scientific reports

Check for updates

OPEN

Using interpretable boosting algorithms for modeling environmental and agricultural data

Fabian Obster[1,2]✉, Christian Heumann[2], Heidi Bohle[3] & Paul Pechan[3]✉

We describe how interpretable boosting algorithms based on ridge-regularized generalized linear models can be used to analyze high-dimensional environmental data. We illustrate this by using environmental, social, human and biophysical data to predict the financial vulnerability of farmers in Chile and Tunisia against climate hazards. We show how group structures can be considered and how interactions can be found in high-dimensional datasets using a novel 2-step boosting approach. The advantages and efficacy of the proposed method are shown and discussed. Results indicate that the presence of interaction effects only improves predictive power when included in two-step boosting. The most important variable in predicting all types of vulnerabilities are natural assets. Other important variables are the type of irrigation, economic assets and the presence of crop damage of near farms.

In this work, we show how interpretable boosting algorithms can be used to predict financial vulnerabilities against multiple hazards based on environmental factors but also based on human, social, and biophysical factors as well as their interactions. For finding interactions we propose a new method based on two-step boosting, which is still interpretable and blends together with component-wise boosting. Interpretability tools like variable importance, effect sizes, and partial effects are utilized to better understand the underlying factors that may cause these vulnerabilities against climatic changes.

Model-based boosting algorithms have been used in environmental sciences for multiple purposes. For example for quantifying several soil parameters based on soil samples[1], predicting the financial wellbeing of farmers based on environmental factors[2], and predicting the number of zoo visitors based on climatic variables[3]. Also non-interpretable boosting algorithms based on classification or regression trees like Adaboost[4] have been used for environmental predictions based on environmental data because of their high predictive power. Applications include landslide susceptibility[5] and predicting the presence of juvenile sea-trouts based on environmental factors[6].

Through the proposed boosting models we want to achieve the following goals:

- *Predictive Power* The model should not only have a good fit for the analyzed data but also for unseen data from the same domain assuming a similar distribution of the variables.
- *Interpretability* We are interested in the question of which variables are associated with the outcome. But we also want to know how the associations look like. In the agronomic case, we want to derive actions to reduce vulnerability against adverse environmental changes. This is only possible if the effect of adaptive measures is known. Only if the associations are known, one can state causal hypotheses and test them with new specific experiments. We also want the effects to be modeled as simply as possible while retaining the power of the model. Linear effects should be prioritized over nonlinear effects and over interaction effects. Black-Boxes should be avoided in this case.
- *Sparsity* We consider high dimensional data sets where the number of variables p is relatively large compared to the number of observations n or even possibly higher if we consider the case with interactions. Out of the many possible variables, we want to know which ones are actually associated with the outcome and which

[1]Department of Business Administration, University of the Bundeswehr Munich, 85577 Neubiberg, Germany. [2]Department of Statistics, LMU Munich, 80539 Munich, Germany. [3]Department of Media and Communication, LMU Munich, 80539 Munich, Germany. ✉email: fabian.obster@unibw.de; paul.pechan@ifkw.lmu.de

ones are not. Therefore, the model should perform variable selection to enforce sparsity. The goal is to find the smallest subset of variables that still has high predictive power. Sparsity also increases interpretability because the scientist and stakeholders only have to look at the truly relevant variables and can disregard the unimportant ones. In the vulnerability setting this could mean that farmers focus on selected variables like the type of irrigation systems rather than not selected variables like financial adaptive measures.

- *Complexity* The model should be as complex as necessary and as simple as possible. Complexity is the characteristic that balances all previously stated points. Out of two explanations with the same predictive power the model should pick the one that is simpler. By simpler, we mean sparser, more interpretable, and without interactions. On the other hand, we do not want to neglect important complexities like non-linearity and interactions. It is important to identify if some variables are modified by others. There could also be non-hierarchical interactions, where a variable has by itself no effect on the outcome, but may have a positive effect in one subset of the data and a negative one in the other. One example could be, that in one region a high variety of crops has a positive effect on vulnerability and in another region a negative effect.
- *Group structure* The variables in the data can be clustered into groups. "Climate change experience" is one example and contains the binary variables "increasing temperature", "increasing drought", "increasing extreme weather" and "decreasing rain". The question is whether the outcome is influenced by each or only by some of the individual variables or if they act as a group due to the similarity. Group structures also increase interpretability, because it is often easier for humans to comprehend the overall effect of an abstract concept than to look at all its facets.

There are many approaches to deal with each of the above specifications. For example, sparsity can be achieved through Lasso Regression[7] or boosted Lasso[8], predictiveness can be achieved through a big variety of models and group structures can be incorporated with the sparse group lasso[9].

In this work we focus on how these goals can be met using boosting algorithms, namely componentwise boosting (mb), componentwise boosting with interactions (mb int), sparse group boosting (sgb), and two-step boosting for interactions (2-boost). We compare their predictive power, effect sizes, and the relative importance of variables/groups. In the following, we describe the used methods for the analysis and discuss how they help to achieve the stated goals using modifications of the generic boosting algorithm.

Methods

Introduction of the data. Randomly selected cherry and peach farmers in the selected regions of Tunisia and Chile. In order to be selected for the survey, farmers had to own the farm, manage and work on the farm and derive the majority of their income from their farming activities. A total of 801 face-to-face interviews were subsequently conducted with farmers who fulfilled the selection criteria—401 peach farmers in Tunisia (201 in Mornag and 200 in Regueb regions) and 400 cherry farmers in Chile (200 in Rengo and 200 in Chillán regions). Mornag, Tunisia (longitude: 10.28805, latitude: 36.68529, altitude: 110 m), hereafter referred to as Northern Tunisia, is located approximately 20 km east of the capital Tunis. Regueb (longitude: 9.78654, latitude: 34.85932; altitude: 230 m), Tunisia, hereafter referred to as Central Tunisia, is located approximately 230 km south of Tunis. Rengo (longitude: −70.86744, latitude: −34.40237, altitude: 570 m), Chile, hereafter referred to as Central Chile, is located approximately 110 km south of Santiago de Chile. Chillán (longitude: −72.10233, latitude: −36.60626, altitude: 120–150 m), Chile, hereafter referred to as Southern Chile, is located approximately 380 km south of Santiago de Chile. The approximately one-hour-long interviews were carried out with farmers directly on their farms. The interviews were carried out after harvest completion in the fall of 2018 by Elka Consulting in Tunisia and in the spring 2019 by Qualitas AgroConsultores in Chile. All methods were carried out in accordance with relevant guidelines and regulations. Informed consent for the data collection was provided by the survey participants. No personality-identifiable data was collected, assuring full anonymity. Department of Communication and Media Research, University of Munich had been consulted about the participation of human subjects in the survey research. Guidance was sought from our institute about the survey implementation and data use that included participation of human subjects. Experimental protocol was approved by University of Munich. A descriptive description of the data[10] and further mixed methods analysis on vulnerability[11] with similar data was performed.

Code availability. The R code of the analysis can be found at https://github.com/FabianObster/boostingEcology.

Independent variables. The analyzed variables can be clustered into groups, including

- Climate experience group (Increasing temperature, decreasing rain, increasing drought, increasing extreme weather)
- Natural asset group (geographical regions)
- Social asset group (reliance on/use of information, trust in information sources, community, science or religion)
- Human asset group (age, gender, education)
- Biophysical asset group (farm size, water management systems used on the farm, diversity of crops used, adaptive measures)
- Economic asset group (farm debt, farm performance, reliance on orchard income)
- Goals group (Keep tradition alive, work independently)

- Harm group (Climate threatens farm, Optimism)
- Spatial group (Crop damage near farms, Crop damage of farms in Country)

An overview of all variables and the belonging groups can be found in Tables 4 and 5. There, also the number of farmers in each category can be found (Tables 1, 2).

Outcome variables. The outcome variables measure financial vulnerability against the 5 climate hazards, increasing winter temperatures, increasing summer temperatures, decreasing rainfall, increasing drought, and increasing extreme weather based on self-assessment of the farmers. For each of the hazards, a binary variable indicating if a farmer is vulnerable to the hazard is defined as the outcome variable. The main category includes farmers, who are not financially vulnerable and the reference category includes farmers who are financially vulnerable. The number of farmers in each category can be found in Table 3.

Interaction variables. 22 variables were used as variables that may have an interaction effect with the other variables on the outcome. The interaction variables include regions as well as socio-demografic variables amongst others and are indicated in bold in Tables 4 and 5. Together with all other variables, this yields 1366 interaction terms and over 4000 possible model parameters to estimate. Since there are 801 farmers in the data,

AUC sgb	AUC mb	AUC 2-boost	AUC mb int	Outcome vulnerability
0.656	0.619	0.608	0.587	Summer temperature
0.707	0.708	0.713	0.705	Winter temperature
0.852	0.852	0.852	0.500	Decreasing rainfall
0.768	0.768	0.768	0.500	Drought
0.776	0.778	0.783	0.773	Extreme weather

Table 1. AUC values for the sparse group boosting (sgb), component-wise boosting (mb), parallel boosting with interaction (mb int) and two-step boosting with interactions (2-boost) for all vulnerability outcomes evaluated on the test data.

