Forest Ecosystems

RESEARCH Open Access

Performance of statistical and machine learning-based methods for predicting biogeographical patterns of fungal productivity in forest ecosystems

Albert Morera[1,2]*, Juan Martínez de Aragón[3], José Antonio Bonet[1,2], Jingjing Liang[4] and Sergio de-Miguel[1,2]

Abstract

Background: The prediction of biogeographical patterns from a large number of driving factors with complex interactions, correlations and non-linear dependences require advanced analytical methods and modeling tools. This study compares different statistical and machine learning-based models for predicting fungal productivity biogeographical patterns as a case study for the thorough assessment of the performance of alternative modeling approaches to provide accurate and ecologically-consistent predictions.

Methods: We evaluated and compared the performance of two statistical modeling techniques, namely, generalized linear mixed models and geographically weighted regression, and four techniques based on different machine learning algorithms, namely, random forest, extreme gradient boosting, support vector machine and artificial neural network to predict fungal productivity. Model evaluation was conducted using a systematic methodology combining random, spatial and environmental blocking together with the assessment of the ecological consistency of spatially-explicit model predictions according to scientific knowledge.

Results: Fungal productivity predictions were sensitive to the modeling approach and the number of predictors used. Moreover, the importance assigned to different predictors varied between machine learning modeling approaches. Decision tree-based models increased prediction accuracy by more than 10% compared to other machine learning approaches, and by more than 20% compared to statistical models, and resulted in higher ecological consistence of the predicted biogeographical patterns of fungal productivity.

(Continued on next page)

* Correspondence: morera.marra@gmail.com
[1]Department of Crop and Forest Sciences, University of Lleida, Av. Alcalde Rovira Roure 191, E-25198 Lleida, Spain
[2]Joint Research Unit CTFC-AGROTECNIO-CERCA Center, Av. Rovira Roure 191, 25198 Lleida, Spain
Full list of author information is available at the end of the article

(Continued from previous page)

Conclusions: Decision tree-based models were the best approach for prediction both in sampling-like environments as well as in extrapolation beyond the spatial and climatic range of the modeling data. In this study, we show that proper variable selection is crucial to create robust models for extrapolation in biophysically differentiated areas. This allows for reducing the dimensions of the ecosystem space described by the predictors of the models, resulting in higher similarity between the modeling data and the environmental conditions over the whole study area. When dealing with spatial-temporal data in the analysis of biogeographical patterns, environmental blocking is postulated as a highly informative technique to be used in cross-validation to assess the prediction error over larger scales.

Keywords: Modeling, Regression, Biogeography, Climate, Forest, Fungi, Mushrooms

Background

Understanding the biogeographical patterns of organisms in natural ecosystems and predicting their distribution is a fundamental challenge in environmental sciences (Ehrlén and Morris 2015). This entails a deep understanding of their distribution across space and time underpinning ecological mechanisms, which becomes increasingly complex with an increasing amount of factors driving these patterns and the possible interactions and nonlinear dependencies between them (Dixon et al. 1999; Ye et al. 2015). Such complex interrelationships require advanced data analytic methods and modeling tools to yield realistic predictions of natural ecosystem attributes and processes.

Statistical methods traditionally used for this purpose aim at accounting for several elements that govern these natural mechanisms, trying to reach a parsimonious and robust understanding of ecological patterns (Wood and Thomas 1999). However, since conventional parametric approaches may over-simplify nonlinear relationships between variables and over- or under-estimate the influence of some drivers, conventional parametric approaches may result in poor predictions and/or descriptions of reality (Ye et al. 2015), especially for the analyses of large databases. To overcome potential limitations of classic statistical approaches in big data analysis, the increased computing power has led to recent considerable growth in the use of analytical methods based on artificial intelligence such as machine learning (Christin et al. 2019).

Machine learning algorithms are increasingly being used in species distribution and ecological niche modeling (Prasad et al. 2006; Cutler et al. 2007; Hannemann et al. 2015; Liang et al. 2016; Prasad 2018; Gobeyn et al. 2019), forest resources (Stojanova et al. 2010; Görgens et al. 2015) and climate change studies (Thuille 2003; Bastin et al. 2019), among others. To determine to what extent these “new” methodologies can contribute to improving our understanding and prediction capacity within the field of environmental sciences, comparative studies are required between those models that have been used historically and those fed by artificial intelligence algorithms (Özçelik et al. 2013; Diamantopoulou et al. 2015; Hill et al. 2017; Bonete et al. 2020). Yet, many machine learning algorithms have been developed in recent years, and each of them may be more or less appropriate depending on the specific tasks and research objectives (Thessen 2016). This highlights the need for systematic studies allowing for discerning the most suitable methodology according to a given research objective and data. Although several studies have analysed the performance of different analytical approaches (Hill et al. 2017; Bonete et al. 2020; Mayfield et al. 2020), existing ecological research addressing systematic assessments and comparisons of alternative modeling and predictive methods is scarce, making it difficult to provide clear methodological recommendations about the suitability of different approaches. Besides, in the field of environmental sciences, often, extrapolations in biophysically differentiated areas are required, which makes it necessary to take even more into account the data spatial dependencies. Due to data spatial autocorrelation, random cross-validations leads to over-optimistic error estimates (Bahn and McGill, 2012; Micheletti et al., 2013; Juel et al. 2015; Gasch et al. 2015; Roberts et al. 2017; Meyer et al. 2018; Meyer et al. 2019a), which makes it necessary to use proper, complementary validation methods such as spatial cross-validation (Le Rest et al. 2014; Pohjankukka et al. 2017; Roberts et al. 2017; Meyer et al. 2018; Valavi et al. 2018). Moreover, spatial dependencies in the data can lead to a misinterpretation of some predictors outside the sampling range (Meyer et al. 2018, 2019a).

Biogeographical patterns of fungal dynamics over large scales are a highly relevant question in ecology given the key role of fungi in forest ecosystems (Stokland et al. 2012; Mohan et al. 2014), especially in fungi-tree symbiosis. However, due to their great diversity and differential ecological requirements (Glassman et al. 2017), as well as the difficulty of monitoring their dynamics and the large array of potential drivers (Büntgen et al. 2013), little is known of fungal dynamics over large scales. The prediction of biogeographical patterns of fungal dynamics requires large fungal datasets with a correct taxonomic identification of the specimens and a consistent sampling methodology across space and time to avoid sample bias (Hao et al. 2020).

In particular, the spatially-explicit prediction of fungal productivity, i.e. mushroom fruiting patterns, is a key feature of fungal dynamics as it is tightly related to the supply of multiple provisioning, regulating and cultural ecosystem services (Boa 2004). However, the high correlation between mushroom production and meteorological conditions among other drivers (Taye et al. 2016; Alday et al. 2017; Collado et al. 2019) makes the prediction of mushroom production challenging, especially in Mediterranean ecosystems where there is a high interannual variability of climatic conditions. The long period of potential fruiting of different mushroom species, as a result of their adaptation to the recurrent climatic patterns of a dry summer followed by a wet autumn (Barnard et al., 2014), makes mushroom yields dependent on a large number of variables. Precipitation and temperatures on a weekly scale can be the factors that lengthen, shorten or shift the fruiting period (Gange et al. 2007; Kauserud et al. 2008; Kauserud et al., 2009; Büntgen et al. 2012), and also those that modulate mushroom production to a higher degree (Karavani et al. 2018). The large number of variables involved and their presumed interactions may often yield a misconception that fungal productivity is highly stochastic or very difficult to predict. Previous research to estimate mushroom productivity over large scales has been mainly based on mixed-effects modeling (de-Miguel et al. 2014; Sánchez-González et al. 2019). Despite being a valid approach, it may have certain limitations that are worth assessing in comparison with alternative methods that remain unexplored.

This study compares different statistical and machine learning models in estimating mushroom productivity at the landscape level, together with a systematic methodology to determine the best approach to predict mushroom production in forest ecosystems. Using climatic and biophysical data together with in situ fungal records collected weekly over more than 20 years on a hundred permanent plots, we developed spatially explicit, high-resolution continuous maps of mushroom productivity that were also used in the selection of the most suitable methods for predicting this ephemeral and important forest resource. Specifically, we compared two statistical models, namely, generalized linear mixed models (GLMM) and geographically weighted regression models (GWR), as well as four alternative state-of-the-art machine learning algorithms such as random forest (RF), extreme gradient boosting (XGB), support vector machine (SVM) and artificial neural network (ANN).

