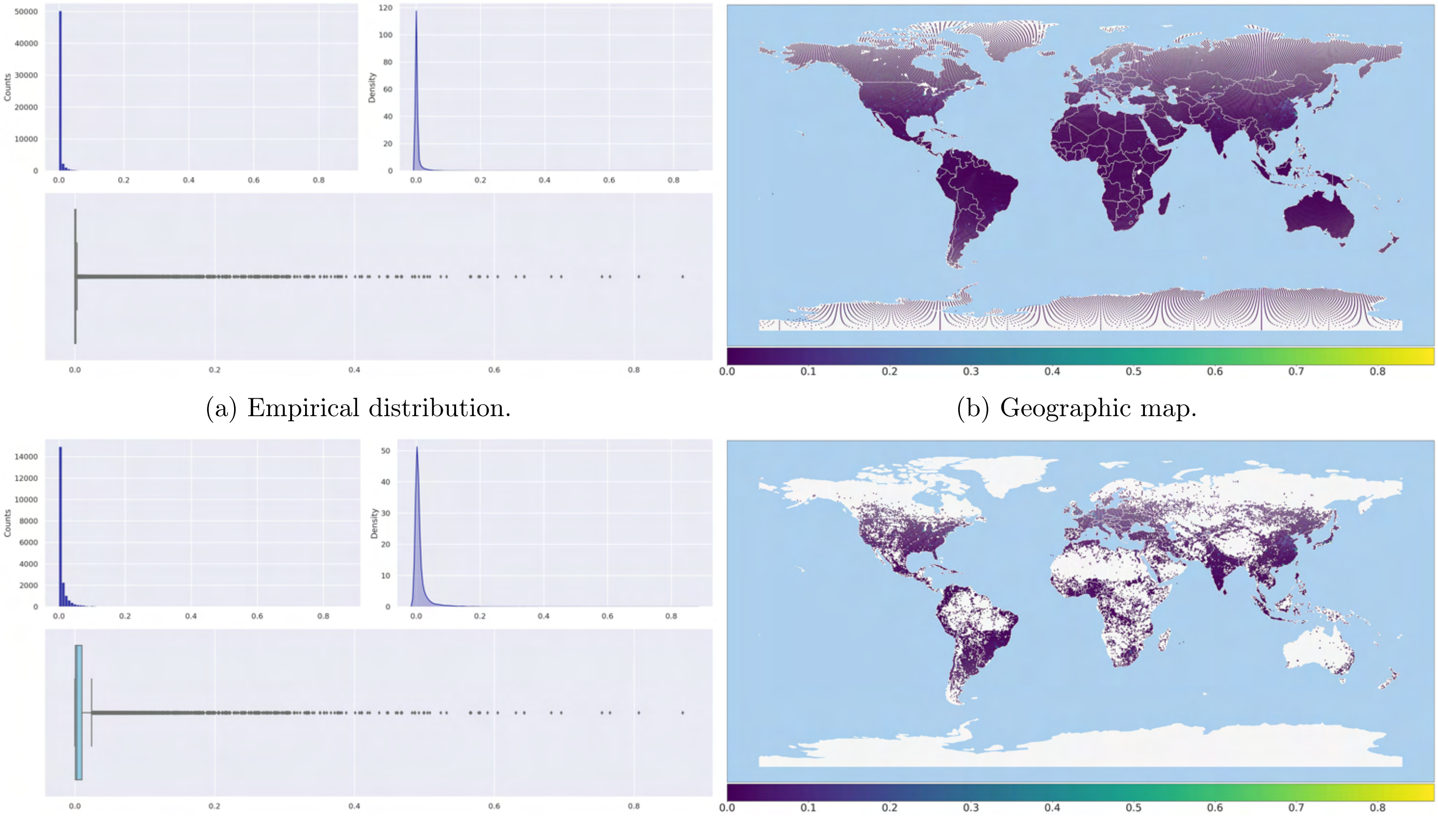


Figure A.22: Geographic map of the distribution of croplands fractions. Enlargement of Figure A.21b.



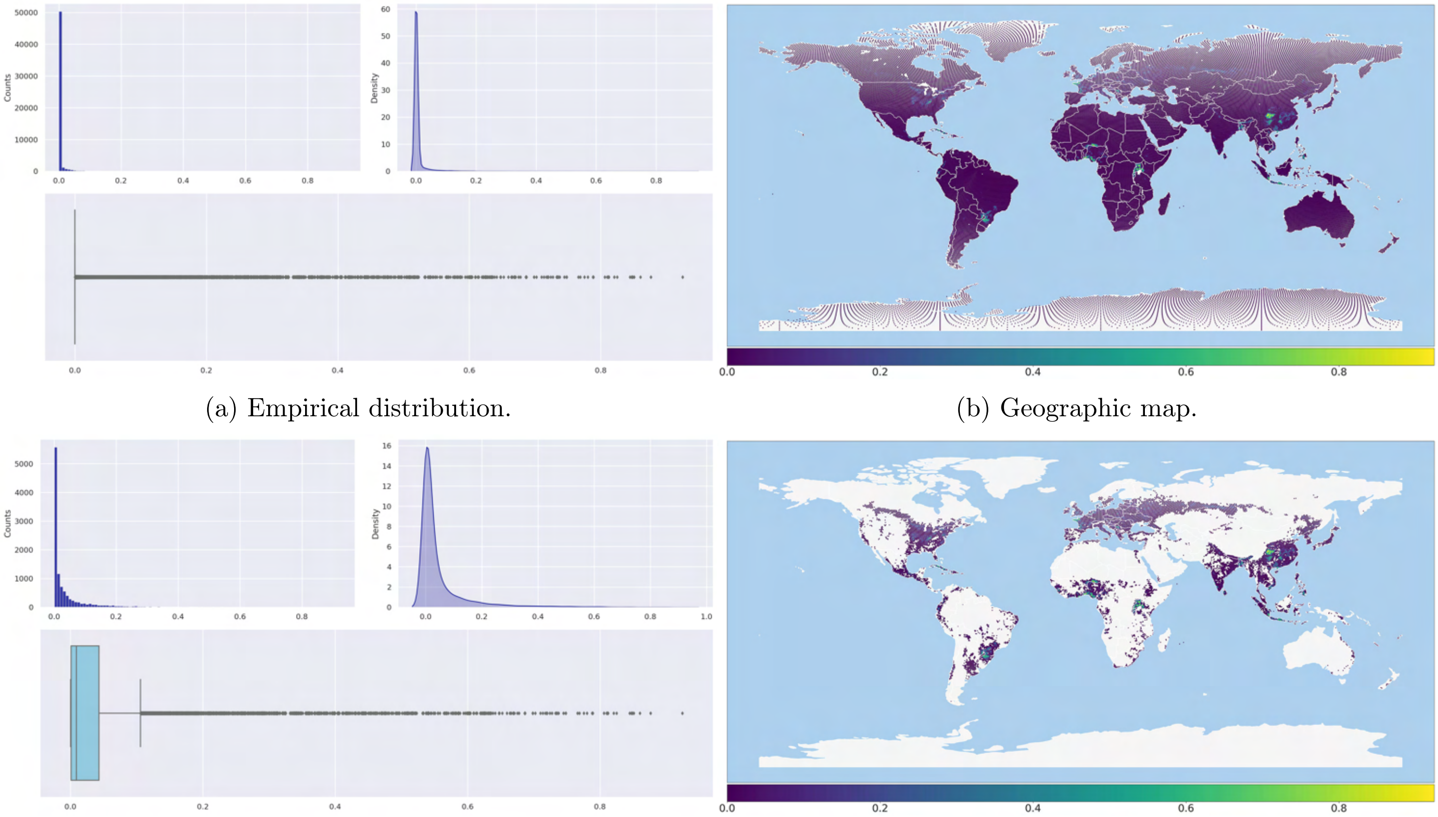
(a) Empirical distribution.

(b) Geographic map.

(c) Empirical distribution of positive fractions.

(d) Geographic map of positive fractions.

Figure A.23: IGBP13 - *Urban and built-up lands.*



(a) Empirical distribution.

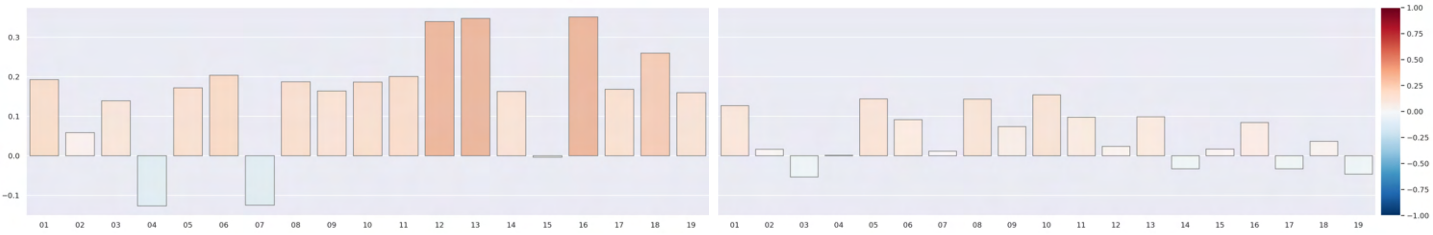
(b) Geographic map.

(c) Empirical distribution of positive fractions.

(d) Geographic map of positive fractions.

Figure A.24: IGBP14 - *Croplands/Natural vegetation mosaics.*

A.2.3 Pairwise relationships between responses and bioclimatic predictors



(a) Spearman coefficients.

(b) Pearson coefficients.

Figure A.25: Correlation coefficients between the response variable IGBP12 and the bioclimatic predictors.