Model	Number selected interaction terms	1-Sparsity in percent	Outcome vulnerability
mb int	13	0.95	Summer temperature
2-boost	0	0	Summer temperature
mb int	38	2.78	Winter temperature
2-boost	12	0.88	Winter temperature
mb int	48	3.51	Decreasing rainfall
2-boost	1	0.07	Decreasing rainfall
mb int	27	1.98	Drought
2-boost	16	1.17	Drought
mb int	32	2.34	Extreme weather
2-boost	10	0.73	Extreme weather

Table 2. Comparison of the number of selected interaction terms based on two-step estimation (2-boost) and the parallel estimation (mb int) and the percentage of selected interactions (1-Sparsity) of the 1366 interaction terms.

Variable	Category	n
No summer temperature vulnerability	Yes	358
No winter temperature vulnerability	Yes	579
No decreasing rainfall vulnerability	Yes	451
No drought vulnerability	Yes	492
No extreme weather vulnerability	Yes	453

Table 3. Overview over outcome variables. Financial vulnerability against climate hazards. The "n" column gives the number of farmers who are not financially vulnerable to each of the hazards.

Variable name	Category	n	Group name
Agronomic measures	Yes	647	Biophysical asset group
Economic measures	Yes	464	Biophysical asset group
Use of river irrigation	Yes	138	Biophysical asset group
Use of well irrigation	Yes	231	Biophysical asset group
Farm size	Yes	283	Biophysical asset group
Orchard size	Yes	318	Biophysical asset group
More than one variety grown	Yes	508	Biophysical asset group
Other products	Yes	571	Biophysical asset group
Technological measures	Yes	721	Biophysical asset group
Increasing temperature	Yes	629	Climate experience group
Decreasing rainfall	Yes	659	Climate experience group
Increasing drought	Yes	671	Climate experience group
Increasing extreme weather	Yes	542	Climate experience group
Income invested > 80 Percent	Yes	137	Economic asset group
Income invested <40 percent	Yes	358	Economic asset group
High financial wellbeing	Yes	346	Economic asset group
Low financial wellbeing	Yes	148	Economic asset group
Farm debt load	High	96	Economic asset group
Dependent on farm	Yes	528	Economic asset group
Family farm engagement	Yes	203	Economic asset group
Adaptive measures efficacy	High	490	Efficacy group
Work independent	Yes	635	Goals group
Keep tradition alive	Yes	460	Goals group
Provide good living environment	Yes	466	Goals group
Be in profitable business	Yes	320	Goals group
Climate change is harmful	Yes	258	Harm group
High optimism	Yes	446	Harm group
High certainty	Yes	470	Harm group
Climate threatens farm	Yes	629	Harm group
Climate risks > benefits	Yes	648	Harm group
Climate change acceptance	Yes	676	Human asset group
Human cause climate change	Yes	685	Human asset group
Climate extremes	Yes	755	Human asset group
Age > 50	Yes	438	Human asset group
Gender	F	121	Human asset group
Gender	M	680	Human asset group
Education	Yes	459	Human asset group
Years of farm possession	Yes	577	Human asset group
Prior ownership (family)	Yes	399	Human asset group
Years of farm managing	Yes	437	Human asset group
Natural assets	CentralChile	200	Natural asset group
Natural assets	CentralTunisia	200	Natural asset group
Natural assets	NorthernTunisia	201	Natural asset group
Natural assets	SouthernChile	200	Natural asset group
Adapive measures near farms	1	424	Norms group
Adapive measures near farms	2	151	Norms group
Adapive measures near farms	3	226	Norms group
High optimism	Yes	446	Perception group

Table 4. Overview over variables and groups. The 22 variables used as interaction variables (potential moderators) are bold. The number of observations within each category of each variable is in the n column. For binary variables, only one category is presented and the remaining category is "no" if the shown category is "yes" and "low" if the shown category is "high".

finding interactions results in a "p > n" problem, where the number of variables in the design matrix is greater than the number of observations.

Variable	Category	n	Group
Use of newspapers	Yes	95	Social asset group
Use of farming journals	Yes	161	Social asset group
Use of TV	Yes	415	Social asset group
Use of radio	Yes	219	Social asset group
Use of internet	Yes	319	Social asset group
Use of extension workers	Yes	346	Social asset group
Use of government workers	Yes	166	Social asset group
Use of neighbours	Yes	313	Social asset group
Use of industry	Yes	192	Social asset group
Use of farm associations	Yes	97	Social asset group
Trust in newspapers	Yes	174	Social asset group
Trust in farming journals	Yes	291	Social asset group
Trust in TV	Yes	329	Social asset group
Trust in radio	Yes	241	Social asset group
Trust in internet	Yes	319	Social asset group
Trust in extension workers	Yes	433	Social asset group
Trust in government workers	Yes	268	Social asset group
Trust in neighbours	Yes	319	Social asset group
Trust in industry	Yes	215	Social asset group
Trust in farm associations	Yes	184	Social asset group
Trust in government institutions	Yes	312	Social asset group
Trust in other farmers	Yes	351	Social asset group
Trust in religion	Yes	317	Social asset group
Trust in fate	Yes	360	Social asset group
Crop damage near farms	Yes	643	Spatial group
Crop damage farms in Country	Yes	673	Spatial group
Climate change occurs	Yes	592	Spatial group

Table 5. Overview over variables and groups continued. The 22 variables used as interaction variables (potential moderators) are bold. The number of observations within each category of each variable is in the n column. For binary variables, only one category is presented and the remaining category is "no" if the shown category is "yes" and "low" if the shown category is "high".

General setup, model formulation and evaluation. All analyses were performed with R[12] and the boosting models were fitted with the package "mboost"[13].

Since all outcome variables are binary, we use the Ridge penalized negative log-likelihood of the binomial distribution as a loss function and a logit link, which yields

$$h(\beta, X_i) = P(y_i == 1) = \frac{1}{1 + \exp(-X_i^T \beta)},$$

$$l(y, h) = -\left[\sum_{i=1}^{n} y_i \log(h(\beta, X_i)) + (1 - y_i) \log(1 - h(\beta, X_i))\right] + \lambda \|\beta\|_2^2.$$

Before performing any analysis the data was split into 70 percent training data and 30 percent test data, which was only used for the final evaluation. Variable importance and partial effects were computed using the whole data after the predictive analysis. Model evaluation was based on the area under the receiver operator curve (ROC) and computed using the test data. The area under the ROC (AUC) takes both the true positive and the false positive rate into account by considering all possible thresholds of predicted probabilities computed by a prediction model. While the AUC is often used for discriminatory performance, it is also limited by not assessing calibration and in the presence of strong class imbalances.

In the analysis, we use multiple boosting models for multiple purposes. All boosting models were fitted with the R package "mboost"[14]. For early stopping, the stopping parameter was determined using a 10-fold cross-validation performed at every boosting step. The first and most simple one is component-wise model-based boosting (mb) with ridge-regularized linear effects of all variables, such that the degrees of freedom are all equal to one. This model allows us to perform variable selection and allows for a comparison between all variables regarding their relative importance. For the second model, we used sparse group boosting with a mixing parameter $\alpha = 0.5$, which balances group selection and individual variable selection. This way it is possible to see if variables are important on their own for the outcome, or if they rather act as groups of variables.

To find interactions in the data we use two approaches. The first one is the standard approach by defining linear effects and interaction effects at the same time in each iteration. Then the model can decide weather it selects the main effects or the interaction effects. In the second approach we use a two-stage boosting model. As the first step we use the already fitted mb model, which only uses individual linear base-learners. The second step uses solely interactions. This way linear base-learners are prioritized over interaction base-learners since they are fitted first.

This remaining part of the methods section is more technical and may be skipped by the application-oriented reader.

Generic boosting algorithm. We will start with the general formulation of the boosting algorithm which can also be described as a functional gradient descent algorithm. The goal is to find a function f^* that minimizes some Loss function $l(y, f)$. Here, we only consider differentiable convex loss functions. The loss function has two arguments. The first argument $y \in \{1, ..., n\}$ is the outcome variable with n observations. The second argument f is a real-valued function $f : \mathbb{R}^{n\times p} \mapsto \mathbb{R}$, which is a function of the data $X \in \mathbb{R}^{n\times p}$.

Another way of fitting sparse regression models is through the method of boosting. The fitting strategy is to consecutively improve a given model by adding a base-learner to it. Throughout this article, we refer to a base-learner as a subset of columns of the design matrix associated with a real-valued function. To enforce sparsity, each base-learner only considers a subset of the variables at each step[15]. In the case of component-wise $\mathcal{L}^2$ boosting, each variable will be a base-learner. In the case of a one-dimensional B-Spline, a base-learner is the design matrix representing the basis functions of the B-Spline. The goal of boosting in general is to find a real valued function that minimizes a typically differentiable and convex loss function $l(\cdot, \cdot)$. Here we will consider the negative log-likelihood as a loss function to estimate f^* as

$$f^*(\cdot) = \arg\min_{f(\cdot)} \mathbb{E}[l(y, f)].$$

General functional gradient descent Algorithm[16].