Methods

Study area and sampling plots

The study area was Catalonia region, northeastern Spain, in the western Mediterranean basin. The forest ecosystem types considered in this study were the main Mediterranean pine forest ecosystems that represent the majority of the forest area of the study region, namely, pure stands of *Pinus halepensis*, *P. sylvestris*, *P. pinaster*, *P. nigra* and *P. uncinata* and mixed stands of *P. halepensis* and *P. nigra*, and of *P. sylvestris* and *P. nigra*. We used a dataset that contains information from 98 permanent monitoring plots for fungal dynamics sampled on a weekly basis during the main mushroom fruiting period, between August and the end of December and from 1997 to 2019. The plots were distributed randomly and proportionally to the relative surface of the different pine forest ecosystems (Bonet et al. 2010) (Fig. S1). Data were aggregated to an annual basis to create predictive models to estimate annual mushroom productivity. More information about the sampling methods and data can be found in Bonet et al. (2004), Martínez de Aragón et al. (2007) and Table S3.

Climate and biophysical data

Meteorological data for each sampling plot was obtained from the interpolation and altitudinal correction of daily weather of 201 meteorological stations from the Catalan Meteorological Service and the Spanish Meteorological Agency. Interpolation was conducted with "meteoland" R package (v0.8.1; De Cáceres et al. 2018) that uses a modification of the DAYMET methodology (Thornton et al. 1997; Thornton and Running 1999). Likewise, to determine the typical climatic conditions across the whole study region, we used the mean of the interpolated daily weather variables for the period between 1991 and 2016 with 1-km resolution. We computed the accumulated monthly rainfall from August to October and the mean, maximum and minimum monthly temperatures for the same period, coinciding with the main mushroom fruiting period.

The total area covered by the different pine forest ecosystems was retrieved from the CORINE habitats map (Commission of the European Community 1991). The biophysical variables such as elevation, slope, aspect and stand basal area were obtained at 20-m resolution from the first cover of the LIDARCAT Project (http://territori.gencat.cat/es/detalls/Article/Mapes_variables_biofisiques_arbrat) based on different LiDAR flights between 2008 and 2011 with a point density of 0.5 points·m^{-2}.

Analytical methods

We used and compared six different analytical methods to predict annual mushroom productivity. Two analytical approaches were based on statistical methods, namely, generalized linear mixed-effects models (GLMM) and geographically weighted regression (GWR), whereas the other four analytical methods were based on alternative machine learning approaches, namely, random forest (RF), extreme gradient boosting (XGB), support vector machine (SVM) and artificial neural network (ANN).

Statistical models fitting

We used a two-stage modeling approach to take into account the high frequency of "zero" production values in many sample plots over time (Hamilton and Brickell 1983; de-Miguel et al. 2014; Karavani et al. 2018; Collado et al. 2018). The high occurrence of these values arise from the small size of the plots and the stochastic nature of mushroom emergence (de-Miguel et al. 2014).

The first stage determines the probability of mushroom emergence, according to binomial data of presence/absence, using a logistic regression $\pi(X) = E(Y|X)$ and a logit link function to represent the conditioned mean of Y given X (Eqs. 1 and 2).

$$\pi(x_k) = E(Y|\mathrm{x}_k) = \frac{1}{1+e^{-g(x_k)}} \tag{1}$$

$$g(x_k) = \ln\left(\frac{\pi(x_k)}{1-\pi(x_k)}\right) = \beta_0 + \beta_k x_k \tag{2}$$

where π (x_k) is the probability of mushroom occurrence, g (x_k) the logit transformation of π (x_k), Y is the dependent variable (mushroom presence/absence), x_k is k^{th} independent variable, β_0 is the intercept parameter and β_k is the regression coefficient for k^{th} independent variable.

The second stage was based on the modeling of the production of non-zero production values at logarithmic scale using linear regression $y = E$ $(\log(Y)|X) + \varepsilon$. Logarithmic transformation allows to limit the production range in the interval $[0, \infty)$, depending on the values of X (Eq. 3). The proportional bias of the logarithmic regression was corrected with the Snowdon's bias correction factor (Snowdon 1991) based on the ratio of the arithmetic sample mean and the mean of the back-transformed predicted values from the regression (Eq. 4):

$$\ln(\mathrm{prod}) = \beta_0 + \beta_k x_k + \varepsilon \tag{3}$$

$$CF = Y/\hat{y} \tag{4}$$

where ln (prod) represents the non-zero production of mushrooms at logarithmic scale, β_0 is the intercept parameter, β_k the regression coefficient for k^{th} independent variable, ε the random error of the deviation of the observations from the conditioned mean of ln(Y) and $Y/\hat{y}$ is the ratio between the mean of observed and the mean of predicted values of the sapling units.

Finally, the total production of mushrooms was obtained from the product of the probability of appearance and the conditioned production of non-zero values (Eq. 5).

$$\mathrm{yield} = \pi(x_k)\, e^{\ln(\mathrm{yield})}\, CF \tag{5}$$

where π (x_k) is the probability of mushroom occurrence, ln (yield) represents the production of mushrooms at logarithmic scale and CF is bias correction factor.

Generalized linear mixed models Due to mushrooms sampling methodology, where annually data was taken from a network of permanent plots, we used GLMM (de-Miguel et al. 2014; Karavani et al. 2018; Collado et al. 2018). This method can consider the spatial and temporal autocorrelation among observations (Pinheiro and Bates 2000) adding random effects to segment the data into different groups according to year and plot. In the proposed mixed-effects models only random effects on model interception were considered. All the models were fitted using the "glmer" function from the "lme4" R package (v1.1–21; Bates et al. 2015).

Geographically weighted regression GWR is a non-stationary modeling technique that describes the spatially varying relationships between the dependent variable and the explanatory variables (Wheeler and Páez 2009). Coefficients of a GWR-based model are given by the spatial location of data and can be estimated for any new location. This means that given a grid, the estimated coefficients for each point in space vary continuously as a function of the spatial heterogeneity of the relationships.

Coefficients for each regression point were calibrated using the data around itself. Due to the annual sampling methodology and the geographical distribution of plots, some plots were grouped denser in some areas and less dense in others. Consequently, we used an adaptive window according to the spatial density of our plots (Georganos et al. 2017). Occurrence and conditional production models, as well as the optimal data value for adjusting the adaptive window, were obtained from the "ggwr" and "gwr.set" functions, respectively, of the R package "spgwr" (v0.6–32; Bivand and Yu, 2017).

Hyperparameters optimization of machine learning models

To optimize the performance of machine learning models, it was necessary to tune their respective hyperparameters (Hutter et al. 2011; Bergstra and Bengio 2012; Duarte and Wainer 2017). This needs to be conducted prior to the training of the final predictive models and also needs to consider the spatial and temporal dependencies of both the modeling and prediction datasets (Schratz et al. 2019). Since hyperparameters tuning based on a resampling method such as k-fold cross-validation may lead to an incorrect tuning for models that aim to predict in environmentally different areas (Roberts et al. 2017), hyperparameters tuning was conducted based on several alternative resampling techniques, namely, k-fold cross-validation, spatial cross-validation, and environmental cross-validation (Roberts et al. 2017).

We used an optimization algorithm based on a search grid (Bergstra and Bengio 2012) implemented in the R package "mlr3tuning" (v0.5.0; Becker et al. 2020) that selects the best hyperparameters configuration according to a given metric. The search space was defined from the Cartesian product of the discretized values of a set of n hyperparameters to be tuned in each model (Table 1). First, a set of hyperparameters configurations from the search grid was randomly selected and evaluated according to each resampling strategy. A search grid of resolution 25 was defined and tested with a total of 250 different hyperparameters configurations. Otherwise, we used 10 folds in each resampling strategy using the R package "mlr3" (v0.9.0; Lang et al. 2019) for the core computational operations and the extensions "mlr3spatiotempcv" (v0.1.1; Schratz and Becker 2021) and "mlr3learners" (v0.4.3; Lang et al. 2020a) for the resampling and the use of the different models, respectively. In addition, the R packages "paradox" (v0.6.0; Lang et al. 2020b) and "mlr3keras" (v0.1.3; Pfisterer et al. 2021) were also used in hyperparameters tuning.