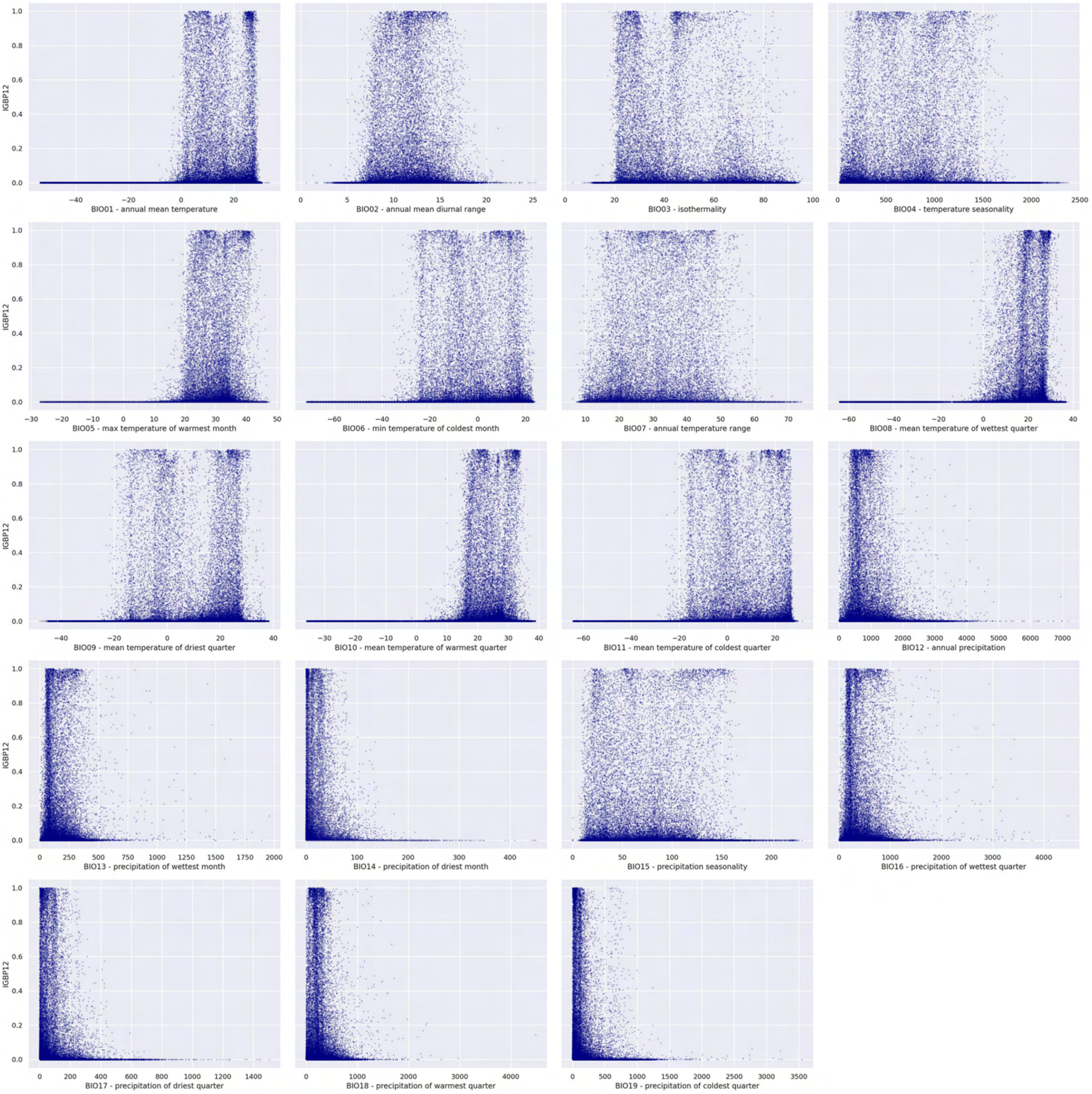


Figure A.26: Pairwise scatter plots between the response variable IGBP12 and the bioclimatic predictors.

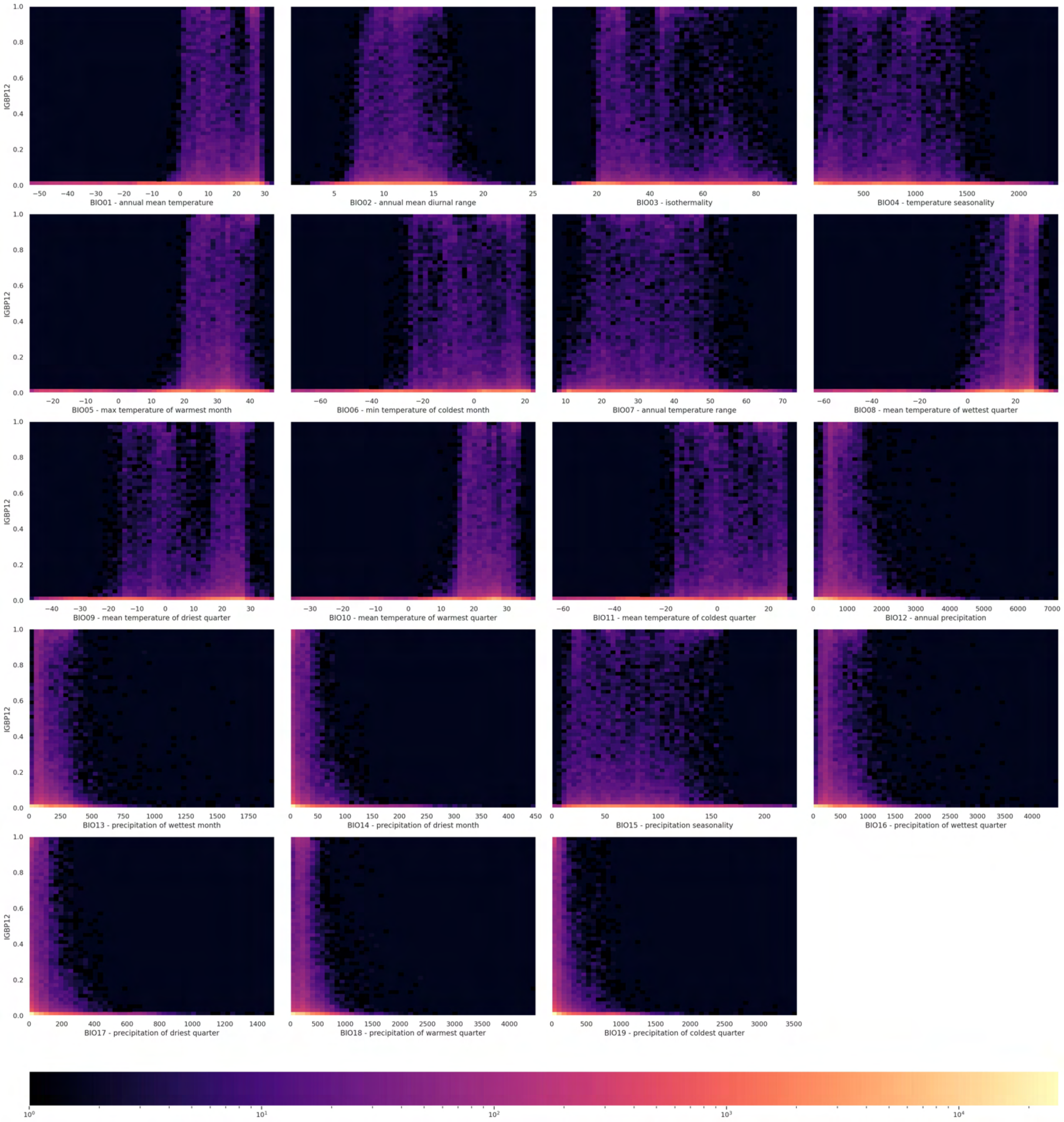
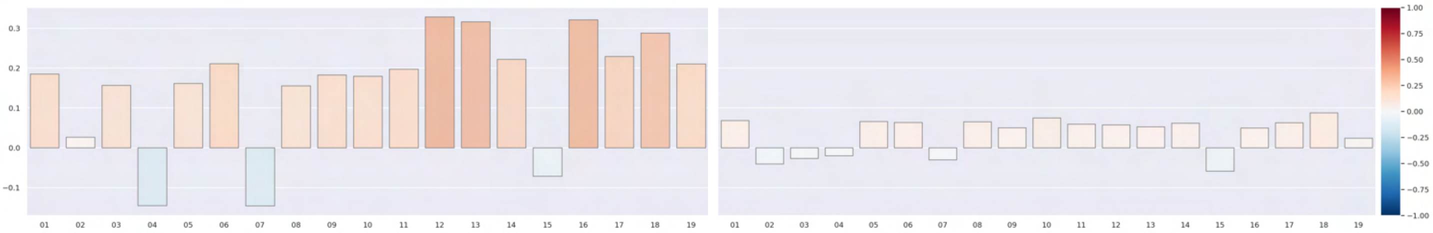


Figure A.27: Pairwise 2D histograms between the response variable IGBP12 and the bioclimatic predictors.



(a) Spearman coefficients.

(b) Pearson coefficients.

Figure A.28: Correlation coefficients between the response variable IGBP13 and the bioclimatic predictors.