1. Define base-learners of the structure $h : \mathbb{R}^{n\times p} \to \mathbb{R}$
2. Initialize $m = 0$ and $\widehat{f}^{(0)} \equiv 0$ or $\widehat{f}^{(0)} \equiv \overline{y}$
3. Set $m = m + 1$ and compute the negative gradient $\frac{\partial}{\partial f} l(y, f)$ and evaluate it at $\widehat{f}^{[m-1]}$. Doing this yields the pseudo-residuals $u_1, ..., u_n$ with

$$u_i^{[m]} = \frac{\partial}{\partial f} l(y_i, f)|_{f=\widehat{f}^{[m-1]}},$$

   for all $i = 1, ..., n$
4. Fit the base-learner h with the response $(u_1^{[m]}, ..., u_n^{[m]})$ to the data. This yields $\widehat{h}^{[m]}$, which is an approximation of the negative gradient
5. Update

$$\widehat{f}^{[m]} = \widehat{f}^{[m-1]} + \eta \cdot \widehat{h}^{[m]}$$

   here η can be seen as learning rate with $\eta \in]0, 1[$
6. Repeat steps 3,4 and 5 until $m = M$

Boosted ridge regression. The loss function $l(\cdot, \cdot)$ can be set to any function. In the case of interpretable boosting, the negative log-likelihood is a reasonable choice. The log-likelihood can also be modified using a Ridge penalty. By introducing the hyperparameter $\lambda > 0$, one can modify the loss function l. Let h be a function of a parameter vector $\beta \in \mathbb{R}^p$ and the design matrix $X \in \mathbb{R}^{n\times p}$, then

$$l_{\text{Rige}}(u, h) = l(u, h) + \lambda \|\beta\|_2^2$$

is the Ridge penalized loss function. By increasing λ, the parameter vector β can be shrunken towards zero. Closely related to λ are the degrees of freedom. Let S be the approximated generalized ridge hat matrix as in Proposition 3 in[17]. We remark that in the special case of ordinary least squares ridge regression we have $S = X(X^TX + \lambda I)^{-1}X^T$. Generally, the degrees of freedom can be defined as

$$\text{df}(\lambda)) = \text{tr}(2\text{S} - (\text{S})^{\text{T}}\text{S})).$$

It is recommended to set the regularization parameter for each base-learner, such that the degrees of freedom are equal for all base-learners. Thus, the regularization parameter enables using complex base-learners like polynomial effects and simple effects like linear effects at the same time. Since the more complex base-learners are regularized more than the simpler ones it is possible to prioritize simple and more interpretable base-learners over complex ones, introducing an inductive bias towards interpretability, as we demanded in the problem statement.

Component-wise and group component-wise boosting. In step 4 of the general functional gradient descent algorithm, the function h is applied. Instead of just one function, one can also use a set of R functions

$\{(h_r)_{r\leq R}\}$. Then the update in step 5 is only performed with the function that has the lowest loss function applied to the data, meaning $r^* = \arg\min_{r\leq R}\mathbb{E}[l(u, h_r)]$. In the case of component-wise boosting, for each variable in the dataset, a function is used that is only a function of this variable and not the others. This way in each step only one variable is selected. Then through early-stopping, or setting M relatively smaller compared to the number of variables in the dataset, a sparse overall model can be fitted. This addresses the sparsity requirement in the problem statement section. In the case of grouped variables, one can also define base-learners as groups of variables, which are a function of only the variables belonging to one group. These could be all item variables that belong to a specific construct like in sociological data[18] or all climate change-related variables in agricultural and environmental data[2]. This allows group variable selection, where only a subset of groups is selected, yielding a group/construct-centric analysis rather than on an individual-variable basis. This way, the group structure can be taken into account.

Sparse group boosting. It is also possible to use individual and group-based base-learners at the same time. Then at each step, either an individual variable or a group of variables is selected. Using a similar idea as in the sparse group lasso[9], the sparse group boosting can be defined[19]. We do this again by modifying the degrees of freedom. Each variable will get its own base-learner, and each group of variables will get one base-learner, containing all variables of the group. Let G be the number of groups and p_g the number of variables in group g. Then, for the degrees of freedom of an individual base-learner $x_j^{(g)} \in \mathbb{R}^{n\times 1}$ we will use

$$\mathrm{df}\left(\lambda_j^{(g)}\right) = \frac{1}{p_g}\cdot\alpha.$$

For the group base-learner we use

$$\mathrm{df}\left(\tilde{\lambda}^{(g)}\right) = \frac{1}{p_g}\cdot(1-\alpha).$$

The mixing parameter $\alpha \in [0, 1]$ allows to change the prioritization of groups versus individual variables in the selection process. If $\mathrm{df}(\lambda) = 0$ means $\lambda \to \infty$, $\alpha = 1$ yields component-wise boosting, and $\alpha = 0$ yields group boosting.

Two-step boosting. In the generic boosting algorithm, a single set of functions is applied sequentially to the data. While there is variable selection within the set of functions, the set itself does not change during the boosting procedure. We describe a modification of the general that allows more flexibility, namely a two-step version of boosting. A similar idea of two-step boosting, called hierarchical boosting has been used in genetic research[20] in transfer learning[21], and also deep learning applications[22]. In most cases, hierarchical boosting is used, if the outcome variable consists of a hierarchical class structure[23]. In contrast to the data analyzed in the literature, the data we analyze here does not contain hierarchical class structures. Hence, we do not use hierarchical boosting as in most cases presented in the literature, but for hierarchical and non-hierarchical interaction detection.

We formulate and generalize the two-step boosting. Let K be the number of steps and for every step $k \leq K$ let H_k be the set of base-learners.

K-step boosting algorithm.

1. Set K as the number of steps
2. For every step $k \leq K$ define the set of base-learners H_k to be used and set M_k to the number of boosting iterations
3. Initialize $m_0 = 0$ and $\widehat{f}^{(0)} \equiv 0$ or $\widehat{f}^{(0)}$
4. For $k \leq K$ repeat:
5. For $m_k \leq M_k$ perform steps 2-6 of the general boosting algorithm
6. Set Initialization $m_k = 0$ and $u^{[0]} = u^{[M_{k-1}]}$

One may ask why it is necessary to run multiple boosting algorithms after each other if it is possible to just use more base-learners in parallel in the original method. Previous research has shown high predictive powers in some combinations of steps. However, as described in the problem statement for us predictive power is only one part of the requirements and not necessarily desirable if it comes at the cost of interpretability and understanding of the data. Also, the sequential nature of the algorithm reduces computational improvements through parallelization, as not all base-learners can be fitted in the same boosting iteration in parallel. The k-step boosting algorithm can also be seen as a special case of the general boosting algorithm, where the base-learners themselves are boosting algorithms.

Variable importance. For each of the previously described boosting methods, it is possible to compute a variable importance measure. In each step, the log-likelihood is computed, which means that one can compute the reduction of log-likelihood attributed to the base-learner being selected in the step. After the fitting for each base-learner the total reduction of likelihood can be computed. This way, one can compute the percentage of reduction in the negative log-likelihood attributed to each base-learner, regardless of the type of base-learner. The variable importance allows us to compare the relative importance of variables compared to each other and

is distinct from the concept of significance or p values which tests a hypothesis of a parameter not being zero based on a set of assumptions. Hence a variable can be important in boosting while not being significant base on classical regression and vice versa.

Partial effect and effect sizes. For boosted generalized linear models, partial effects can be computed[13]. Similar to classical logistic regression, odds ratios for all base-learners can be computed by first summing up all linear predictors for one base-learner. These odds ratios can then be interpreted similarly to effect sizes in logistic regression. Based on the linear predictor one can also compute predicted probabilities for categories of variables if all other base-learners are set to average values. This way partial effects can be plotted, both for individual variable base-learners and for interaction-base-learners. Thus model-based boosting models are by themselves interpretable compared to other machine learning models where only post-hoc explanations can be derived. One can also track which variable was selected in each boosting iteration and thus understand how the model works internally.

Applications

Predictability. Referring to Table 1 and Fig. 1 we can see that the AUC values are comparable between the boosting models except for the interaction model with parallel estimation. Averaging the AUC values across the five vulnerabilities, sgb yields 0.752, mb and 2-boost yield 0.745, and mb int 0.613. For precipitation and drought vulnerability, the parallel estimation of interactions resulted in no variables being selected and therefore the AUC takes a value of 0.5. In 2-boost, also no variables were selected in the second estimation resulting in the same model as mb, which had the highest AUC for these outcome vulnerabilities. For summer temperature vulnerability, sgb had the highest AUC, and for winter temperature and extreme weather 2-boost had the highest AUC. Comparing the predictability of the individual vulnerabilities with each other, we see, that vulnerability against decreasing rainfall can be predicted better with the given variables, followed by vulnerability against increasing extreme weather, decreasing drought, increasing winter temperature, and summer temperature.