Model and variable selection and evaluation

Statistical model evaluation and variable selection was based on the current knowledge of forests and mushroom ecology, the statistical significance of model parameters ($p < 0.05$ or $t > |1.96|$), the variance inflation factor (VIF) to quantify the severity of multicollinearity and the parsimony principle. To check the sensitivity/specificity of the binomial classification models we used Receiver Operator Characteristic (ROC) curves, using the R package "ROCR" (v1.0–7; Sing et al. 2005).

Table 1 Hyperparameters ranges and types for each machine learning model

Algorithm	Hyperparameter ID	Type	Lower	Upper
RF	mtry	Integer	1	N° of predictors
	min.node.size	Integer	1	100
	num.trees	Integer	2	500
XGB	nrounds	Integer	1	100
	gamma	Numeric	1	25
	max_depth	Integer	1	15
	eta	Numeric	0.1	1
SVM	cost	Numeric	1	50
	gamma	Numeric	0.1	1
ANN	epochs	Integer	1	100
	batch_size	Integer	1	N° observations

"Hyperparameter id" corresponds to the names specified in the R package used to train each model. RF (random forest), XGB (extreme gradient boosting), SVM (support vector machine), and ANN (artificial neural network) models were trained using the R packages "ranger" (v0.12.1; Wright and Ziegler 2017), "xgboost" (v0.90.0.2; Chen et al. 2019), "e1071" (v1.7–2; Meyer et al. 2019b) and "keras" (v2.3.0.0.0; Allaire and Chollet 2019), respectively

To assess whether GWR improved GLMM due to the non-stationary nature of data and to avoid introducing an improvement that was not attributed to the type of modeling, the same explanatory variables as in GLMM were used. To test the non-stationarity of the independent variables of GWR models, the local parameters were compared with global GLMM coefficients. The probability of incorporating non-stationary variables increases if the estimated coefficient of the variable in GLMM (± standard error) is outside the 1st and 3rd quartile of the GWR model coefficient (Propastin 2009).

For each of the four machine learning algorithms, two models were adjusted. The first ones were trained using a total of 15 biophysical variables, while the second ones were trained using a subset of them (Table S1). This subset was determined from the same five variables used in the statistical models, including climate predictors only. This allowed us to assess separately the prediction accuracy due to the analytical method, and the prediction accuracy due to differences between predictors or the number of explanatory variables. The final machine learning models were trained using 100% of the sampled data (henceforth referred to as "modeling data") and the optimal hyperparameters settings. We used the optimized hyperparameters from an environmental blocking (Roberts et al. 2017) to train the final models. The relationship between predictors and mushroom productivity was assessed based on partial dependence plots (PDPs), a low-dimensional graphical rendering between variable pairs, in order to determine whether this relationship lacked ecological sense. The importance of the predictors of the models was determined from a sensitivity analysis using the R package "rminer" (v1.4.2; Cortez 2016).

Evaluation of the predictive performance and mapping

Since the main purpose of this study was to develop models to accurately predict mushroom productivity, this entails the evaluation of the predictive performance of the resulting models also outside of the range of the training (or fitting) region. With the aim of determining the similarities between modeling data and the environmental conditions of the whole study area, a principal component analysis (PCA) based on both datasets altogether was used. Using the location of the modeling data within the space described by the first two components of the PCA, a density map was created based on a two-dimensional kernel density estimation and implemented in the "kde2d" function of the R package "MASS" (Venables et al. 2002). By overlapping this density map and the location of each pixel of the study area within the space defined by the two principal components of the

PCA, a similarity value (ranging from 0 to 1) was obtained for each pixel of the study area based on the modeling data density. This similarity value was classified in three categories: low [0, 0.1], medium (0.1, 0.3] and high (0.3, 1] similarity. This classification mainly aimed at detecting the areas with very low similarity, i.e. with very different climatic conditions compared to the modeling data. This whole procedure was performed separately using the 5 climatic predictors of the 5-variable models and the 12 climatic predictors of the 15-variable models, respectively.

To evaluate and compare the predictive accuracy of the models, different resampling strategies were used (see the section of Hyperparameters optimization of machine learning models). The MSE and bias2 of the models were estimated by averaging, respectively, over the MSE and bias2 obtained from each of the 10 folds for each cross-validation strategy.

To generate the landscape-level mushroom productivity maps we used the predictions of the final trained models. These maps were constructed with a resolution of 1 km in accordance with the resolution of the climatic data. The resulting maps were evaluated on the basis of the scientific and expert knowledge about biogeographical fungal productivity patterns in order to assess whether they followed ecologically logical patterns (related to climatic conditions). Thus, we would expect smoothed estimates across the territory driven by the variations of the most important predictors of each model.

Results

Relationships between dependent and explanatory variables

Statistical models showed a statistically significant and positive relationship of mushroom productivity with rainfall in August, September and October (both in conditioned production and occurrence models). On the other hand, conditioned production and occurrence models also showed a statistically significant and negative relationship with the maximum temperature of August and October, respectively (Tables S1, S2, S3). Yet, the coefficients of GWR models varied according to geographical location (Table S3 and Fig. S2), describing certain non-stationarity in both precipitation and temperature.

Within GLMM models, PDPs showed an almost linear relationship between the amount of precipitation between August and October and mushroom productivity in the model fit data range. In contrast, GWR showed an accelerated growth in productivity by increasing rainfall, which was accentuated in those locations with a higher precipitation regression coefficient. Besides, and similarly in GLMM and GWR, the maximum temperature in August showed a decelerated decrease in productivity by increasing temperature, while the maximum temperature of October, even though it showed a negative relation, resulted in little relevance in mushroom productivity for the range of values of the fitting data (Fig. 1).

Different machine learning models resulted in rather similar relationships between variables although, due to the particularities of each algorithm, the patterns changed slightly between approaches. In contrast to the relationships in GLMM and GWR models, some of the machine learning models did not show monotonically increasing or decreasing relationships between dependent and explanatory variables. This monotony was often broken at the extremes of the range of values of the predictor variables, where the amount of data to train the models was lower (Fig. 1). Moreover, machine learning methods also showed differences in the importance assigned to different predictors. Thus, XGB identified some variables as very important compared to other predictors. Specifically, in the models trained with 15 variables, XGB showed a greater importance of precipitation in August, September and October, minimum

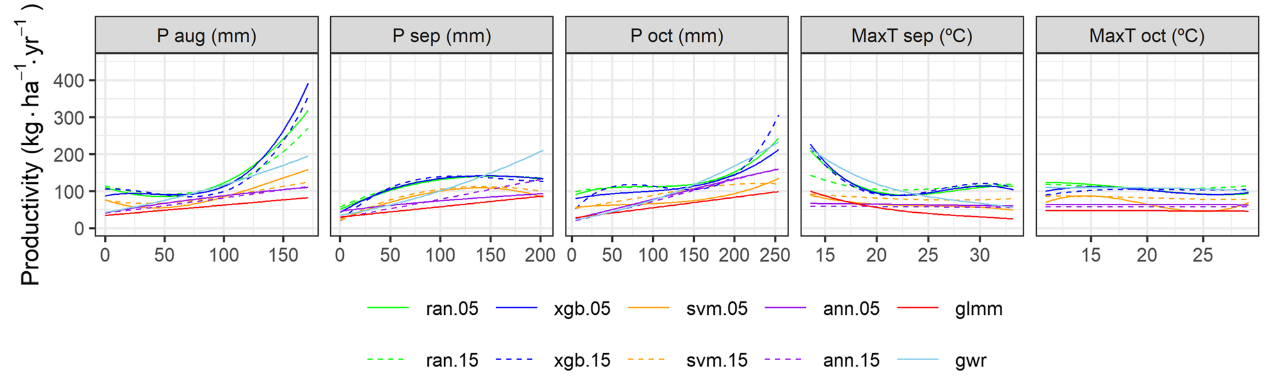

Fig. 1 Relationship between annual mushroom productivity and August, September and October precipitation and maximum temperatures in September and October (these variables are the variables used in the statistical models and the five variables machine learning models). ran (random forest), xgb (extreme gradient boosting), svm (support vector machine), ann (artificial neural network), glmm (generalized linear mixed models) and gwr (geographically weighted regression). 05 and 15 refer to the models trained with five and fifteen variables, respectively

temperature in October and aspect. In addition, precipitation of August resulted in having further greater importance in the models trained with five variables. Conversely, the importance detected by RF and SVM to the whole array of predictors was more homogeneous. RF showed a greater importance to the same variables as XGB, while the most important variables in SVM were precipitation in September and October, average temperature in August and September, and minimum temperature in August (Fig. 2).