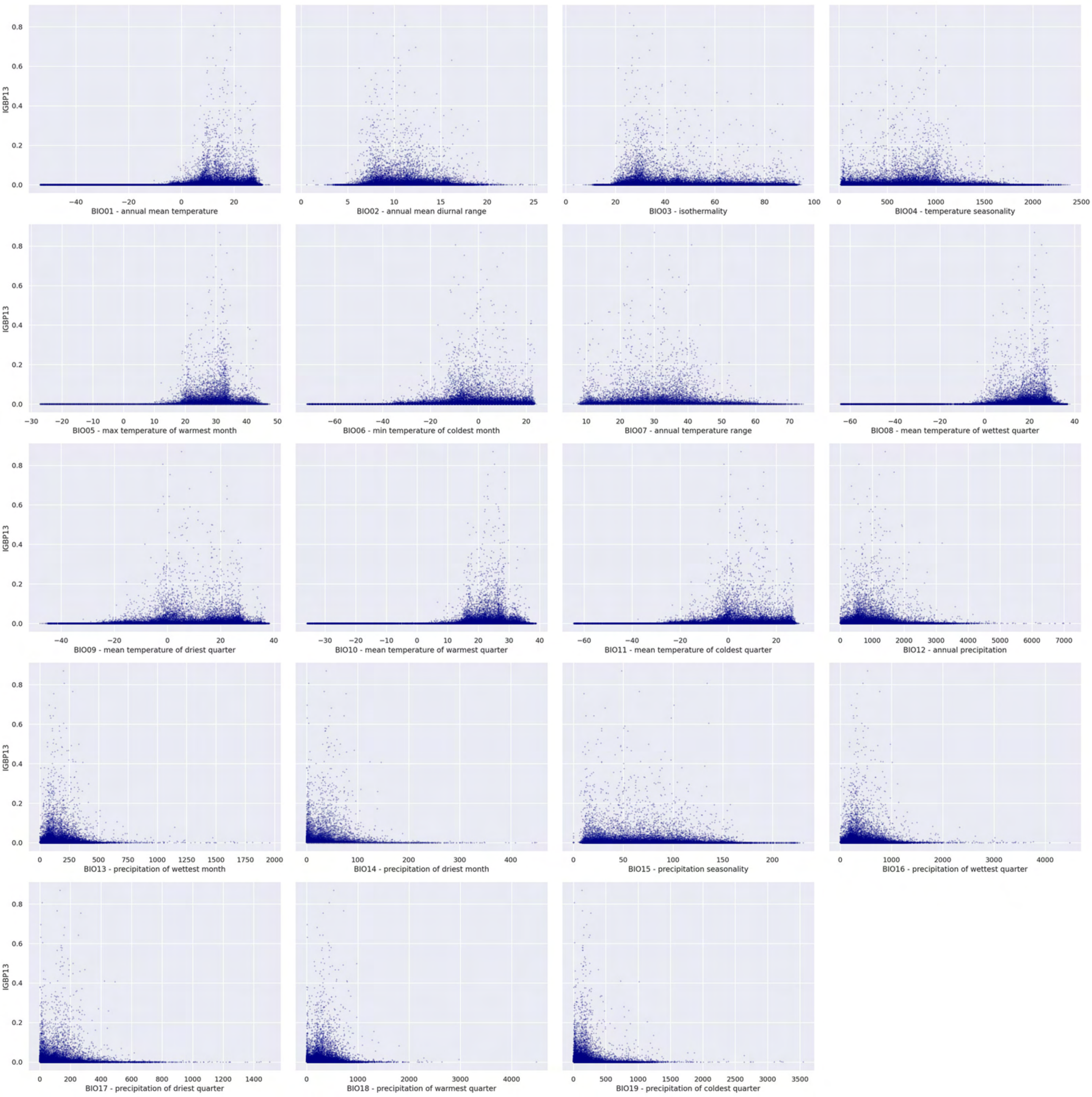


Figure A.29: Pairwise scatter plots between the response variable IGBP13 and the bioclimatic predictors.

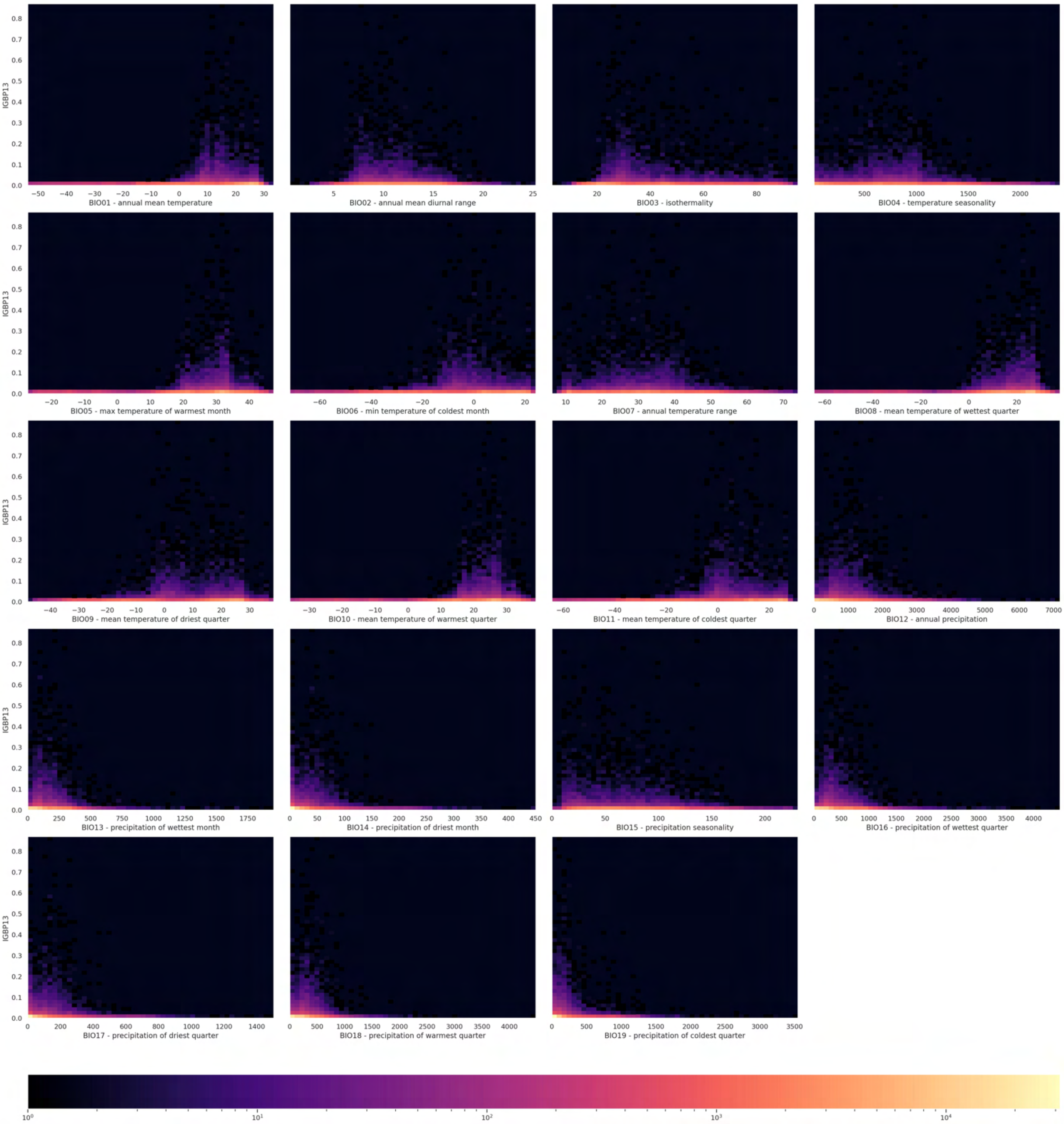
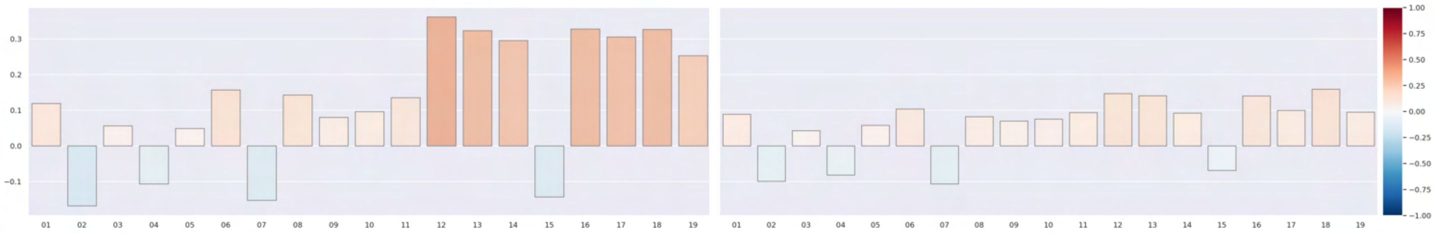


Figure A.30: Pairwise 2D histograms between the response variable IGBP13 and the bioclimatic predictors.



(a) Spearman coefficients.

(b) Pearson coefficients.

Figure A.31: Correlation coefficients between the response variable IGBP14 and the bioclimatic predictors.