Importance of individual variables and groups. Comparing the variable importance of the sparse group boosting (sgb) and componentwise boosting (mb) in Fig. 2, it becomes apparent, that while there is some overlap, also some variables differ. The single most important variable for all outcomes is "Natural assets" indicating the four regions of the farm. However, the relative importance of the natural assets is higher for sgb than for mb for all five vulnerabilities. Groups seem to be more important in explaining increasing temperature vulnerability than the other vulnerabilities, as the economic asset group is the second most important variable for summer temperature vulnerability and the goals group is the second most important variable for winter temperature vulnerability. The spatial group is the third most important variable for decreasing rainfall vulnerability but the relative importance is minor compared to the most important variable.

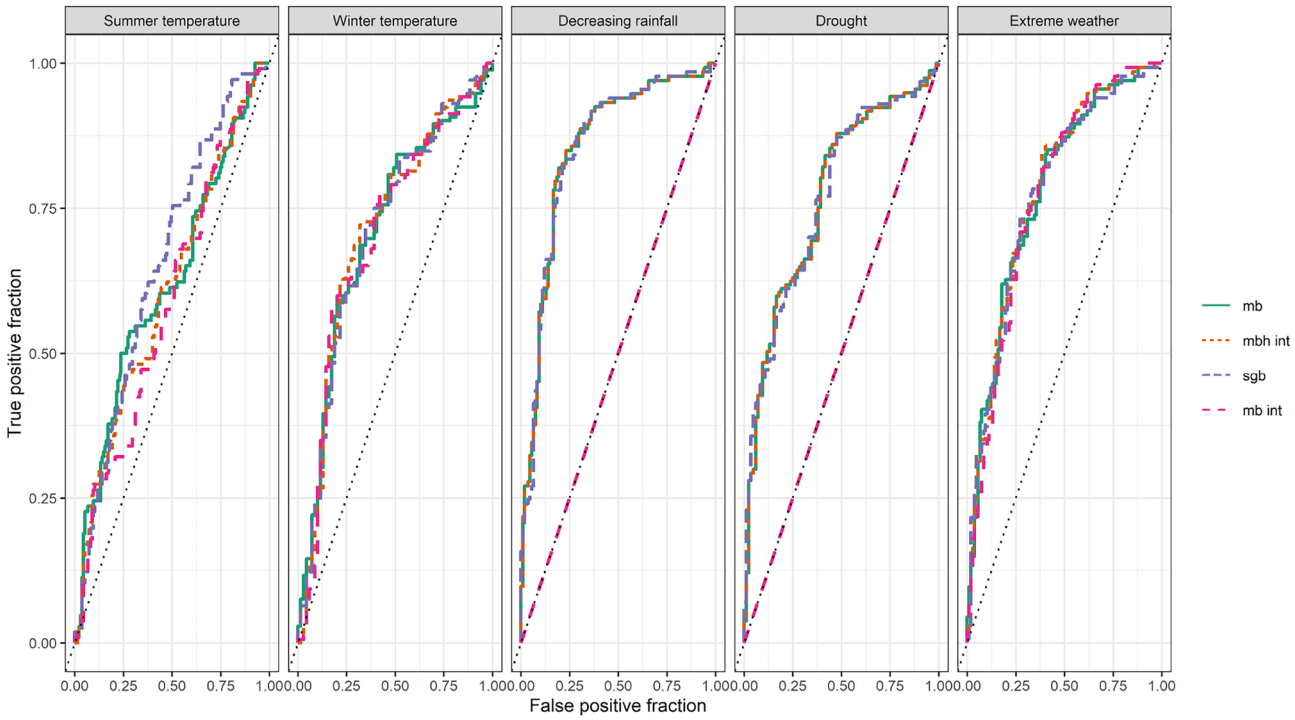

Figure 1. ROC-curves for the sparse group boosting (sgb), component-wise boosting (mb), parallel boosting with interaction (mb int) and two-step boosting with interactions (2-boost) for all vulnerability outcomes evaluated on the test data.

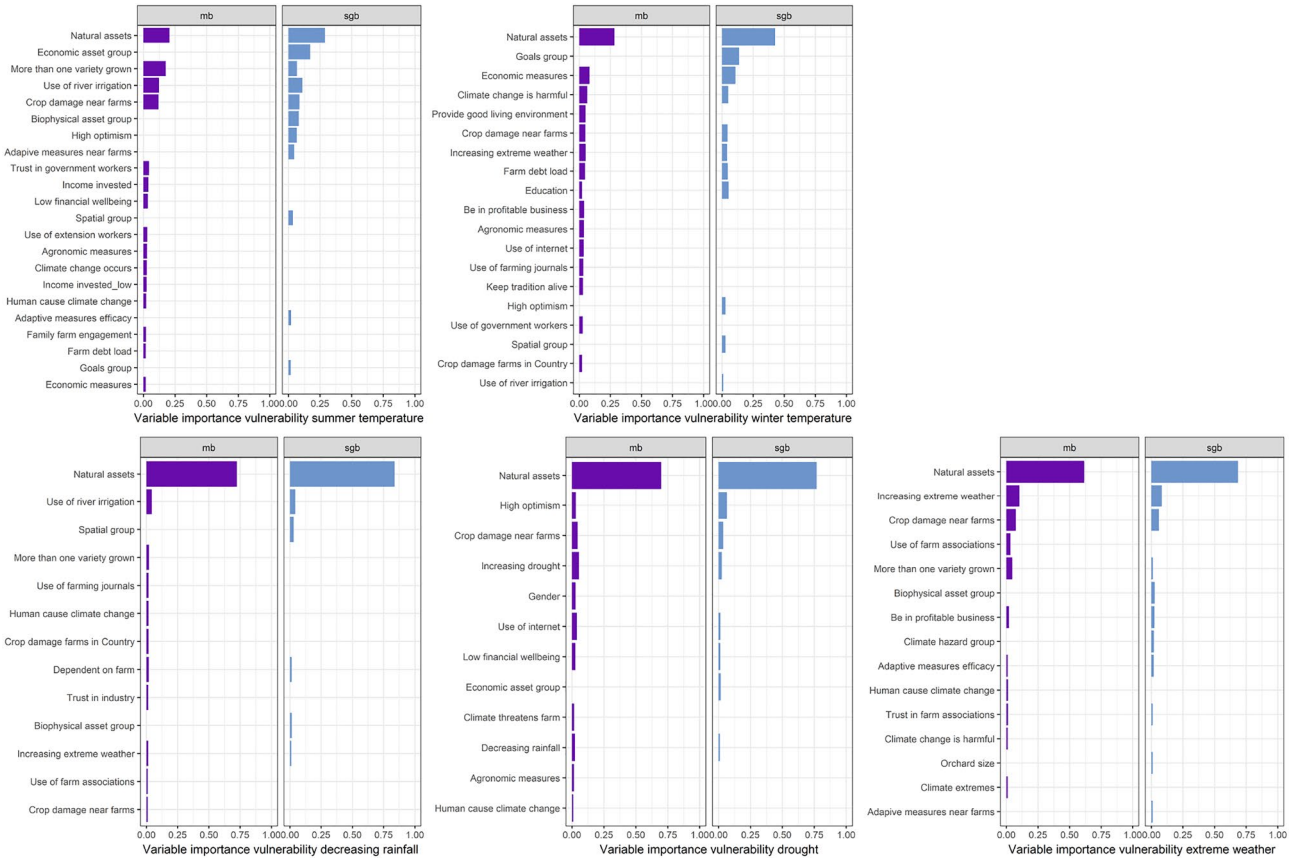

Figure 2. Comparison of variable importance based on component-wise boosting (mb) and sparse group boosting (sgb) for each vulnerability. The ordering of variables is based on the sum of relative importance for both models. only variables with a relative contribution of at least one percent and at most 15 variables per model are shown.

Importance of interactions. In the predictability section, we have already seen some differences between the two-step and the parallel estimation for interaction effects. For predictions, only models trained on the training data were used for model evaluation on the test data. For the variable importance in Fig. 3 and Sparsity in Table 2 the whole data was used. The parallel estimation selected only interaction effects and no main effects (individual variables), whereas the two-step estimation selected both.

Referring to Table 2 it becomes apparent that the selection of variables differs substantially. Overall, the two-step estimation in 2-boost yields much fewer interactions. For summer temperature vulnerability, no interaction term was selected, whereas for the parallel estimation, 13 interaction effects were selected. For decreasing rainfall vulnerability the differences are also substantial. The two-step estimation selected only one interaction, namely the one between Agronomic measures and trust in TV was selected and mb int selected 48. For drought vulnerability, the difference was the smallest with 27 interactions for the parallel and 16 for the two-step estimation. The percentage of selected interactions was four out of five times below one percent for 2-boost and for mb int it was above one percent four out of five times.