GLMM and GWR fitted models and their coefficients are shown in the supplementary material in Table S1 to S3. Likewise, optimal machine learning hyperparameters can be found in the supplementary material in Tables S4 and S5.

Predictive accuracy of different methods

In general, ML models showed better predictive accuracy, in terms of MSE, compared to the statistical models. Within ML models, ANN models showed a higher error than the other algorithms. Decision tree-based models, namely, RF and XGB, showed no differences between the 15- and five-variable models when k-fold CV-based resampling was used. Using an environmental blocking, RF models, as well as SVM and ANN, showed lower accuracy when using five variables instead of 15. Contrary, the prediction error using XGB increased significantly when using 15 variables (instead of five) in the environmental CV, resulting in the lowest accuracy among all machine learning models and equaling the error of the statistical models. Decision tree-based models reduced significantly the bias between predicted and observed values, especially when conducting k-fold CV. On the other hand, the error estimated from a spatial CV with the SVM and ANN models trained with 15 variables was lower than in the five-variable models. Using a k-fold CV, the error of the SVM models was higher when using

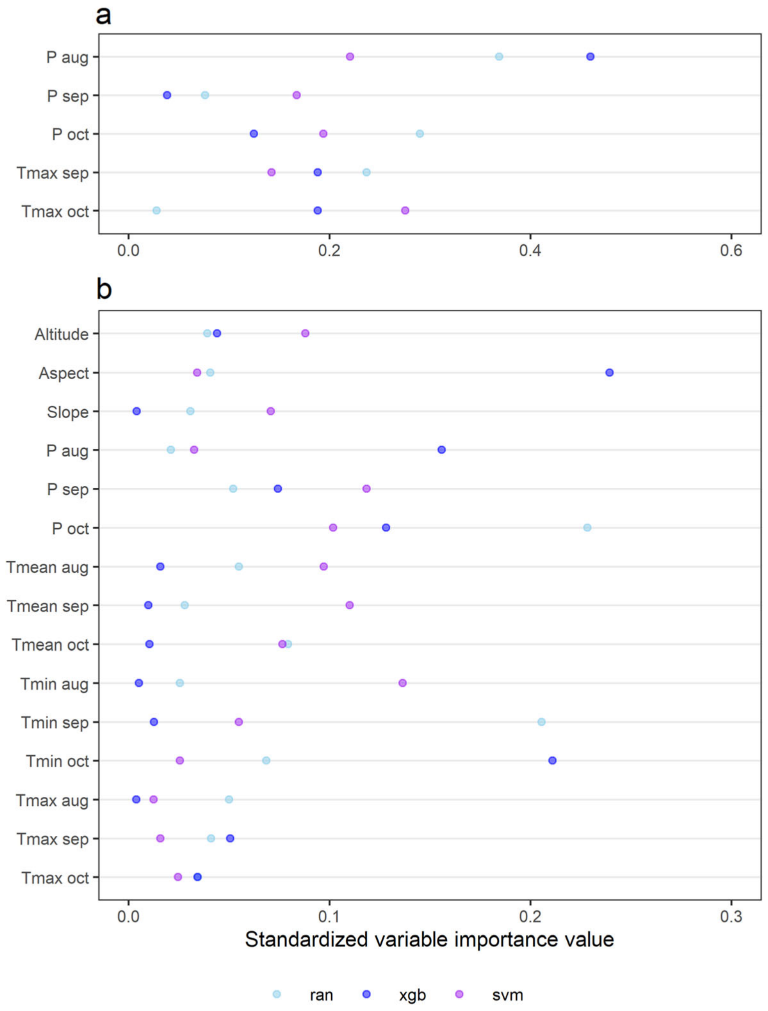

Fig. 2 Standardized variable importance value used to train random forest (ran), extreme gradient boosting (xgb) and support vector machine (svm) models with five **a** and 15 **b** variables. Variable importance values represent the contribution of each variable in the prediction of annual mushroom productivity

more predictors, whereas in the ANN models it was higher when using more parsimonious models. Although GWR improved GLMM predictive accuracy in the k-fold and spatial CV mainly due to a notable bias reduction, this was not the case in the environmental CV, where the error was similar for both statistical modeling approaches (Table 2).

Decision tree-based models increased the predictive accuracy reducing MSE by up to 10% compared to XGB, almost by 40% compared to ANN and GLMM, and by 20% compared to GWR when using k-fold CV. Similar trends were also obtained using environmental and spatial CV.

Mapping and accuracy of predictions at the landscape level

The spatially explicit predictions from each model at the landscape level resulted in rather similar general patterns between modeling approaches (Fig. 3). Namely, they predicted higher productivity in the northern areas of the study region, characterized by higher altitudes, i.e., Pyrenees mountain range. Also, the different models reproduced similar patterns within these areas according to variations in local topography. In addition, RF, XGB and SVM models trained with 15 predictors yielded higher estimates of mushroom productivity in coastal areas compared to the same algorithms based on a subset of five predictors. Those coastal areas represented the least similar bioclimatic conditions compared to the modeling data when using 12 predictors (Fig. 4 and Fig. S3), therefore increasing the area of extrapolation beyond the range of the modeling data. Specifically, the similarity map based on 12 predictors, shows that the number of pixels with low and medium similarity increased by 58% (359 km^2) and 50% (847 km^2), respectively, compared to the similarity map based on five predictors. On the other hand, pixels with high similarity decreased by 28% (1206 km^2).

Table 2 Mean squared error (MSE) and squared bias (bias2) of the different machine learning and statistical models depending on different resampling strategies, namely, k-fold, environmental, and spatial cross-validation. ran (random forest), xgb (extreme gradient boosting), svm (support vector machine), ann (artificial neural network), glmm (generalized linear mixed models) and gwr (geographically weighted regression). 05 and 15 refer to the models trained with five and fifteen variables, respectively

	Environmental cv		Spatial cv		k-fold cv	
	MSE	Bias^2	MSE	Bias^2	MSE	Bias^2
ran.05	22,941	88	18,096	37	12,677	1
ran.15	19,875	33	18,356	10	12,148	2
xgb.05	21,778	178	17,433	62	13,744	1
xgb.15	28,654	148	18,473	93	13,231	1
svm.05	30,901	2556	19,930	1226	14,140	657
svm.15	22,910	1214	21,032	797	12,824	392
ann.05	28,950	5206	25,021	4511	20,128	3087
ann.15	26,815	5619	29,516	8033	24,487	5946
glmm	28,318	9528	24,460	5228	21,086	3394
gwr	28,214	9789	20,590	2553	16,078	204

RF, XGB and SVM trained with 15 variables also resulted in less smoothed predictions of mushroom yield across the territory compared to estimates based on the subset of five predictors. Furthermore, SVM produced illogical predictions below 0 kg·ha^{-1}·year^{-1} in a few spatially localized areas when five variables were used, and scattered throughout the territory when using 15 predictors. In contrast, ANN resulted in very smoothed estimates across the territory, contrary to the maps obtained from all the other machine learning methods (Fig. 3).

In addition, mushroom productivity predictions based on RF, XGB, SVM and GWR ranged between 0, in the less productive areas, and approximately 300 and 400 kg·ha^{-1}·year^{-1} (with some maximum peaks reaching 500 and 600 kg·ha^{-1}·year^{-1}). Slightly lower productivity was detected using GLMM and ANN for the most productive sites, i.e. not exceeding 200 kg·ha^{-1}·year^{-1} in any point of the study area.

Discussion

To our knowledge, this is the first study addressing a systematic evaluation of the predictive performance of alternative statistical and machine learning models to predict fungal productivity, and one of the few systematic comparisons between these different predictive approaches within the field of ecological research. This was conducted using one of the largest datasets (if not the largest one) for fungal productivity monitoring, based on consistent sampling methodology and taxonomic identification of mushrooms over more than 20 years on nearly a hundred permanent sampling plots, randomly distributed throughout the study region, which contributes to overcoming most of the practical problems related to the existence of available data for modeling fungal resources (Hao et al. 2020).