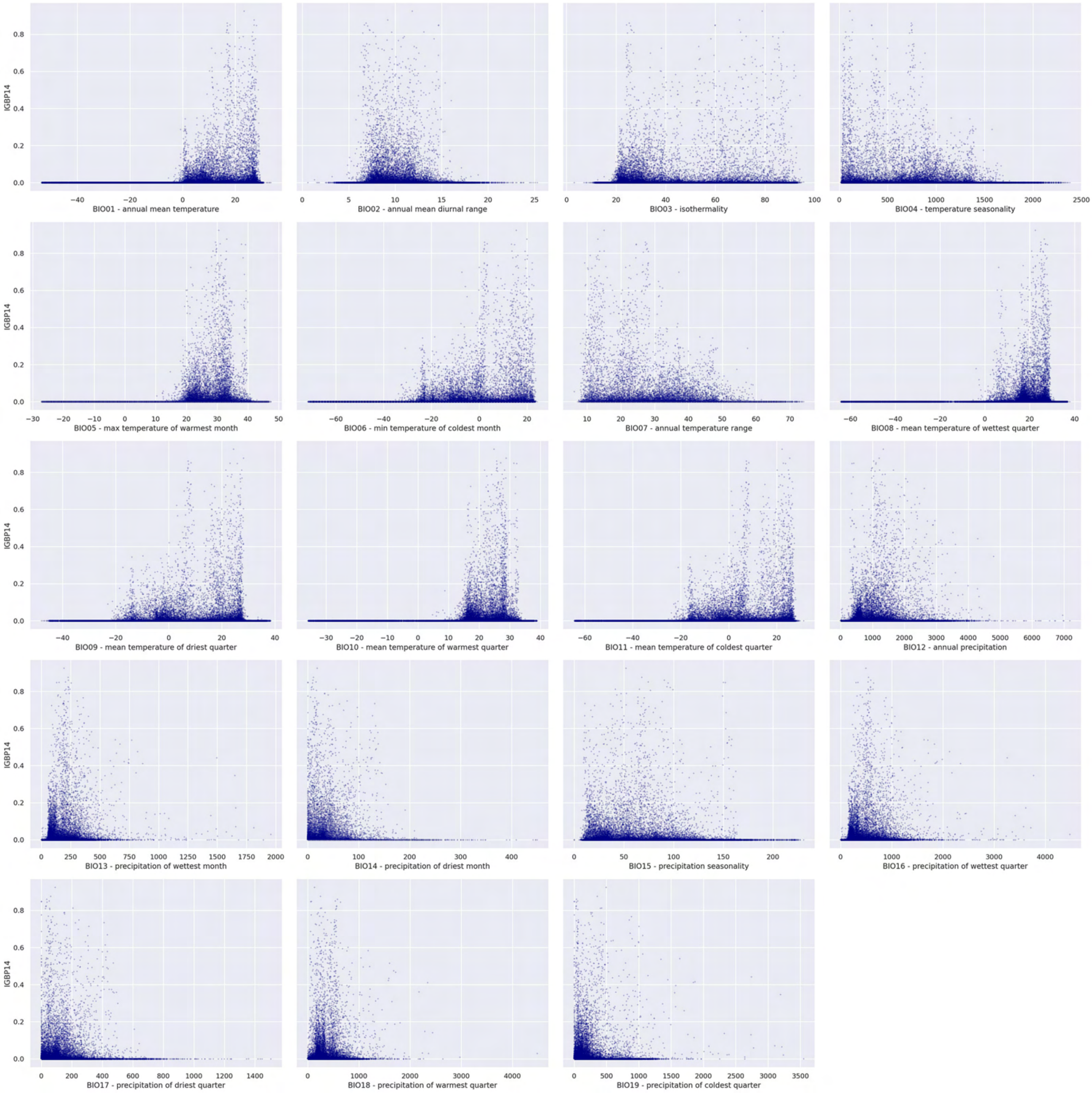


Figure A.32: Pairwise scatter plots between the response variable IGBP14 and the bioclimatic predictors.

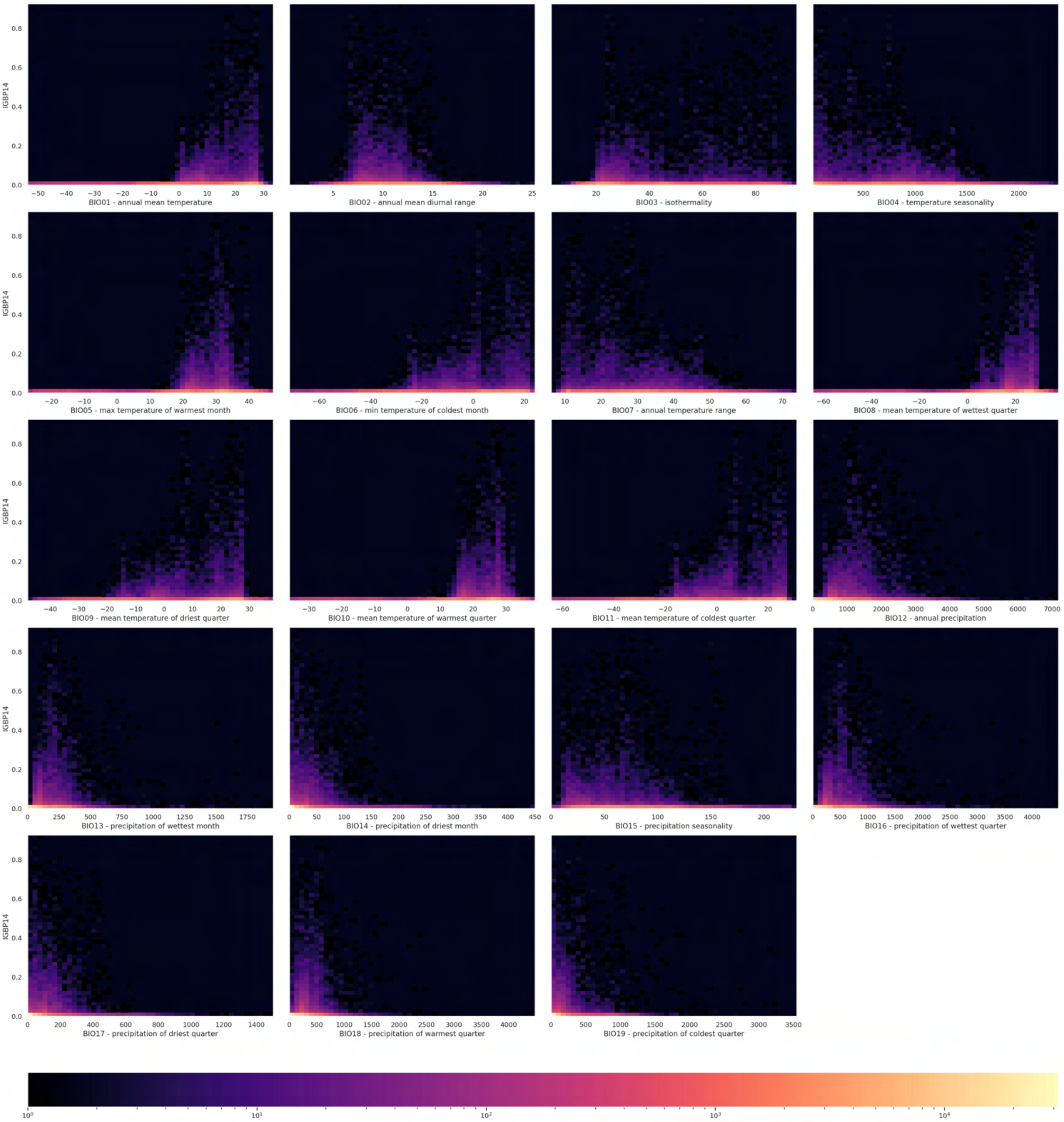


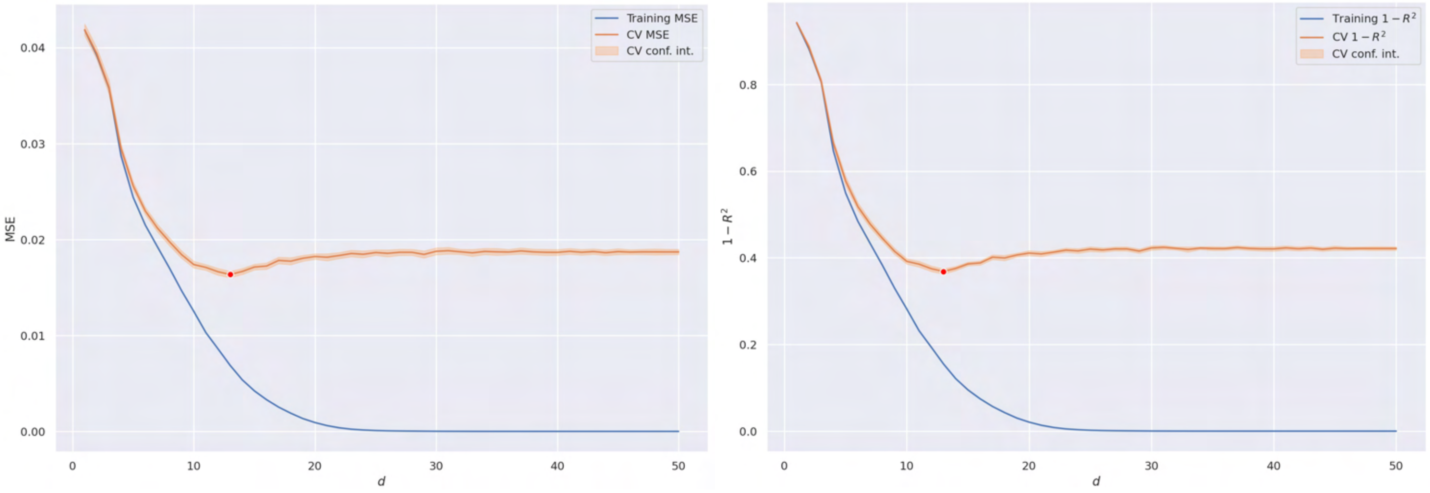
Figure A.33: Pairwise 2D histograms between the response variable IGBP14 and the bioclimatic predictors.

Appendix B

Regression analysis - Plots and tables

B.1 Decision trees

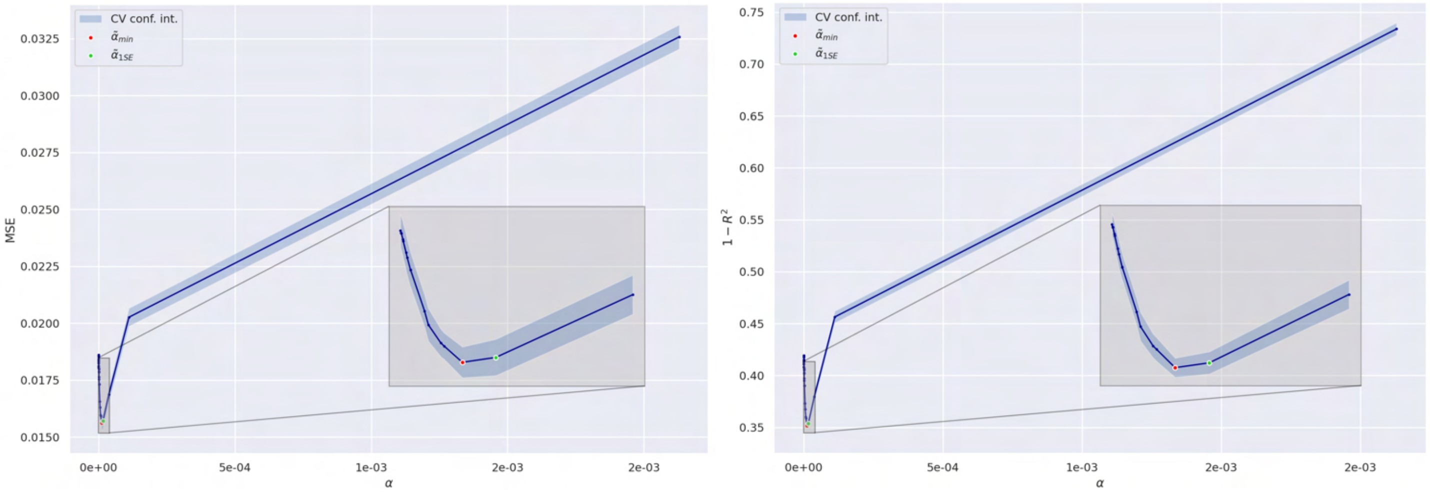
B.1.1 Tree growing and pruning



(a) MSE.