Not only does the sparsity differ, but also the selected interactions themselves. Referring to Fig. 3, for winter temperature vulnerability the two interactions "Natural assets"-"Be profitable business" and "Country"-"Farm debt load" have high relative importance based on both models. But apart from those two, there is almost no overlap. For example for decreasing rainfall vulnerability, the only selected interaction between "Agronomic measures"-"Trust in TV" has a relative importance of 1 based on 2-boost and is not selected based on mb int, which in turn selected 48 other interactions.

In Figs. 4, 5, 6, 7 and 8 we plotted the four most important interaction effects for each of the vulnerabilities found in mb int and 2-boost based on a classical logistic regression only using one interaction term at a time. There, the probability of no vulnerability is plotted based on the joint categories of the interaction. This is done once for the data in Chile, Tunisia, and the whole data. Exemplary, we interpret the two common interaction effects "Natural assets"-"Be profitable business" and "Country"-"Farm debt load" for winter temperature vulnerability, which was selected by both models. In the northern region of Chile, having compared to not having the goal of being a profitable business is associated with a higher probability of not being vulnerable to increasing winter temperatures. In the southern Region of Chile, the association is reversed, meaning that having compared to not having the goal of being a profitable business is associated with a lower probability of not being vulnerable against increasing winter temperatures. In Tunisia, in both regions, the association of having the goal of being a profitable business is negative but more negative in the Southern region compared to the Northern

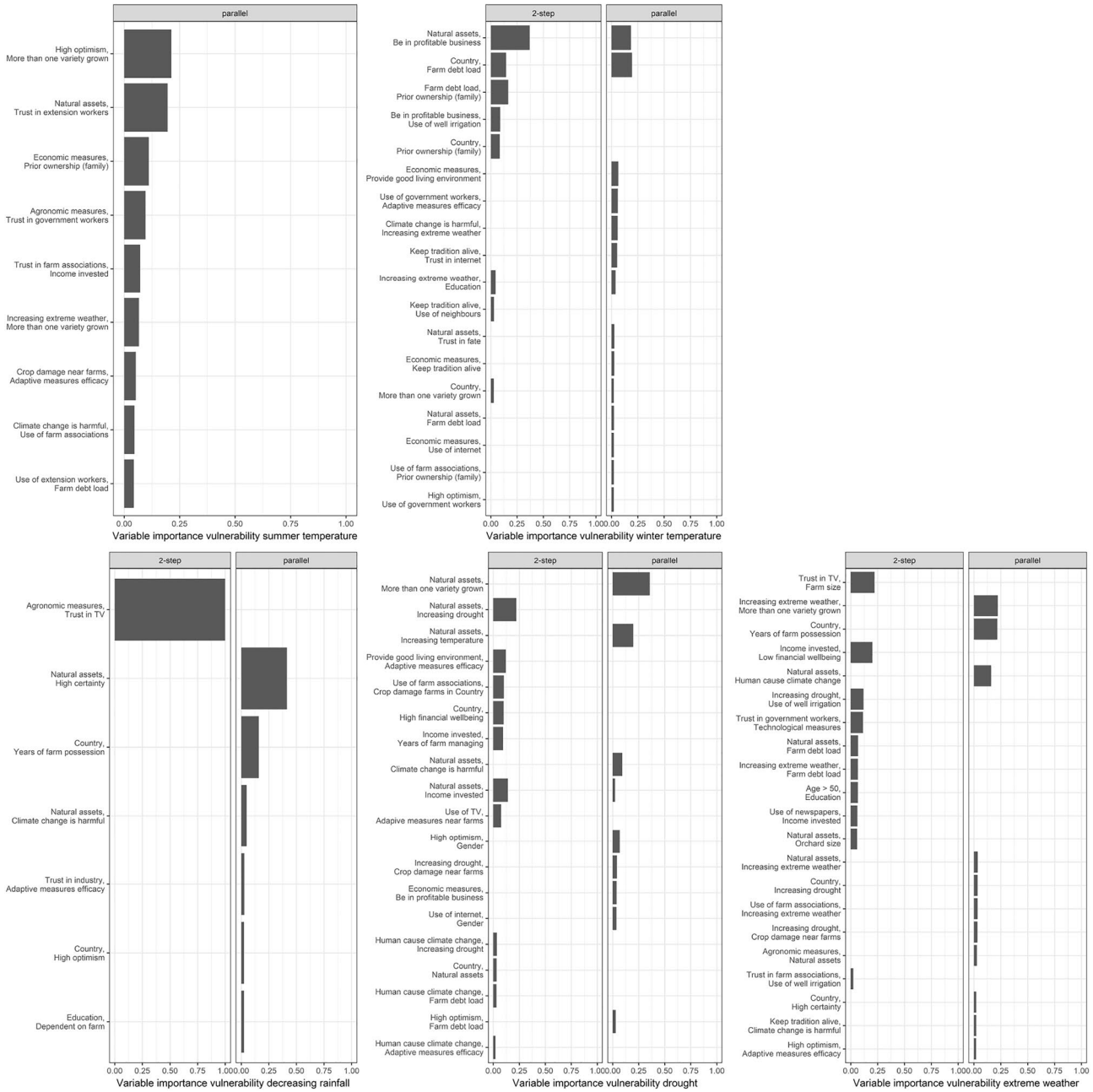

Figure 3. Variable importance of interaction terms in two-step estimation (2-boost) and parallel estimation (2-boost) for each vulnerabilities. The ordering of variables is based on the sum of relative importance for both models Only variables with a relative contribution of at least two percent and at most 15 variables per model are shown.

region. Based on the interaction term "Country"-"Farm debt load", high farm debt load has a positive association with the probability of not being vulnerable to increasing winter temperature, where the association is negative in Tunisia. The positive association in Chile is stronger in the northern region and the negative association in Tunisia is stronger in the southern region.

Discussion

The results indicate that the vulnerability of farmers in Chile and Tunisia against climate hazards can be predicted with the interpretable boosting algorithms and their variations by the variables and groups of variables used in the analysis. All models performed variable selection. The highest predictive power measured in AUC was achieved for vulnerability against decreasing rainfall and the lowest for summer temperature increases regardless of the type of boosting approach. For predicting summer temperature vulnerability the sparse group boosting outperformed all other models indicating that there may be underlying latent variables that cause the effects rather than the individual variables. The group variable importance mainly points to economic and