When dealing with complex ecological interactions between multiple potential explanatory variables, our results show that statistical models, especially GLMM, clearly seem to have lower predictive performance compared to artificial intelligence-based approaches, in line with previous research (e.g. Smoliński and Radtke 2016 and Schratz et al. 2019). They were less accurate and produced large over- or underestimation of mushroom productivity (Table 2), making them unreliable for such purposes compared to other alternatives. On the other hand, statistical models can be good candidates for detecting the most appropriate

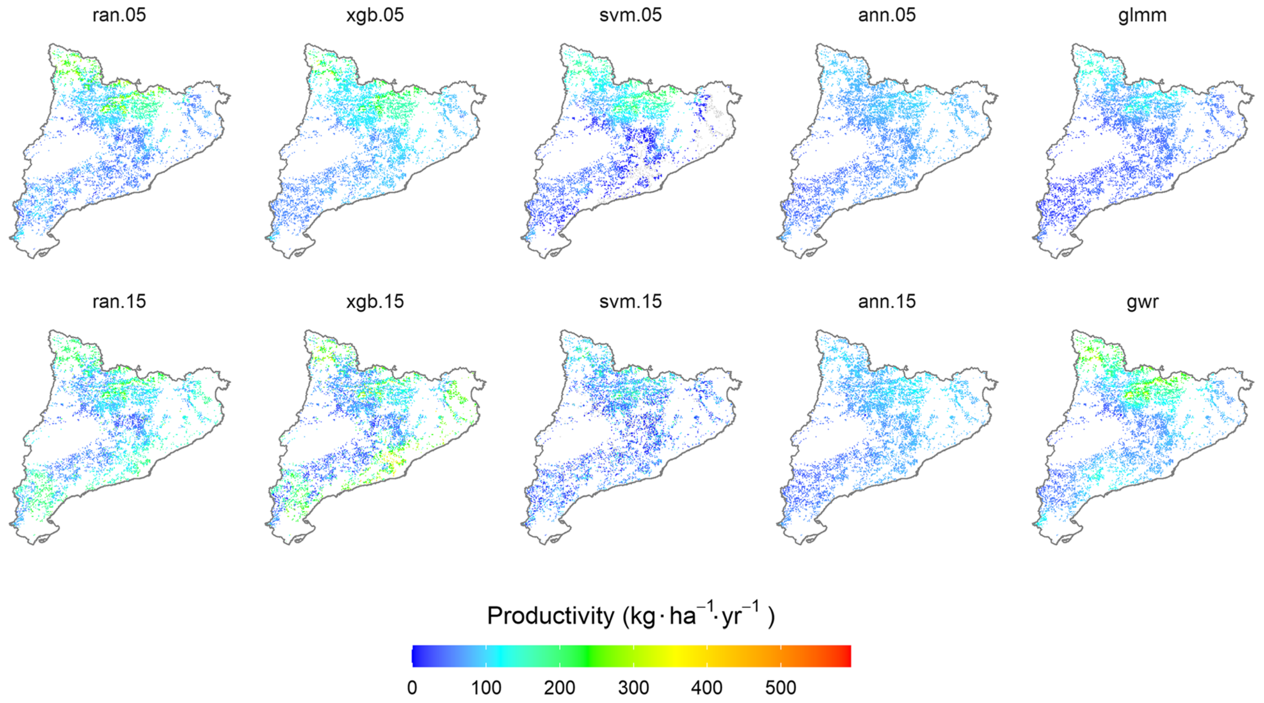

Fig. 3 Landscape-level prediction of total annual mushroom productivity, using ran (random forest), xgb (extreme gradient boosting), svm (support vector machine), ann (artificial neural network), glmm (generalized linear mixed models) and gwr (geographically weighted regression). 05 and 15 labels refer to the models trained with five and fifteen variables, respectively

variables to be used in machine learning models and unravel environmental-ecological relationships between them (Shmueli 2010; Schratz et al. 2019), since the inherent statistical assumptions that shape these models allow the relationships between data in a set of probability distributions to be correctly approximated. Fitting GWR parameters using a subset of data according to their geographical location corrected for the strong underestimation of fungal productivity produced by GLMM models using k-fold

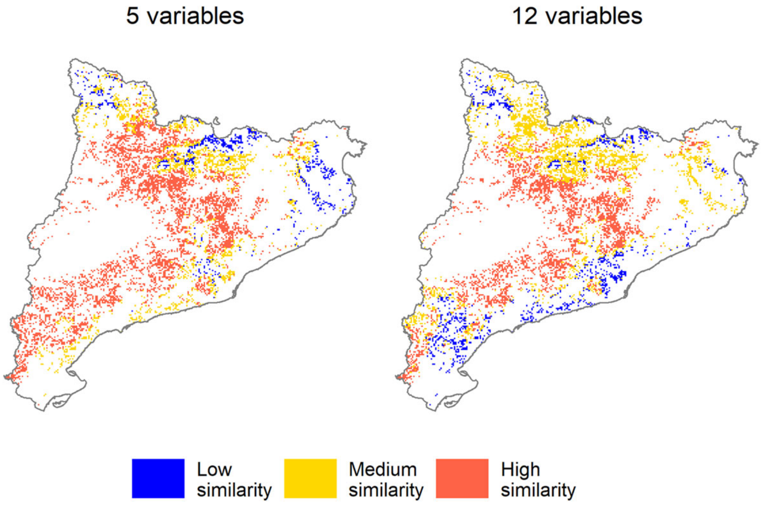

Fig. 4 Similarity between the climate conditions of the whole study area and the modeling data. The spatial similarity was based on the number of sampled data with an environment similar to the prediction environment. The per-pixel similarity was obtained by overlapping the pixel position and a density map of the sampled data in a two-dimensional space defined by the first two principal components of a principal component analysis (PCA). The PCA was performed with five and 12 variables according to the machine learning models' climate predictors (Fig. S3)

CV. However, it was not possible to correct for the bias in mushroom productivity prediction in environmentally differentiated areas. By considering spatial parameters, we were able to find non-stationary patterns across the territory, denoting that climatic conditions do not affect equally at a landscape level.

As demonstrated here, choosing a subset of variables from statistically significant predictors from statistical models can help us to deal with some drawbacks. A problem with selecting a single subset of variables from a machine learning models is that, due to the algorithm itself, the significance is adjusted differently and could be inappropriate for some of them. For example, within decision tree algorithms, XGB determines the variable to be used in each node of the tree among the total of variables of a model, while RF does it within a subset of them, giving greater probability of being chosen to those less important variables (Hastie et al. 2001). On the other hand, the importance of a set of correlated variables can be distributed among the different predictors (giving lower importance to each one of them), but the total importance that this set represents in the predictive performance is remarkable (Toloşi and Lengauer 2011). This can cause that when discarding the less important variables, this set of predictors is omitted, causing a notorious drop in predictive performance. Moreover, in a group of correlated variables where there is only one true predictor (the one that implies real causality), machine learning algorithms could give similar values of importance to the whole set of variables (Archer and Kimes, 2008), actually hiding the true predictor. These problems may be aggravated when using a larger number of variables to train the models, where the probability of finding groups of correlated variables is higher. Consequently, each machine learning algorithm give a different importance to each variable. Therefore, to identify the variables that could best explain the processes that occur in natural ecosystems and/or use the variable importance to select a subset of predictors to train a more parsimonious model, the above considerations should be taken into account.