(b) $1-R^2$.

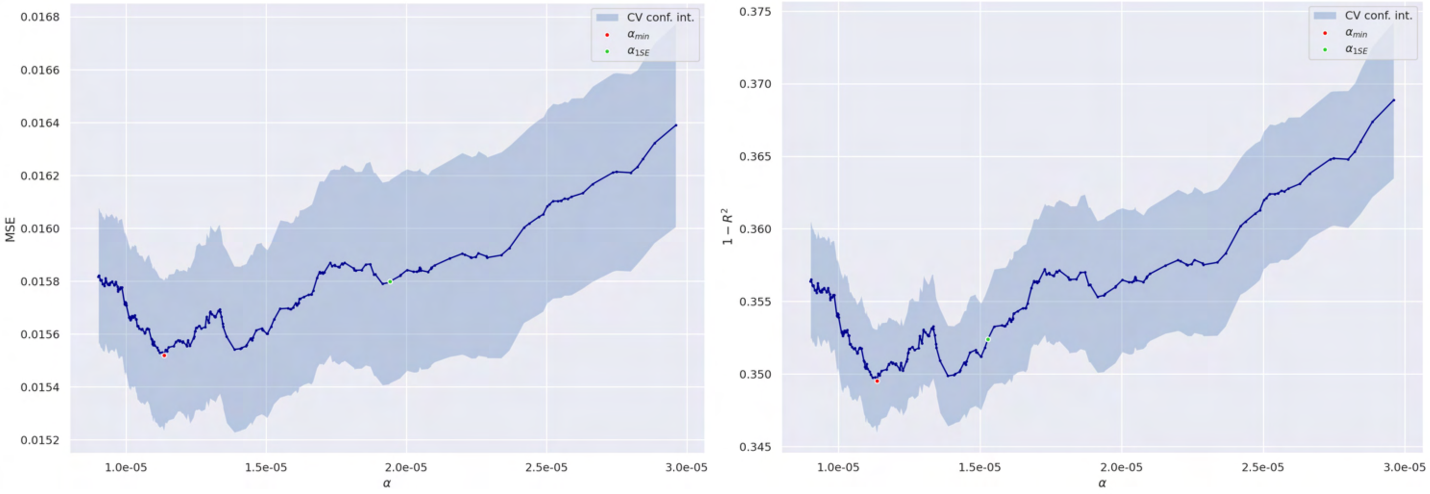
Figure B.1: Evolution of the scoring metrics as the maximum depth d of the tree increases. The generalisation estimate is represented by the orange line and the confidence interval by the light orange area all around. The training score is described by the blue line.



(a) MSE.

(b) $1-R^2$.

Figure B.2: Evolution of the generalisation scoring metrics estimates as function of α when the solution of (1.5) is computed over the randomly sampled complexity parameters during step 1. $\big(\tilde{\alpha}_{min}, \hat{\mathcal{R}}(T_{\tilde{\alpha}_{min}})\big)$ and $\big(\tilde{\alpha}_{1\text{SE}}, \hat{\mathcal{R}}(T_{\tilde{\alpha}_{1\text{SE}}})\big)$ are identified by, respectively, a red and a green dot. The confidence interval of the estimates is represented by the blue area.



(a) MSE.

(b) $1-R^2$.

Figure B.3: Evolution of the generalisation scoring metrics estimates as function of α when the solution of (1.5) is computed over the subset $A \cap [9 \cdot 10^{-6}, 3 \cdot 10^{-5}]$ during step 2. $\big(\alpha_{min}, \hat{\mathcal{R}}(T_{\alpha_{min}})\big)$ and $\big(\alpha_{1\text{SE}}, \hat{\mathcal{R}}(T_{\alpha_{1\text{SE}}})\big)$ are identified by, respectively, a red and a green dot. The confidence interval of the estimates is represented by the blue area.

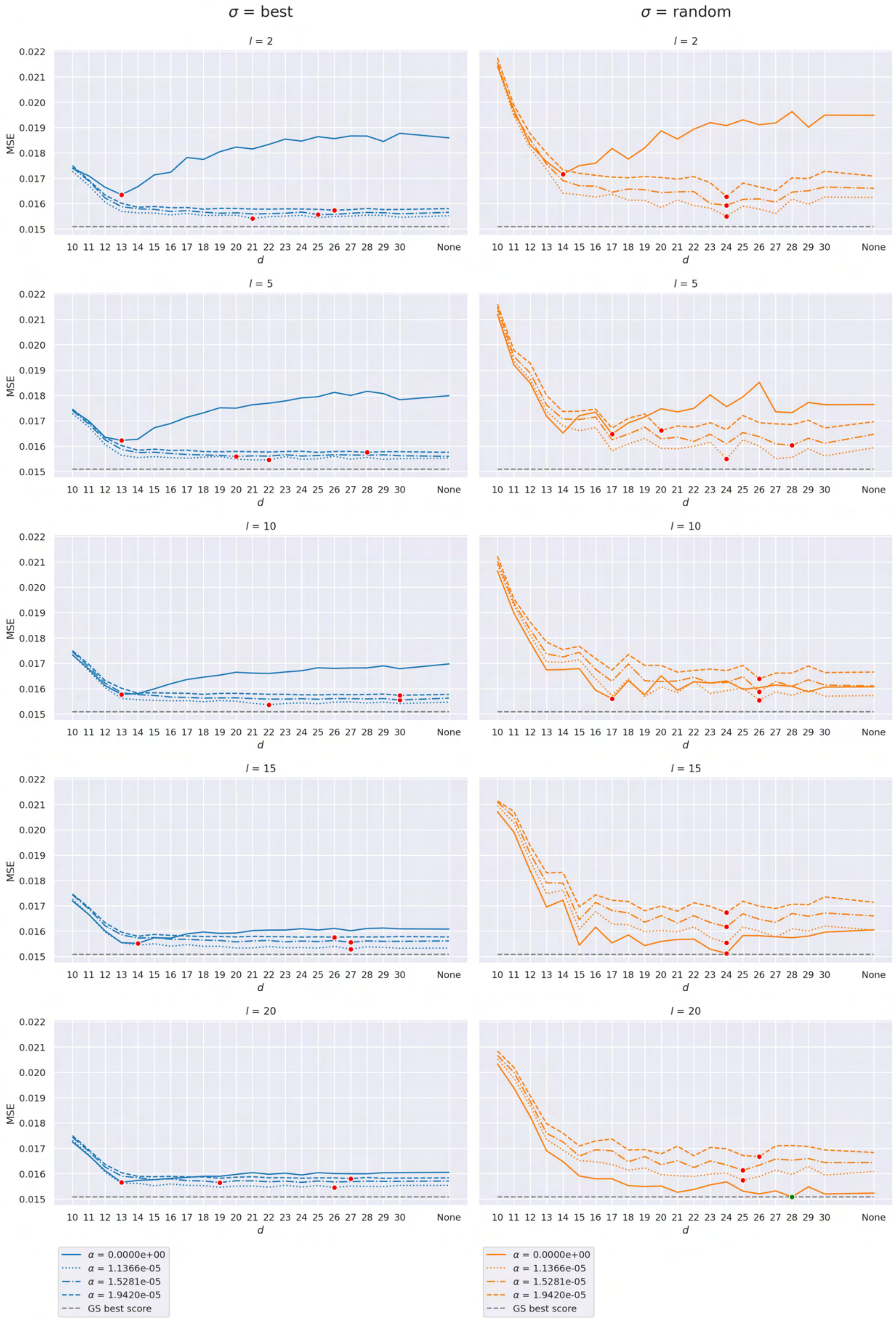


Figure B.4: Generalisation MSE estimates of the regression trees explored when combining different pruning techniques and splitting approaches of the predictor space. Red dots denote the minimum of each curve and, the green dot identifies the score of the best-performing regression tree.