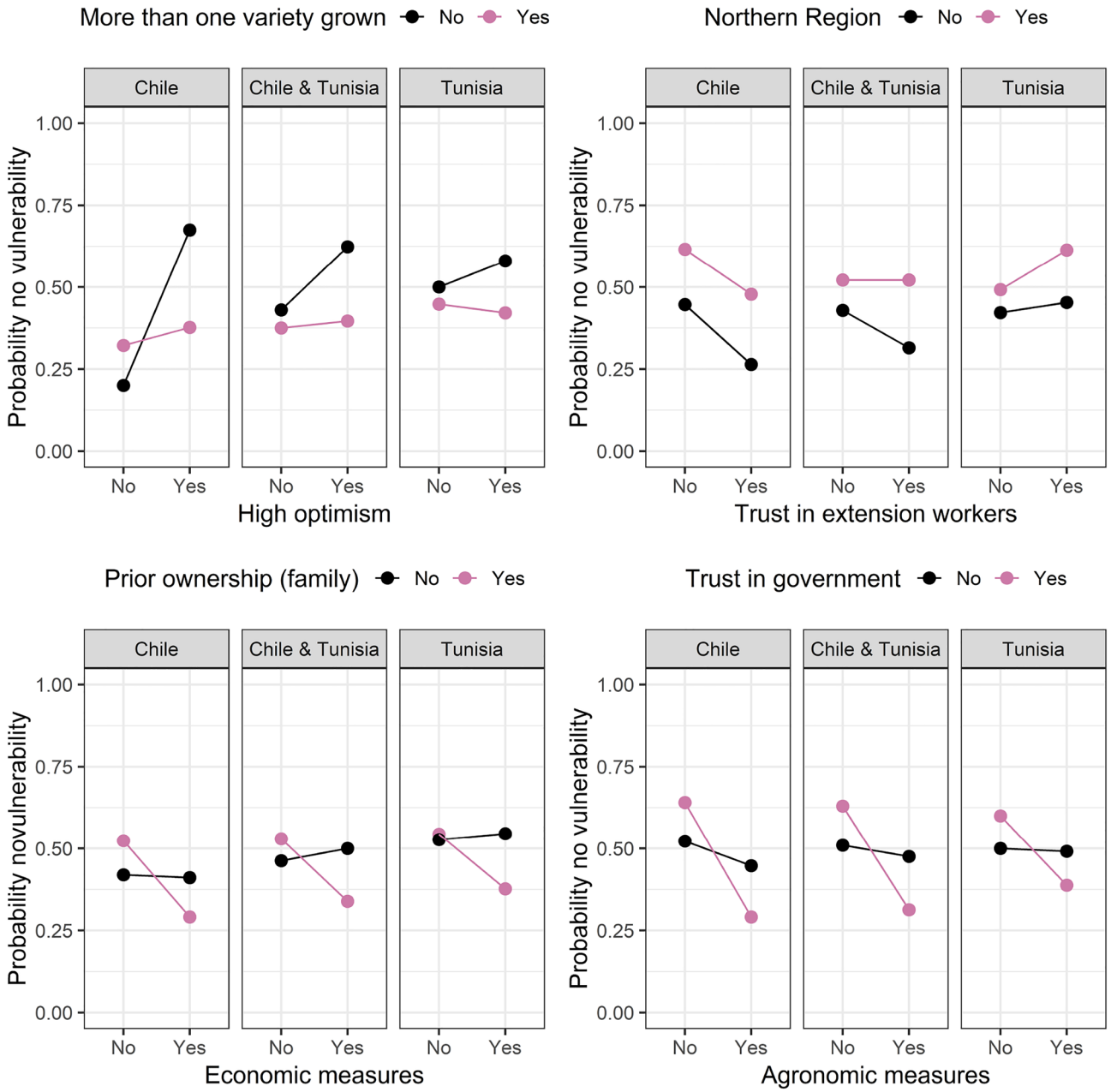

Figure 4. Probability of not being vulnerable against increasing summer temperature based on the categories of the four most important interaction effects found in mb int and 2-boost. Probabilities are based on classical logistic regression only using one interaction term at a time. The results are once stratified by country (Chile, Tunisia) and once estimated on the whole data.

biophysical assets including adaptive measures which may be an underlying determinant for summer temperature vulnerability. The variable importance strongly points to Natural assets consisting of the four different regions in Chile and Tunisia, which are a main determinant of all types of vulnerability. This indicates strong within and between country differences. The interaction analyses also confirm the importance of regionality, as some effects are strongly modulated by Country and North-South comparisons. The modulated effect of debt load by region may be an indication of economic differences between regions and closeness to bigger cities or could be a result of the different climatic zones.

Even though there are strong interaction effects present in the data as seen in the univariate interaction analysis, it is not a simple task to transfer their presence into higher predictive power in a high-dimensional setting. This becomes apparent since the model including interactions base-learners additionally to the main effects performed worse than the same model without interactions in all cases. One of the reasons is probably overfitting, as the number of parameters to estimate exceeds the number of variables by a factor of over four. The result was that the interaction model did not include any main effects and only interactions. We believe that this issue of overfitting becomes more systematic in high-dimensional data than purely random because there

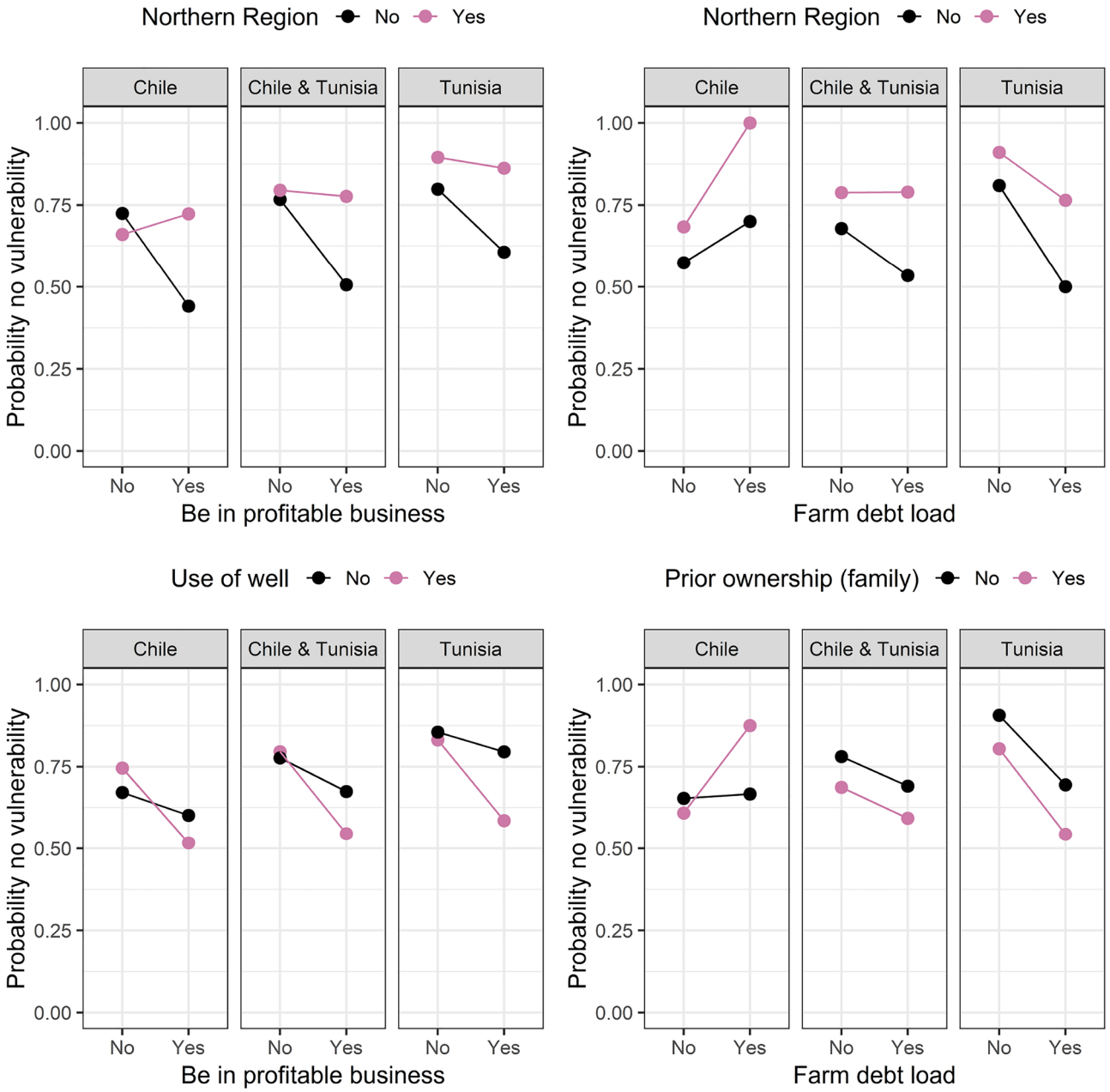

Figure 5. Probability of not being vulnerable against increasing winter temperature based on the categories of the four most important interaction effects found in mb int and 2-boost. Probabilities are based on classical logistic regression only using one interaction term at a time. The results are once stratified by country (Chile, Tunisia) and once estimated on the whole data.

if there are p variables in the dataset, then there are $\mathcal{O}(p^2)$ possible interaction terms. So, with increasing p, the chance of selecting an interaction term over a main effect increases with regardless of the actual effect sizes. This implicit interaction selection bias could be addressed successfully by the proposed two-step boosting approach. The two-step boosting yielded higher predictive power and a higher degree of sparsity with fewer interactions being present in the resulting model. This leads us to believe that this approach is superior to the "classical" parallel estimation by including interaction terms in the main model formula in boosting. The only drawback we see is, that one has to estimate two models instead of just one which slightly increased the programming effort and reduces the potential for further parallelization as the models are fitted sequentially and not in parallel. However, it is common practice and in line with the principle of sparsity to always fit one model that contains only individual variables if one wants to do an interaction analysis[24]. In this case, the two-step boosting is also computationally more efficient because one can build upon the first model and avoid having to refit the main effect.