The fact that the prediction error obtained from RF, SVM, and ANN models was lower when using 15 predictors in environmental blocking, suggests that models using a larger number of predictors may be a better alternative for predicting mushroom productivity at the landscape level. However, the combination of climatic conditions represented by model predictors increases exponentially with increasing number of variables (Hughes, 1968). This makes it more likely that increasing the number of model predictors will increase the mismatch between the modeling data and the climatic conditions across the whole study area. Therefore, models with a larger number of predictors will probably result in greater extrapolation beyond the range of the modeling data, as shown in our study (Fig. 4 and Fig. S3). Thus, in the 15-variable models, extrapolation beyond the range of modeling data occurred over a larger extent of the study area compared to the models based on five predictors. Assuming that k-fold CV estimates model accuracy in areas where climatic conditions are similar to the modeling data, while environmental or spatial CV estimates model prediction error in climatically different areas (Roberts et al. 2017; Meyer et al. 2019a), the assessment of model accuracy for prediction across the study area can be improved based on the similarity in the climatic conditions between the modeling data and the whole study area. Thus, random blocking with five-predictor models informed more appropriately about the magnitude of the prediction error over a larger area compared to 15-predictor models, because areas with high similarity increased when using fewer predictors, i.e. from ~ 2800 to ~ 4000 km^2. Conversely, the prediction error of the less parsimonious models will be given by an environmental blocking in a smaller area than in the models with less predictors. To assess which model is more suitable for prediction, one needs to consider the extent of the study area where the prediction error is quantified through random blocking and through environmental blocking, respectively, and not only whether the model error of more or less parsimonious models is higher or lower in each blocking strategy. In our study, when using models with 15 predictors, the entire coastal areas (east) and the Pyrenees mountain range (north), showed a low to moderate similarity of climatic conditions compared to the modeling data. In contrast, in the 5-predictor models, the coastal areas with low similarity decreased, while the area with high similarity of the Pyrenean mountain range increased considerably. Thus, it seems that parsimony may be a useful model selection criterion not only for statistical methods but also for machine learning algorithms (Coelho et al. 2018).

As noted, statistical models do not seem to be competitive compared to machine learning approaches due to poor predictive performance. Among the machine learning models, the ANN approach had the highest prediction error and also resulted in biogeographic patterns that did not seem to agree with the expected climatic variations throughout the study area. In turn, SVM yielded illogical negative values of mushroom productivity in some areas. Therefore, the best candidate methods are the decision trees-based algorithms, i.e. RF and XGB. Considering the similarities in the climatic conditions between the modeling data and the whole study area, we conclude that the best models will be RF and XGB models trained with five predictors. This study shows that, although machine learning algorithms allow to train models using a large number of variables, it may be wise to conduct a more thorough selection of model predictors prior to training the final models (Kuhn and Johnson 2013). This further contributes to improving the selection of the best modeling approach for prediction and also provides a methodology that, in the face of the current paucity of data to build process-based models (Hao et al. 2020), can be reasonably used in extrapolation. This is especially relevant in a context of global change, where climatic conditions are predicted to change over the years beyond the historical climatic ranges.

Conclusions

This study compares different statistical and machine learning models for predicting fungal productivity biogeographical patterns using a systematic methodology. Decision tree-based models, namely, RF and XGB, performed the best in the prediction of fungal productivity in both environmentally similar and differentiated areas. Therefore, we recommend the use of these algorithms for further research involving the prediction of fungal productivity, both under the current bioclimatic conditions and under climate change scenarios. When using these methods, careful selection of predictors allows for defining more interpretable and computationally less expensive models as well as for reducing the environmental space described by model predictors. Accordingly, the range of environmental conditions represented by the predictors in the modeling data can be more similar to the conditions over the whole study area, leading to reduced extrapolation. As a result, predictions can be more ecologically consistent compared to models with much higher number of predictors. In this regard, the degree of similarity in the range of environmental conditions between the modeling data and the whole study area for prediction is relevant when selecting the most appropriate blocking strategy for estimating model error. In more parsimonious models, where the range of the modeling data may be more representative of the environmental conditions over the whole study area compared to more complex models, the magnitude of the prediction error at the landscape level may be better retrieved through random blocking. In contrast, increasing model complexity may require environmental blocking for a more proper characterization of the prediction error at the landscape level. Model and variable selection should therefore also consider the extent of the area within the study region where the magnitude of the prediction error can be quantified more appropriately from either environmental or random blocking. Maps depicting the similarity between the environmental conditions accounted by the modeling data compared to the environmental conditions of the whole study area, can be useful to identify environmentally different or similar areas to further assist model selection and proper characterization of the prediction error based on alternative resampling techniques. In the end, given the multiple environmental factors driving fungal productivity, we highlight the importance of applying such methods using high-resolution environmental information to properly estimate its biogeographic patterns over large scales.

Abbreviations

ANN: Artificial Neural Network; bias2: Squared bias; GLMM: Generalized Linear Mixed Models; GWR: Geographically Weighted Regression; MAE: Mean Absolute Error; PCA: Principal Component Analysis; PDP: Partial Dependence Plots; RF: Random Forest; MSE: Mean Squared Error; SVM: Support Vector Machine; XGB: Extreme Gradient Boosting

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s40663-021-00297-w.

Additional file 1: Table S1. Fixed GLMM coefficients. β_1, β_2, β_3 significance were calculated from *t*-value, while β_3, β_4, β_5 from *p*-value. **Table S2.** Random GLMM coefficients. **Table S3.** GWR models coefficients. **Table S3.** Summary of mushroom, climate and physical data of the 98 sampled plots. All these variables were used to train the 15-variable machine learning models, while the 5-variable machine learning models and the statistical models only used those marked with *. The response variable is shown with **. **Table S4.** Tuned optimal hyperparameters using a k-fold CV. **Table S5.** Tuned optimal hyperparameters using an environmental CV. **Figure S1.** Study area, distribution of mushroom productivity monitoring plots (red points) and pine forest ecosystems represented by the sample plots (green area) Coordinates system: WGS 84 / UTM zone 31 N. **Figure S2.** GWR coefficient estimates according to geographical location. Coefficient of precipitation amount from August to October (A) and maximum temperature in August (B) in conditioned production model. Coefficient of precipitation amount from August to October (A) and maximum temperature in October (D) in occurrence model (C). **Figure S3.** Similarity in climatic conditions between the modeling data and the whole study area using five or 12 variables. Two-dimensional representation given by the two principal components, namely, PC1 and PC2, of a principal component analysis (PCA) of the modeling data (with a density map) and the environmental conditions over the whole study area (gray dots).

Acknowledgements

Not applicable.

Authors' contributions

JMdA, JAB and SdM contributed in the installation of the sampling plots and data collection. AM, SdM and JL analysed the data. AM wrote the manuscript. All authors participated in the review and editing of the manuscript. SdM supervised the work during the whole process. The authors read and approved the final manuscript.

Funding

This work was supported by the Secretariat for Universities and of the Ministry of Business and Knowledge of the Government of Catalonia and the European Social Fund. This work was partially supported by the Spanish Ministry of Science, Innovation and Universities (Grant No. RTI2018–099315-A-I00). J.A.B. benefitted from a Serra-Húnter Fellowship provided by the Government of Catalonia.

Availability of data and materials

The datasets generated and/or analyzed during the current study are not publicly available due to legal constraints because they are owned by different institutions, but are available from the authors on reasonable request.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Author details

[1]Department of Crop and Forest Sciences, University of Lleida, Av. Alcalde Rovira Roure 191, E-25198 Lleida, Spain. [2]Joint Research Unit CTFC-AGROTECNIO-CERCA Center, Av. Rovira Roure 191, 25198 Lleida, Spain. [3]Forest Science and Technology Centre of Catalonia, Ctra. Sant Llorenç de Morunys km 2, 25280 Solsona, Spain. [4]Forest Advanced Computing and Artificial Intelligence Laboratory, Department of Forestry and Natural Resources, Purdue University, West Lafayette, IN 47907, USA.