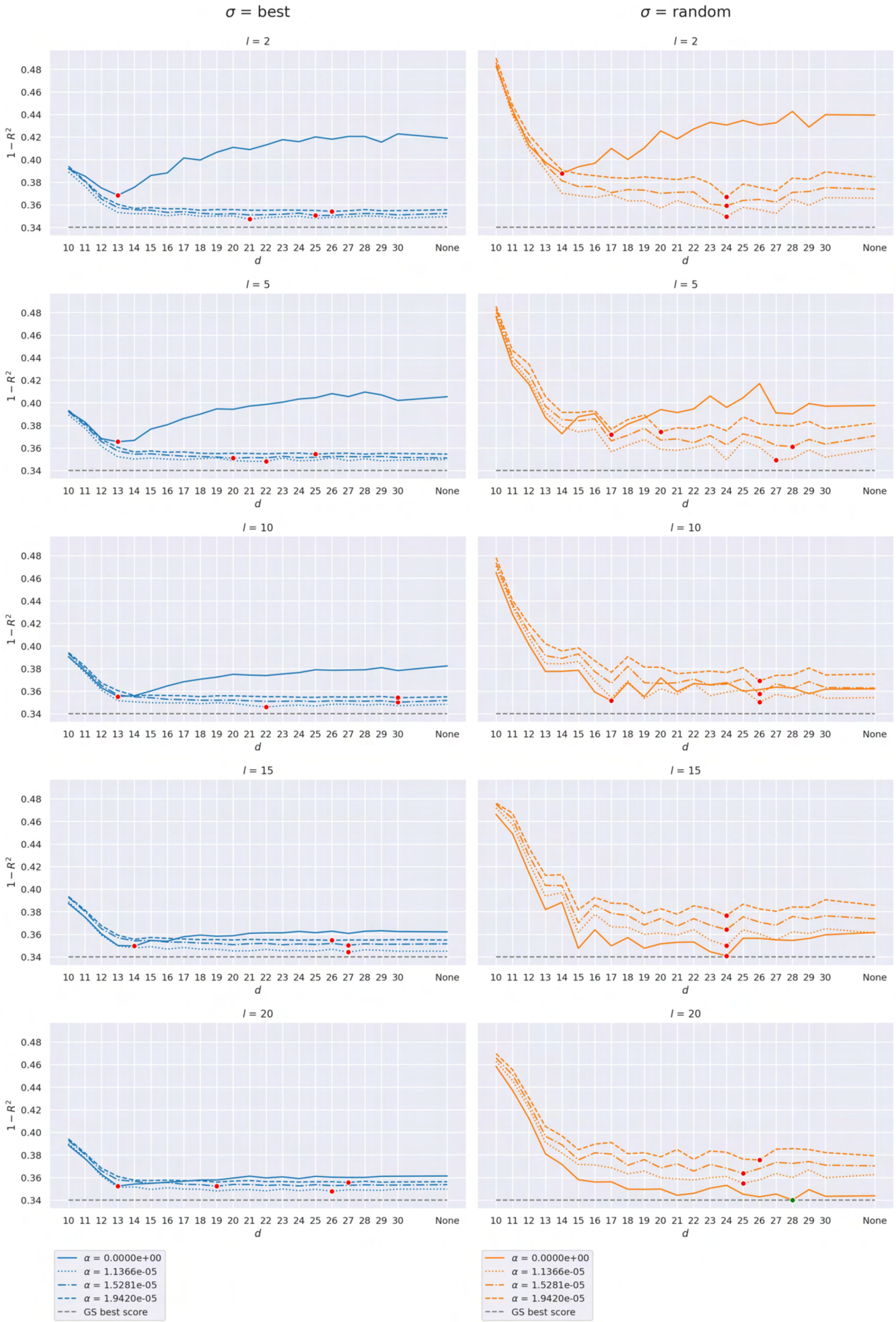


Figure B.5: Generalisation $1-R^2$ estimates of the regression trees explored when combining different pruning techniques and splitting approaches of the predictor space. Red dots denote the minimum of each curve and, the green dot identifies the score of the best-performing regression tree.

					Hyperparameters			Scores		Ranking	
					maximum depth d	min samples for a split l	CCP α (10^{-5})	MSE	$1-R^2$	MSE	$1-R^2$
Scoring	CCP	Pre-pruning	Splitter	Label							
MSE	CCP	PRE-PRUNING	BEST	$T_{1,\text{MSE}}$	27	15	1.13655	0.015288	0.344193	2	2
			RANDOM	$T_{2,\text{MSE}}$	24	5	1.13655	0.015491	0.349476	3	4
		NO PRE-PRUNING	BEST	$T_{3,\text{MSE}}$	None	2	1.13655	0.015520	0.349527	6	5
			RANDOM	$T_{4,\text{MSE}}$	None	2	1.13655	0.016238	0.365762	7	7
	NO CCP	PRE-PRUNING	BEST	$T_{5,\text{MSE}}$	14	15	0	0.015519	0.349546	5	6
			RANDOM	$T_{6,\text{MSE}}$	28	20	0	0.015085	0.339991	1	1
		NO PRE-PRUNING	BEST	$T_{7,\text{MSE}}$	None	2	0	0.018592	0.418950	8	8
			RANDOM	$T_{8,\text{MSE}}$	None	2	0	0.019482	0.439420	9	9
$1-R^2$	CCP	PRE-PRUNING	BEST	$T_{1,1-R^2}$	27	15	1.13655	0.015288	0.344193	-	-
			RANDOM	$T_{2,1-R^2}$	27	5	1.13655	0.015506	0.349198	4	3
		NO PRE-PRUNING	BEST	$T_{3,1-R^2}$	None	2	1.13655	0.015520	0.349527	-	-
			RANDOM	$T_{4,1-R^2}$	None	2	1.13655	0.016238	0.365762	-	-
	NO CCP	PRE-PRUNING	BEST	$T_{5,1-R^2}$	14	15	0	0.015519	0.349546	-	-
			RANDOM	$T_{6,1-R^2}$	28	20	0	0.015085	0.339991	-	-
		NO PRE-PRUNING	BEST	$T_{7,1-R^2}$	None	2	0	0.018592	0.418950	-	-
			RANDOM	$T_{8,1-R^2}$	None	2	0	0.019482	0.439420	-	-

- **Scoring** column specifies the scoring metric according to which the regression tree has been selected.
- CCP and **Pre-pruning** columns specifies whether CCP and pre-pruning techniques have been applied.
- **Splitter** column specifies the adopted splitting approach of the predictor space.
- **Hyperparameters** columns contain the selected values of the pruning hyperparameters.
- **Scores** columns contain the generalisation scores estimates computed via 10-fold cross-validation.
- **Ranking** columns contain the rank of each model according to the **Scores** columns; the absence of a rank value means that the corresponding model is a duplicate in the table, that is, it has been chosen by both scoring metrics.

Table B.1: Summary of the relevant regression trees obtained by combining different pruning techniques and splitting approaches of the predictor space.

					Scores					Ranking				
Scoring	CCP	Pre-pruning	Splitter	Label	MSE	$1-R^2$	MSE $(y>0)$	$1-R^2$ $(y>0)$	MSE $(y=0)$	MSE	$1-R^2$	MSE $(y>0)$	$1-R^2$ $(y>0)$	MSE $(y=0)$
MSE	CCP	PRE-PRUNING	BEST	$T_{1,\text{MSE}}$	0.014919	0.359000	0.038134	0.445343	0.000966	2	2	4	4	2
			RANDOM	$T_{2,\text{MSE}}$	0.015288	0.367868	0.038341	0.447763	0.001432	5	5	5	5	4
		NO PRE-PRUNING	BEST	$T_{3,\text{MSE}}$	0.014818	0.356578	0.037926	0.442916	0.000930	1	1	3	3	1
			RANDOM	$T_{4,\text{MSE}}$	0.015873	0.381947	0.039529	0.461639	0.001654	6	6	6	6	7
	NO CCP	PRE-PRUNING	BEST	$T_{5,\text{MSE}}$	0.014983	0.360549	0.037275	0.435308	0.001586	3	3	1	1	6
			RANDOM	$T_{6,\text{MSE}}$	0.015022	0.361474	0.037379	0.436528	0.001584	4	4	2	2	5
		NO PRE-PRUNING	BEST	$T_{7,\text{MSE}}$	0.017475	0.420501	0.043132	0.503715	0.002054	8	8	8	8	8
			RANDOM	$T_{8,\text{MSE}}$	0.019013	0.457507	0.046848	0.547110	0.002283	9	9	9	9	9
$1-R^2$	CCP	PRE-PRUNING	BEST	$T_{1,1-R^2}$	0.014919	0.359000	0.038134	0.445343	0.000966	-	-	-	-	-
			RANDOM	$T_{2,1-R^2}$	0.016743	0.402895	0.042568	0.497120	0.001222	7	7	7	7	3
		NO PRE-PRUNING	BEST	$T_{3,1-R^2}$	0.014818	0.356578	0.037926	0.442916	0.000930	-	-	-	-	-
			RANDOM	$T_{4,1-R^2}$	0.015873	0.381947	0.039529	0.461639	0.001654	-	-	-	-	-
	NO CCP	PRE-PRUNING	BEST	$T_{5,1-R^2}$	0.014983	0.360549	0.037275	0.435308	0.001586	-	-	-	-	-
			RANDOM	$T_{6,1-R^2}$	0.015022	0.361474	0.037379	0.436528	0.001584	-	-	-	-	-
		NO PRE-PRUNING	BEST	$T_{7,1-R^2}$	0.017475	0.420501	0.043132	0.503715	0.002054	-	-	-	-	-
			RANDOM	$T_{8,1-R^2}$	0.019013	0.457507	0.046848	0.547110	0.002283	-	-	-	-	-

- **Scoring** column specifies the scoring metric according to which the regression tree has been selected.
- CCP and **Pre-pruning** columns specifies whether CCP and pre-pruning techniques have been applied.
- **Splitter** column specifies the adopted splitting approach of the predictor space.
- **Scores** columns contain the generalisation scores computed on the test set. $(y > 0)$ denotes that only the positive fractions of the test response have been considered for the computation; similarly $(y = 0)$ denotes only the null fractions.
- **Ranking** columns contain the rank of each model according to the **Scores** columns; the absence of a rank value means that the corresponding model is a duplicate in the table, that is, it has been chosen by both scoring metrics.

Table B.2: Generalisation scores of the relevant regression trees computed on the test set. This table is the continuation of Table B.1.