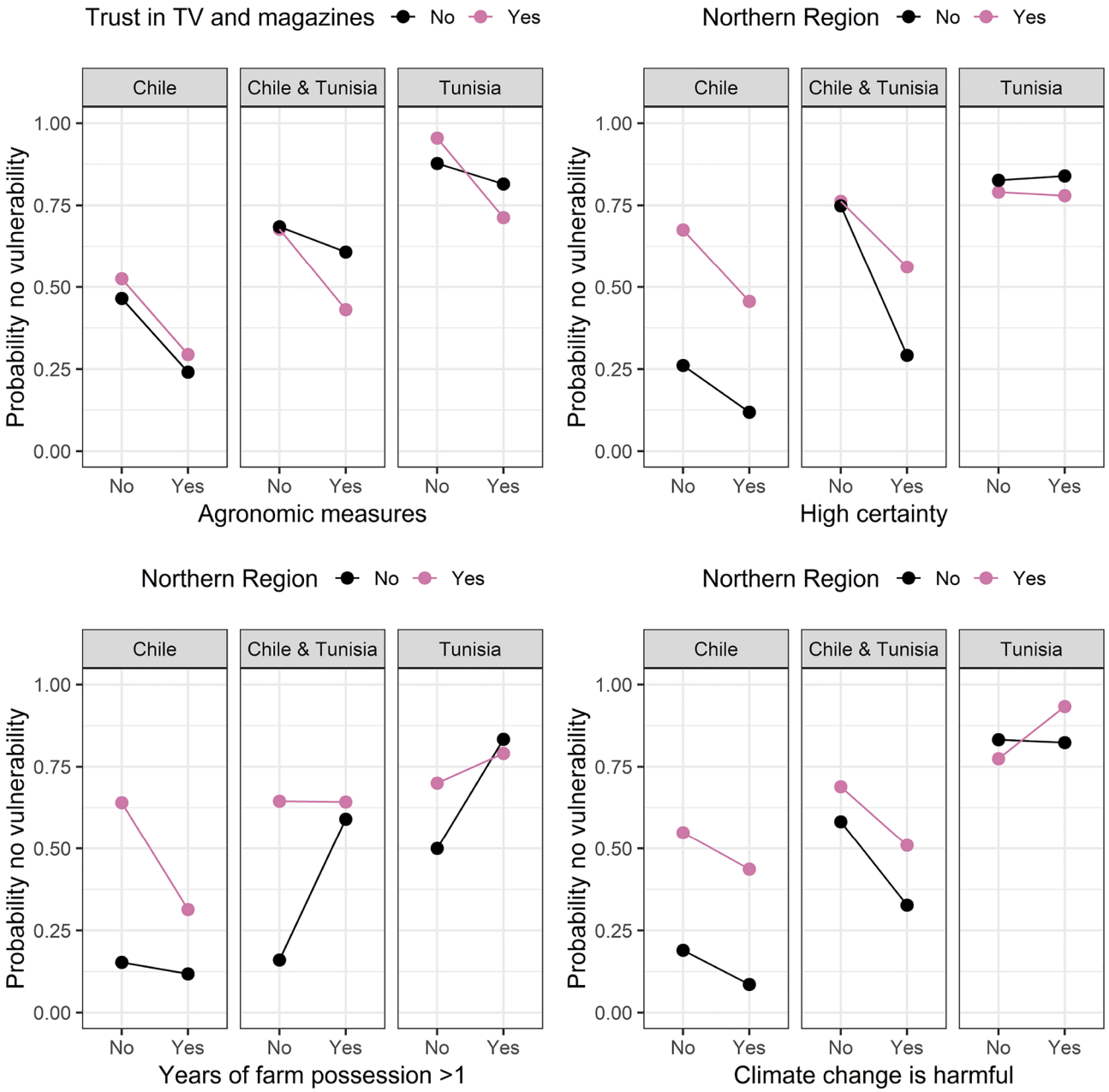

Figure 6. Probability of not being vulnerable against decreasing rainfall based on the categories of the four most important interaction effects found in mb int and 2-boost. Probabilities are based on classical logistic regression only using one interaction term at a time. The results are once stratified by country (Chile, Tunisia) and once estimated on the whole data.

In environmental research, consistently finding associations in high-dimensional datasets requires new methods to advance knowledge. These new methods allow more flexibility but often come at the cost of classical statistical inference, including *p* values and estimations of standard errors as in the case of boosting[25].

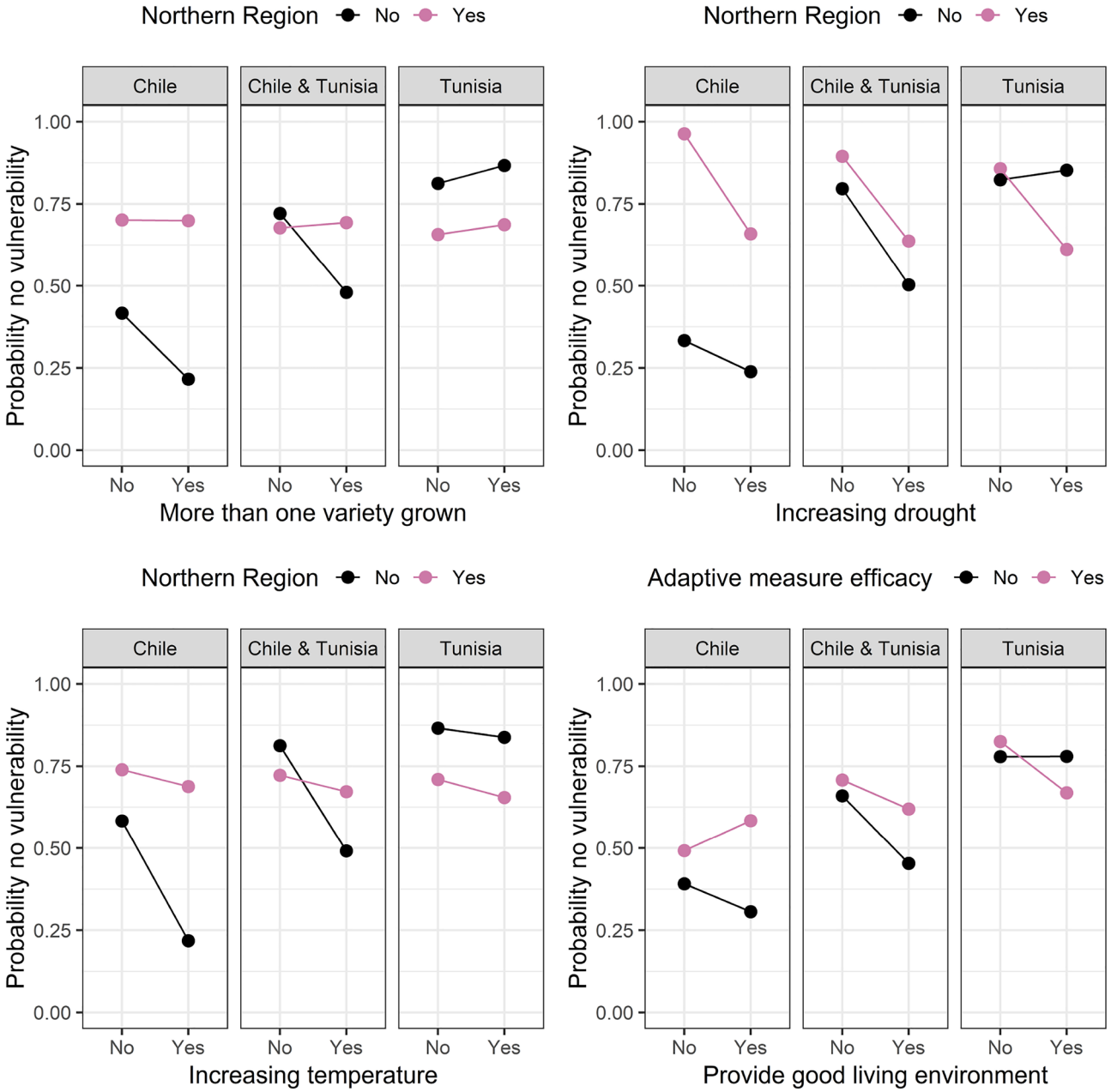

Figure 7. Probability of not being vulnerable against drought based on the categories of the four most important interaction effects found in mb int and 2-boost. Probabilities are based on classical logistic regression only using one interaction term at a time. The results are once stratified by country (Chile, Tunisia) and once estimated on the whole data.

Often, there are multiple plausible explanations for a phenomenon. The here proposed methods can enable direct comparison of a large number of explanations, estimating their explanatory importance for the outcome. This approach can accelerate understanding, particularly for newer phenomena like climate change, by gathering all variables that may be associated with the outcome and sampling observations for them. Starting with a relatively small sample size, one can estimate the relative importance of hypotheses and prioritize future research based on the results.

Using an apriori interpretable method, such as those previously described, provides the great advantage of being able to assess the predictability of a given set of explanations for an outcome. In contrast, post-hoc interpretability tools applied to a black box provide only a simplified explanation of how black-box predictions may be derived, without being able to assess how good the explanations themselves are at predicting the outcome.