Received: 27 November 2020 Accepted: 22 February 2021
Published online: 15 March 2021

References

Allaire JJ, Chollet F (2019) Keras: R Interface to 'Keras'. R package version 2.2.5.0. https://CRAN.R-project.org/package=keras. Accessed 10 Nov 2020

Alday JG, Martínez de Aragón J, de-Miguel S, Bonet JA (2017) Mushroom biomass and diversity are driven by different spatio-temporal scales along Mediterranean elevation gradients. Sci Rep 7(1). https://doi.org/10.1038/srep45824

Archer KJ, Kimes RV (2008) Empirical characterization of random forest variable importance measures. Comput Stat Data Anal 52(4):2249–2260. https://doi.org/10.1016/j.csda.2007.08.015

Bahn V, McGill BJ (2012) Testing the predictive performance of distribution models. Oikos 122(3):321–331. https://doi.org/10.1111/j.1600-0706.2012.00299.x

Barnard RL, Osborne CA, Firestone MK (2014) Changing precipitation pattern alters soil microbial community response to wet-up under a Mediterranean-type climate. ISME J 9(4):946–957. https://doi.org/10.1038/ismej.2014.192

Bastin J-F, Finegold Y, Garcia C, Mollicone D, Rezende M, Routh D, Constantin MZ, Crowther TW (2019) The global tree restoration potential. Science 365(6448):76–79. https://doi.org/10.1126/science.aax0848

Bates D, Mächler M, Bolker B, Walker S (2015) Fitting linear mixed-effects models using lme4. J Stat Softw 67:1–48. https://doi.org/10.18637/jss.v067.i01

Becker M, Lang M, Richter J, Bischl B, Schalk D (2020) mlr3tuning: Tuning for 'mlr3'. R package version 0.5.0. https://CRAN.Rproject.org/package=mlr3tuning

Bergstra J, Bengio Y (2012) Random search for hyper-parameter optimization. J Mach. Learn Res 13:281–305

Bivand R, Yu D (2017) Spgwr: geographically weighted regression. R Pack Version 0.6–32. https://CRAN.R-project.org/package=spgwr. Accessed 10 Nov 2020

Boa E (2004) Wild edible fungi: a global overview of their use and importance to people (non-Wood Forest products no. 17). FAO. Forestry Department, Rome, p 148. ISBN: 92-5-105157-7

Bonet JA, Fischer CR, Colinas C (2004) The relationship between forest age and aspect on the production of sporocarps of ectomycorrhizal fungi in *Pinus sylvestris* forests of the Central Pyrenees. Forest Ecol Manag 203(1–3):157–175. https://doi.org/10.1016/j.foreco.2004.07.063

Bonet JA, Palahí M, Colinas C, Pukkala T, Fischer CR, Miina J, Martínez de Aragón J (2010) Modelling the production and species richness of wild mushrooms in pine forests of the Central Pyrenees in northeastern Spain. Can J For Res 40(2):347–356. https://doi.org/10.1139/x09-198

Bonete IP, Arce JE, Figueiredo Filho A, Retslaff FA de S, Lanssanova LR (2020) Artificial neural networks and mixed-effects modeling to describe the stem profile of *Pinus taeda* L. Floresta 50(1):1123. doi:https://doi.org/10.5380/rf.v50i1.61764

Büntgen U, Kauserud H, Egli S (2012) Linking climate variability to mushroom productivity and phenology. Front Ecol Environ 10(1):14–19. https://doi.org/10.1890/110064

Büntgen U, Peter M, Kauserud H, Egli S (2013) Unraveling environmental drivers of a recent increase in Swiss fungi fruiting. Glob Chang Biol 19(9):2785–2794. https://doi.org/10.1111/gcb.12263

Chen T, He T, Benesty M, Khotilovich V, Tang Y, Cho H, Chen K, Mitchell R, Cano I, Zhou T, Li M, Xie J, Lin M, Geng Y, Li Y (2019) Xgboost: extreme gradient boosting. R package version 0.90.0.2. https://CRAN.R-project.org/package=xgboost. Accessed 10 Nov 2020

Christin S, Hervet É, Lecomte N (2019) Applications for deep learning in ecology. Methods Ecol Evol 10:1632–1644. https://doi.org/10.1111/2041-210X.13256

Coelho MTP, Diniz-Filho JA, Rangel TF (2018) A parsimonious view of the parsimony principle in ecology and evolution. Ecography. https://doi.org/10.1111/ecog.04228

Collado E, Bonet JA, Camarero JJ, Egli S, Peter M, Salo K, Martínez-Peña F, Ohenoja E, Martín-Pinto P, Primicia I, Büntgen U, Kurttila M, Oria-de-Rueda JA, Martínez-de-Aragón J, Miina J, de-Miguel S (2019) Mushroom productivity trends in relation to tree growth and climate across different European forest biomes. Sci Total Environ. https://doi.org/10.1016/j.scitotenv.2019.06.471

Collado E, Camarero JJ, Martínez de Aragón J, Pemán J, Bonet JA, de-Miguel S (2018) Linking fungal dynamics, tree growth and forest management in a Mediterranean pine ecosystem. Forest Ecol Manag 422:223–232. https://doi.org/10.1016/j.foreco.2018.04.025

Commission of the European Community (1991) CORINE biotopes manual – habitats of the European Community. DG Environment, Nuclear Safety and Civil Protection, Luxembourg

Cortez P (2016) rminer: data mining classification and regression methods. R package version 1.4.2. https://CRAN.Rproject.org/package=rminer

Cutler DR, Edwards TC, Beard KH, Cutler A, Hess KT, Gibson J, Lawler JJ (2007) Random forests for classification in ecology. Ecology 88:2783–2792. https://doi.org/10.1890/07-0539.1

De Cáceres M, Martin-StPaul N, Turco M, Cabon A, Granda V (2018) Estimating daily meteorological data and downscaling climate models over landscapes. Environ Model Softw 108:186–196. https://doi.org/10.1016/j.envsoft.2018.08.003

de-Miguel S, Bonet JA, Pukkala T, Martínez de Aragón J (2014) Impact of forest management intensity on landscape-level mushroom productivity: a regional model-based scenario analysis. Forest Ecol Manag 330:218–227. https://doi.org/10.1016/j.foreco.2014.07.014

Diamantopoulou MJ, Özçelik R, Crecente-Campo F, Eler Ü (2015) Estimation of Weibull function parameters for modelling tree diameter distribution using least squares and artificial neural networks methods. Biosyst Eng 133:33–45. https://doi.org/10.1016/j.biosystemseng.2015.02.013

Dixon PA, Milicich MJ, Sugihara G (1999) Episodic fluctuations in larval supply. Science 283(5407):1528–1530. https://doi.org/10.1126/science.283.5407.1528

Duarte E, Wainer J (2017) Empirical comparison of cross-validation and internal metrics for tuning SVM hyperparameters. Pattern Recognit Lett. 88:6–11. https://doi.org/10.1016/j.patrec.2017.01.007

Ehrlén J, Morris WF (2015) Predicting changes in the distribution and abundance of species under environmental change. Ecol Lett 18(3):303–314. https://doi.org/10.1111/ele.12410

Gange AC, Gange EG, Sparks TH, Boddy L (2007) Rapid and recent changes in fungal fruiting patterns. Science 316(5821):71. https://doi.org/10.1126/science.1137489

Gasch CK, Hengl T, Gräler B, Meyer H, Magney TS, Brown DJ (2015) Spatio-temporal interpolation of soil water, temperature, and electrical conductivity in 3D + T: the cook agronomy farm data set. Spat Stat 14:70–90. https://doi.org/10.1016/j.spasta.2015.04.001

Georganos S, Abdi AM, Tenenbaum DE, Kalogirou S (2017) Examining the NDVI-rainfall relationship in the semi-arid Sahel using geographically weighted regression. J Arid Environ 146:64–74. https://doi.org/10.1016/j.jaridenv.2017.06.004

Glassman SI, Wang IJ, Bruns TD (2017) Environmental filtering by pH and soil nutrients drives community assembly in fungi at fine spatial scales. Mol Ecol 26:6960–6973. https://doi.org/10.1111/mec.14414

Gobeyn S, Mouton AM, Cord AF, Kaim A, Volk M, Goethals PLM (2019) Evolutionary algorithms for species distribution modelling: a review in the context of machine learning. Ecol Model 392:179–195. https://doi.org/10.1016/j.ecolmodel.2018.11.013

Görgens EB, Montaghi A, Rodriguez LCE (2015) A performance comparison of machine learning methods to estimate the fast-growing forest plantation yield based on laser scanning metrics. Comput Electron Agric 116:221–227. https://doi.org/10.1016/j.compag.2015.07.004

Hamilton DA Jr, Brickell JE (1983) Modeling methods for a two-state system with continuous responses. Can J For Res 13(6):1117–1121. https://doi.org/10.1139/x83-149

Hannemann H, Willis KJ, Macias-Fauria M (2015) The devil is in the detail: unstable response functions in species distribution models challenge bulk ensemble modelling. Glob Ecol Biogeogr 25(1):26–35. https://doi.org/10.1111/geb.12381

Hao T, Guillera-Arroita G, May TW, Lahoz-Monfort JJ, Elith J (2020) Using species distribution models for fungi. Fung Biol Rev. https://doi.org/10.1016/j.fbr.2020.01.002