B.1.2 Interpretation of the results

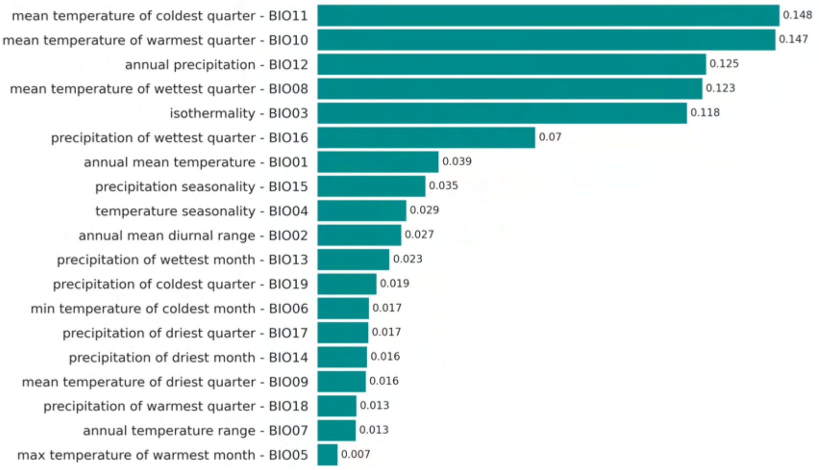
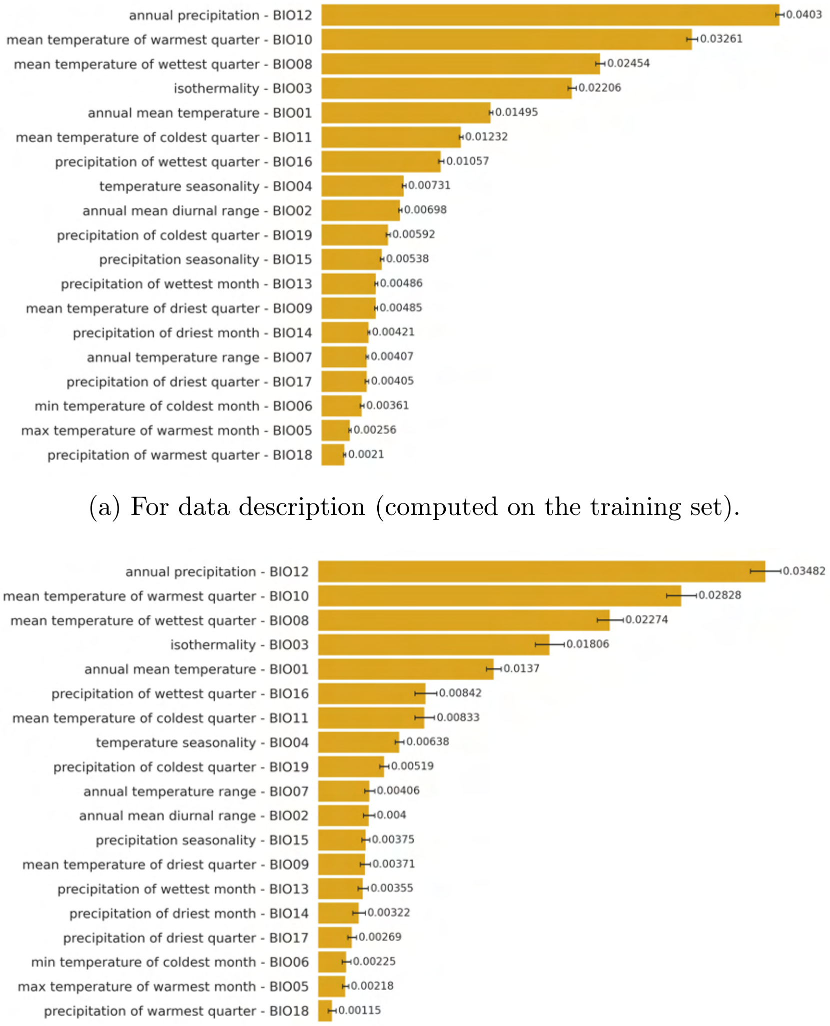


Figure B.6: Ranking of the bioclimatic predictors according to the impurity-based variable importance measure on the regression tree T.



(a) For data description (computed on the training set).

(b) For contribute in generalisation power (computed on the test set).

Figure B.7: Rankings of the bioclimatic predictors according to the permutation importance measure on the regression tree T. Importance values have been averaged over 10 random shufflings. MSE is the scoring metric used. The confidence interval of each value is displayed.

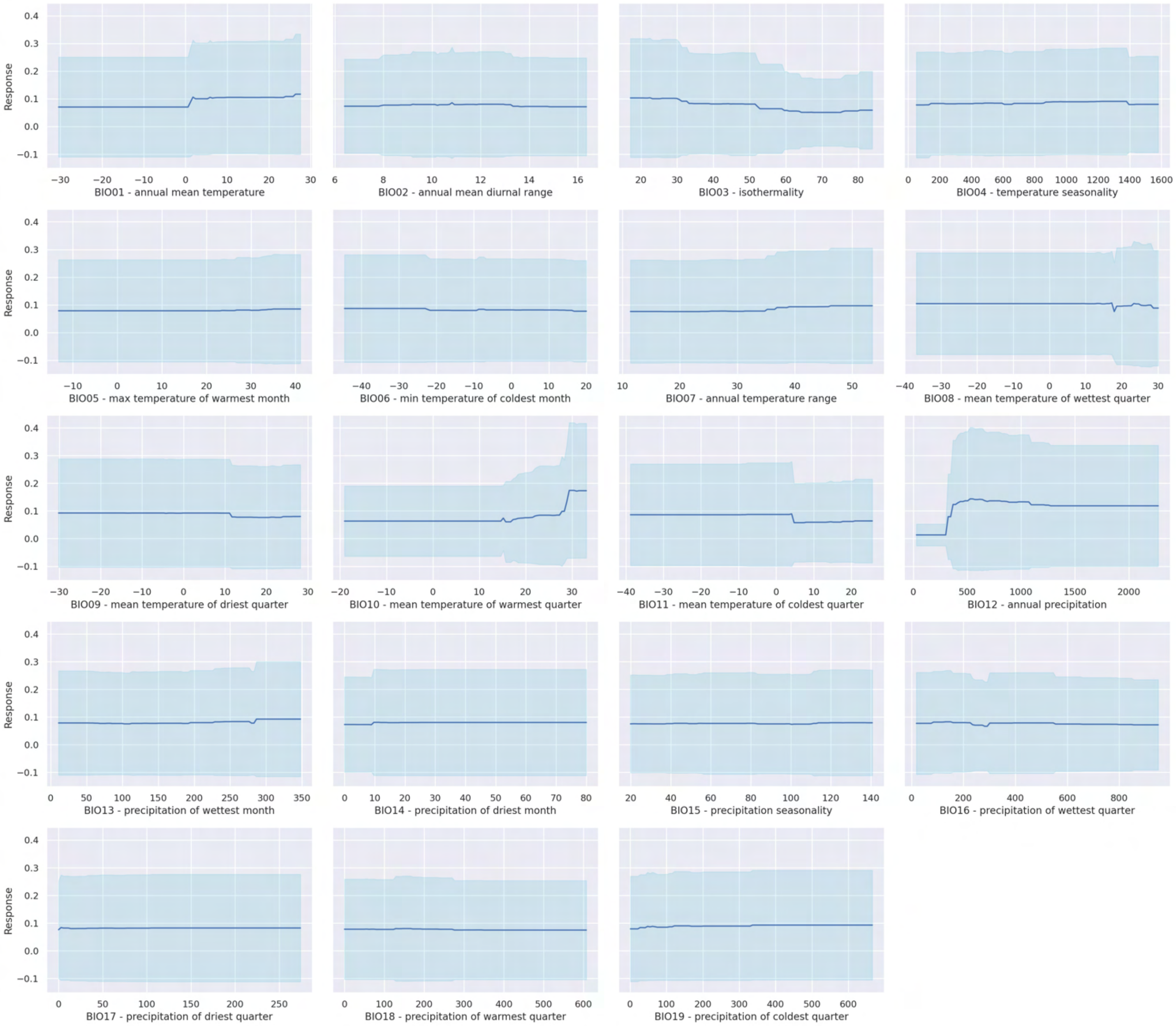
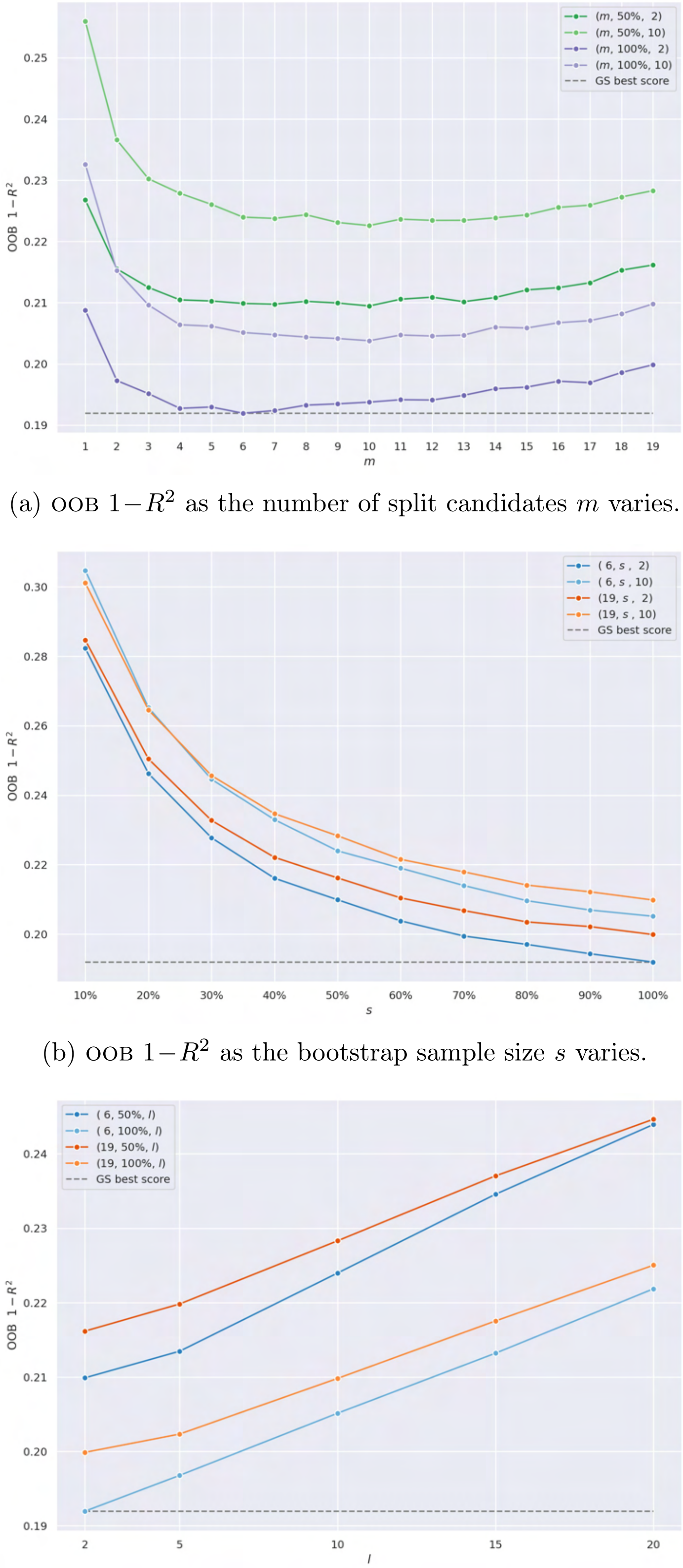


Figure B.8: Partial dependence plots of the regression tree T. Light blue areas represents the confidence intervals.