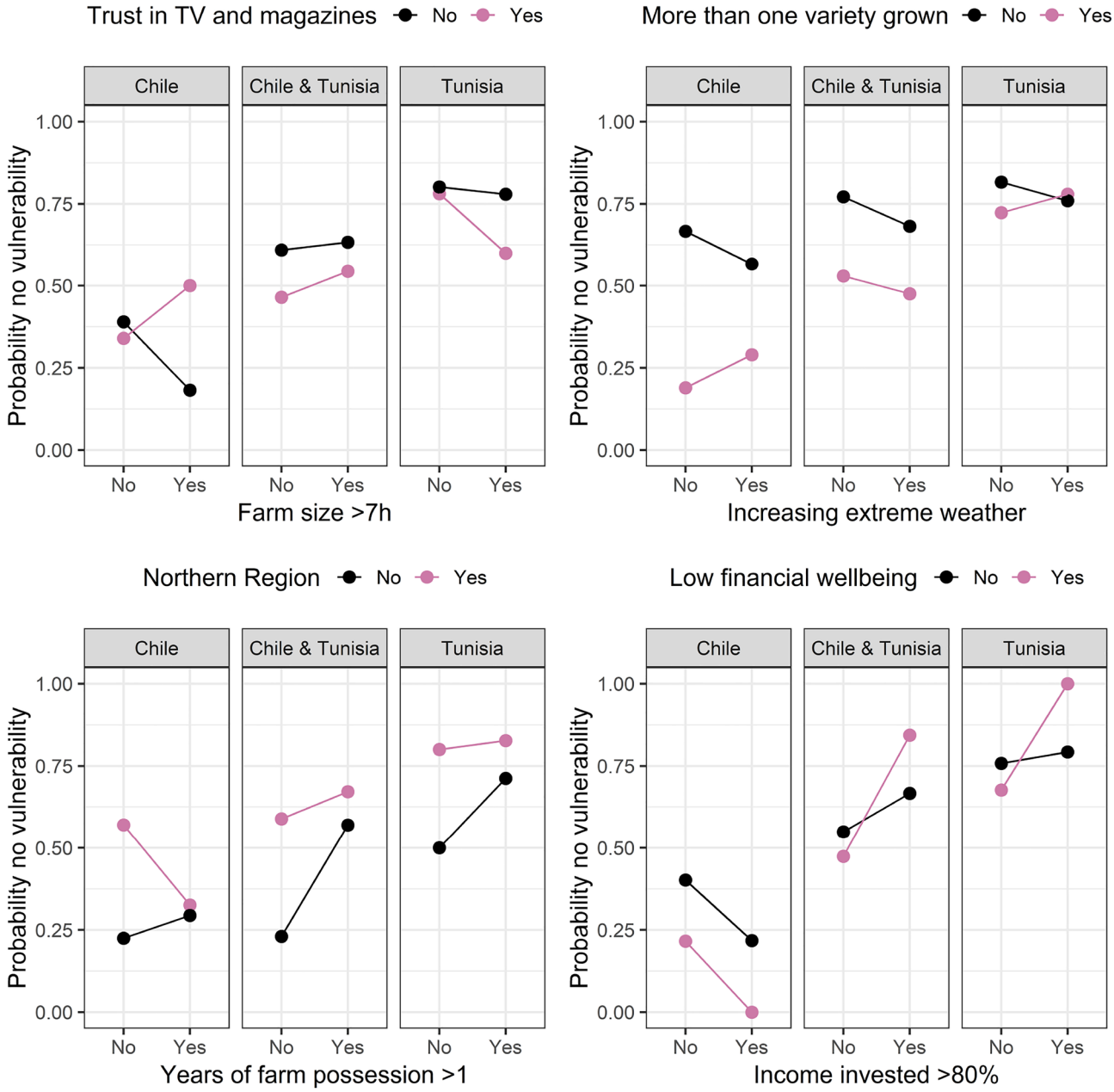

Figure 8. Probability of not being vulnerable against extreme weather based on the categories of the four most important interaction effects found in mb int and 2-boost. Probabilities are based on classical logistic regression only using one interaction term at a time. The results are once stratified by country (Chile, Tunisia) and once estimated on the whole data.

Data availability

The dataset used and analysed during the current study is available from the corresponding author on reasonable request.

Received: 28 April 2023; Accepted: 2 August 2023
Published online: 07 August 2023

References

1. Li, B., Chakraborty, S., Weindorf, D. C. & Yu, Q. Data integration using model-based boosting. *SN Comput. Sci.* **2**, 400. https://doi.org/10.1007/s42979-021-00797-0 (2021).
2. Obster, F., Bohle, H. & Pechan, P. M. Factors other than climate change are currently more important in predicting how well fruit farms are doing financially. arXiv:2301.07685 [cs, stat] (2023).
3. Obster, F., Brand, J., Ciolacu, M. & Humpe, A. Improving boosted generalized additive models with random forests: a zoo visitor case study for smart tourism. *Procedia Comput. Sci.* **217**, 187–197. https://doi.org/10.1016/j.procs.2022.12.214 (2023).
4. Freund, Y. & Schapire, R. E. A decision-theoretic generalization of on-line learning and an application to boosting. *J. Comput. Syst. Sci.* **55**, 119–139. https://doi.org/10.1006/jcss.1997.1504 (1997).

5. Jennifer, J. J. Feature elimination and comparison of machine learning algorithms in landslide susceptibility mapping. *Environ. Earth Sci.* **81**, 489. https://doi.org/10.1007/s12665-022-10620-5 (2022).
6. Froeschke, J. T. & Froeschke, B. F. Spatio-temporal predictive model based on environmental factors for juvenile spotted seatrout in Texas estuaries using boosted regression trees. *Fish. Res.* **111**, 131–138. https://doi.org/10.1016/j.fishres.2011.07.008 (2011).
7. Tibshirani, R. Regression shrinkage and selection via the lasso. *J. R. Stat. Soc.: Ser. B (Methodol.)* **58**, 267–288 (1996).
8. Zhao, P. & Yu, B. Boosted Lasso. Tech. Rep., California Univ Berkeley Dept of Statistics. Section: Technical Reports (2004).
9. Simon, N., Friedman, J., Hastie, T. & Tibshirani, R. A sparse-group lasso. *J. Comput. Graph. Stat.* **22**, 231–245 (2013).
10. Pechan, P. M., Obster, F., Marchioro, L. & Bohle, H. Climate change impact on fruit farm operations in Chile and Tunisia. *agriRxiv* **2023**, 20230025166. https://doi.org/10.31220/agriRxiv.2023.00171 (2023).
11. Pechan, P. M., Bohle, H. & Obster, F. Reducing vulnerability of fruit orchards to climate change. *Agric. Syst.* **210**, 103713. https://doi.org/10.1016/j.agsy.2023.103713 (2023).
12. Team, R. RStudio: Integrated Development Environment for R (2020).
13. Hofner, B., Mayr, A., Robinzonov, N. & Schmid, M. Model-based boosting in R: A hands-on tutorial using the R package mboost. *Comput. Stat.* **29**, 3–35. https://doi.org/10.1007/s00180-012-0382-5 (2014).
14. Hothorn, T., Buehlmann, P., Kneib, T., Schmid, M. & Hofner, B. mboost: Model-based boosting. *CRAN* (2022).
15. Bühlmann, P. & Hothorn, T. Boosting algorithms: Regularization, prediction and model fitting. *Stat. Sci.* **22**, 477–505. https://doi.org/10.1214/07-STS242 (2007).
16. Friedman, J. H. Greedy function approximation: A gradient boosting machine. *Ann. Stat.* **29**, 1189–1232. https://doi.org/10.1214/aos/1013203451 (2001).
17. Tutz, G. & Binder, H. Boosting ridge regression. *Comput. Stat. Data Anal.* **51**, 6044–6059. https://doi.org/10.1016/j.csda.2006.11.041 (2007).
18. Agarwal, N. K. Verifying survey items for construct validity: A two-stage sorting procedure for questionnaire design in information behavior research. *Proc. Am. Soc. Inf. Sci. Technol.* **48**, 1–8. https://doi.org/10.1002/meet.2011.14504801166 (2011).
19. Obster, F. & Heumann, C. Sparse-group boosting—Unbiased group and variable selection. arXiv:2206.06344 [stat] (2022).
20. Pybus, M. *et al.* Hierarchical boosting: A machine-learning framework to detect and classify hard selective sweeps in human populations. *Bioinformatics* **31**, 3946–3952. https://doi.org/10.1093/bioinformatics/btv493 (2015).
21. Wang, C., Wu, Y. & Liu, Z. Hierarchical boosting for transfer learning with multi-source. In *Proceedings of the International Conference on Artificial Intelligence and Robotics and the International Conference on Automation, Control and Robotics Engineering*, ICAIR-CACRE '16, 1–5. https://doi.org/10.1145/2952744.2952756 (Association for Computing Machinery, New York, 2016).
22. Yang, F. *et al.* Feature Pyramid and Hierarchical Boosting Network for Pavement Crack Detection. arXiv:1901.06340 [cs] (2019).
23. Valentini, G. Hierarchical ensemble methods for protein function prediction. *ISRN Bioinform.* **2014**, 901419. https://doi.org/10.1155/2014/901419 (2014).
24. Aguinis, H. & Gottfredson, R. K. Best-practice recommendations for estimating interaction effects using moderated multiple regression. *J. Organ. Behav.* **31**, 776–786. https://doi.org/10.1002/job.686 (2010).
25. Hofner, B., Mayr, A. & Schmid, M. gamboostLSS: An R package for model building and variable selection in the GAMLSS framework. arXiv:1407.1774 [stat] (2014).

Author contributions

P.P. and H.B. accumulated the data, F.O. performed machine learning and statistical modeling, and F.O. analysed the results. All authors reviewed the manuscript.

Funding

Open Access funding enabled and organized by Projekt DEAL. This research was conducted within the project "Phenological And Social Impacts of Temperature Increase - climatic consequences for fruit production in Tunisia, Chile and Germany" (PASIT; grant number 031B0467B of the German Federal Ministry of Education and Research). Open Access funding was enabled by Universität der Bundeswehr München. Additional funding was provided by dtec.bw funded by NextGenerationEU.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to F.O. or P.P.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.