B.2 Random forests

B.2.1 Forest growing and hyperparameters tuning



(a) OOB $1-R^2$ as the number of split candidates m varies.

(b) OOB $1-R^2$ as the bootstrap sample size s varies.

(c) OOB $1-R^2$ as the minimum number of samples required to split an internal node l varies.

Figure B.9: Selected examples of the evolution of the out-of-bag $1-R^2$ of the random forest as a function of a model hyperparameter when the remaining two are fixed.

B.2.2 Interpretation of the results

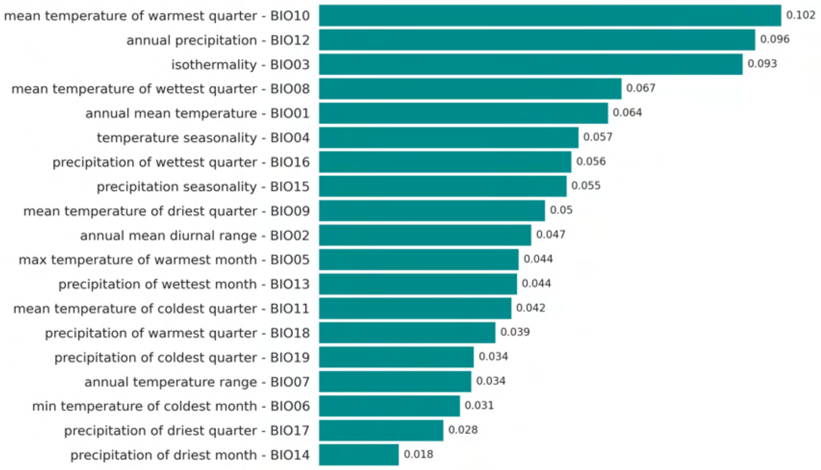
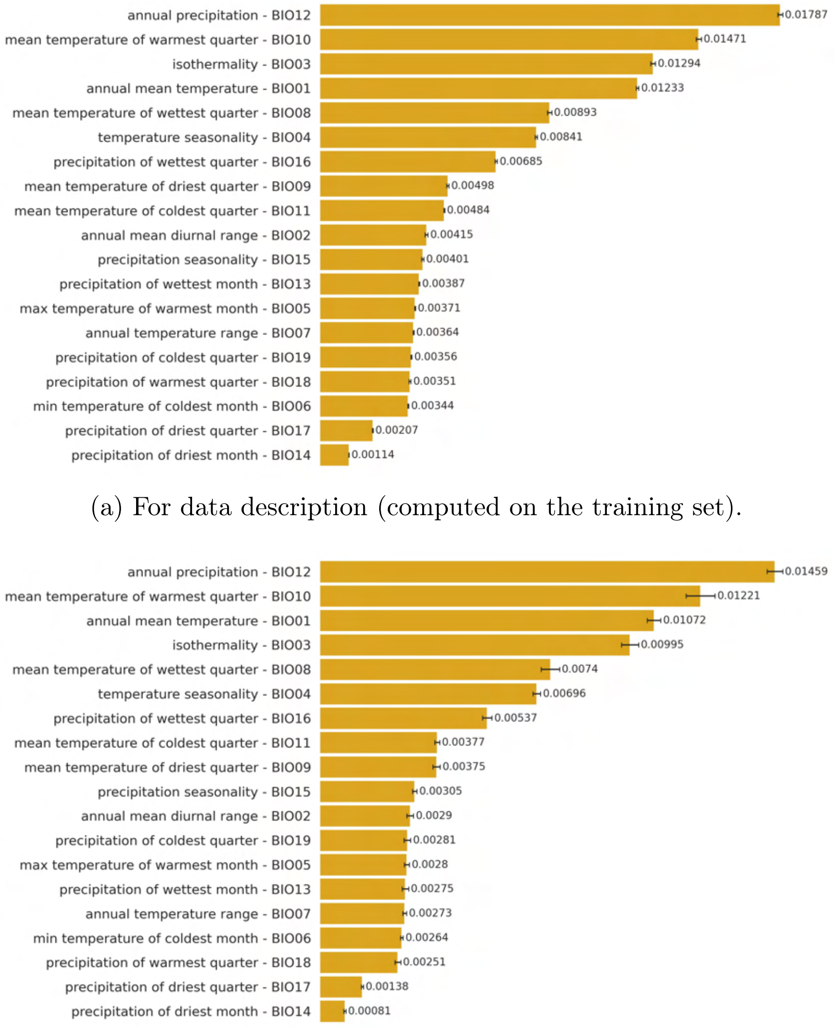


Figure B.10: Ranking of the bioclimatic predictors according to the impurity-based variable importance measure on the random forest F.



(a) For data description (computed on the training set).

(b) For contribute in generalisation power (computed on the test set).

Figure B.11: Rankings of the bioclimatic predictors according to the permutation importance measure on the random forest F. Importance values have been averaged over 10 random shufflings. MSE is the scoring metric used. The confidence interval of each value is displayed.

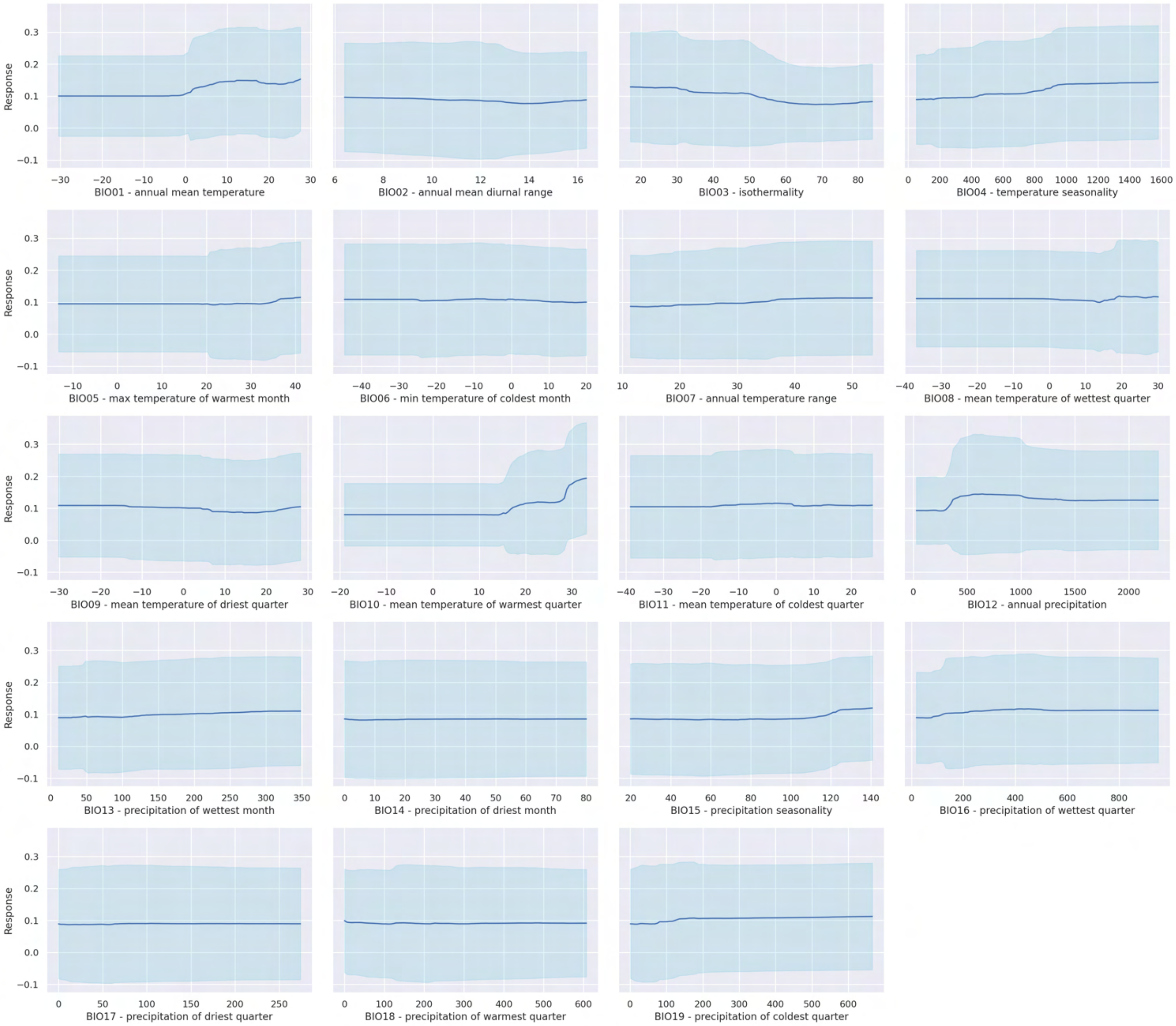


Figure B.12: Partial dependence plots of the random forest F. Light blue areas represents the confidence intervals.