Hastie T, Tibshirani R, Friedman J (2001) The elements of statistical learning: data mining, inference, and prediction. New York: Springer-Verlag. ISBN 0-387-95284-5

Hill L, Hector A, Hemery G, Smart S, Tanadini M, Brown N (2017) Abundance distributions for tree species in Great Britain: a two-stage approach to modeling abundance using species distribution modeling and random forest. Ecol Evol 7:1043–1056. https://doi.org/10.1002/ece3.2661

Hughes G (1968) On the mean accuracy of statistical pattern recognizers. IEEE Trans Inf Theory 14(1):55–63. https://doi.org/10.1109/tit.1968.1054102

Hutter F, Hoos HH, Leyton-Brown K (2011) Sequential model-based optimization for general algorithm configuration. Learn Intell Optim. 507–523. https://doi.org/10.1007/978-3-642-25566-3_40

Juel A, Groom GB, Svenning J-C, Ejrnæs R (2015) Spatial application of random forest models for fine-scale coastal vegetation classification using object based analysis of aerial orthophoto and DEM data. Int J Appl Earth Obs Geoinf 42:106–114. https://doi.org/10.1016/j.jag.2015.05.008

Karavani A, De Cáceres M, Martínez de Aragón J, Bonet JA, de-Miguel S (2018) Effect of climatic and soil moisture conditions on mushroom productivity and related ecosystem services in Mediterranean pine stands facing climate change. Agric Forest Meteorol 248:432–440. doi:https://doi.org/10.1016/j.agrformet.2017.10.024

Kauserud H, Stige LC, Vik JO, Okland RH, Hoiland K, Stenseth NC (2008) Mushroom fruiting and climate change. PNAS 105(10):3811–3814. https://doi.org/10.1073/pnas.0709037105

Kauserud H, Heegaard E, Semenov MA, Boddy L, Halvorsen R, Stige LC, Sparks TH, Gange AC, Stenseth NC (2009) Climate change and spring-fruiting fungi. Proc R Soc B Biol Sci 277:1169–1177. https://doi.org/10.1098/rspb.2009.1537

Kuhn M, Johnson K (2013) Applied predictive modeling. Springer, New York. https://doi.org/10.1007/978-1-4614-6849-3

Lang M, Binder M, Richter J, Schratz P, Pfisterer F, Coors, S, Au, Q, Casalicchio, G, Kotthoff, L, Bischl, B (2019) mlr3: a modern objectoriented machine learning framework in R. J Open Source Softw. https://doi.org/10.21105/joss.01903

Lang M, Au Q, Coors S, Schratz P (2020a) mlr3learners: recommended learners for 'mlr3'. R package version 0.4.3. https://CRAN.Rproject.org/package=mlr3learners

Lang M, Bischl B, Richter J, Sun X, Binder M (2020b). paradox: define and work with parameter spaces for complex algorithms. R package version 0.6.0. https://CRAN.R-project.org/package=paradox

Le Rest K, Pinaud D, Monestiez P, Chadoeuf J, Bretagnolle V (2014) Spatial leave-one-out cross-validation for variable selection in the presence of spatial autocorrelation. Glob Ecol Biogeogr 23(7):811–820. https://doi.org/10.1111/geb.12161

Liang J, Crowther TW, Picard N, Wiser S, Zhou M, Alberti G, Schulze ED, McGuire AD, Bozzato F, Pretzsch H, De-Miguel S, Paquette A, Herault B, Scherer-Lorenzen M, Barrett CB, Glick HB, Hengeveld GM, Nabuurs GJ, Pfautsch S, Viana H, Vibrans AC, Ammer C, Schall P, Verbyla D, Tchebakova N, Fischer M, Watson JV, HYH C, Lei XD, Schelhaas MJ, Lu HC, Gianelle D, Parfenova EI, Salas C, Lee E, Lee B, Kim HS, Bruelheide H, Coomes DA, Piotto D, Sunderland T, Schmid B, Gourlet-Fleury S, Sonke B, Tavani R, Zhu J, Brandl S, Vayreda J, Kitahara F, Searle EB, Neldner VJ, Ngugi MR, Baraloto C, Frizzera L, Balazy R, Oleksyn J, Zawila-Niedzwiecki T, Bouriaud O, Bussotti F, Finer L, Jaroszewicz B, Jucker T, Valladares F, Jagodzinski AM, Peri PL, Gonmadje C, Marthy W, O'Brien T, Martin EH, Marshall AR, Rovero F, Bitariho R, Niklaus PA, Alvarez-Loayza P, Chamuya N, Valencia R, Mortier F, Wortel V, Engone-Obiang NL, Ferreira LV, Odeke DE, Vasquez RM, Lewis SL, Reich PB (2016) Positive biodiversity-productivity relationship predominant in global forests. Science 354(6309):aaf8957. https://doi.org/10.1126/science.aaf8957

Martínez de Aragón J, Bonet JA, Fischer CR, Colinas C (2007) Productivity of ectomycorrhizal and selected edible saprotrophic fungi in pine forests of the pre-Pyrenees mountains, Spain: predictive equations for forest management of mycological resources. Forest Ecol Manag 252(1–3):239–256. https://doi.org/10.1016/j.foreco.2007.06.040

Mayfield H, Smith C, Gallagher M, Hockings M (2020) Considerations for selecting a machine learning technique for predicting deforestation. Environ Model Softw:104741. https://doi.org/10.1016/j.envsoft.2020.104741

Meyer H, Reudenbach C, Hengl T, Katurji M, Nauss T (2018) Improving performance of spatio-temporal machine learning models using forward feature selection and target-oriented validation. Environ Model Softw 101:1–9. https://doi.org/10.1016/j.envsoft.2017.12.001

Meyer D, Dimitriadou E, Hornik K, Weingessel A, Leisch F (2019b) e1071: Misc functions of the Department of Statistics, probability theory group (formerly: E1071), TU Wien. R Pack Version 1.7–2. https://CRAN.R-project.org/package=e1071. Accessed 10 Nov 2020

Meyer H, Reudenbach C, Wöllauer S, Nauss T (2019a) Importance of spatial predictor variable selection in machine learning applications - moving from data reproduction to spatial prediction. Ecol Model 411:108815. https://doi.org/10.1016/j.ecolmodel.2019.108815

Micheletti N, Foresti L, Robert S, Leuenberger M, Pedrazzini A, Jaboyedoff M, Kanevski M (2013) Machine learning feature selection methods for landslide susceptibility mapping. Math Geosci 46(1):33–57. https://doi.org/10.1007/s11004-013-9511-0

Mohan JE, Cowden CC, Baas P, Dawadi A, Frankson PT, Helmick K, Hughes E, Khan S, Lang A, Machmuller M, Taylor M, Witt CA (2014) Mycorrhizal fungi mediation of terrestrial ecosystem responses to global change: mini-review. Fungal Ecol 10:3–19. https://doi.org/10.1016/j.funeco.2014.01.005

Özçelik R, Diamantopoulou MJ, Crecente-Campo F, Eler U (2013) Estimating Crimean juniper tree height using nonlinear regression and artificial neural network models. Forest Ecol Manag 306:52–60. https://doi.org/10.1016/j.foreco.2013.06.009

Pfisterer F, Poon J, Lang M (2021) mlr3keras: mlr3 Keras extension. R package version 0.1.3. https://github.com/mlr-org/mlr3keras

Pinheiro JC, Bates DM (2000) Mixed-effects models in S and S-PLUS, 1st edn. Springer, New York. https://doi.org/10.1007/b98882

Pohjankukka J, Pahikkala T, Nevalainen P, Heikkonen J (2017) Estimating the prediction performance of spatial models via spatial k-fold cross validation. Int J Geogr Inf Sci 31(10):2001–2019. https://doi.org/10.1080/13658816.2017.1346255

Prasad AM (2018) Machine learning for macroscale ecological niche modeling - a multi-model, multi-response ensemble technique for tree species management under climate change. Mach Learn Ecol Sust Nat Res Manag: 123–139. https://doi.org/10.1007/978-3-319-96978-7_6

Prasad A, Iverson L, Liaw A (2006) Newer classification and regression tree techniques: bagging and random forests for ecological prediction. Ecosystems 9(2):181–199. https://doi.org/10.1007/s10021-005-0054-1

Propastin PA (2009) Spatial non-stationarity and scale-dependency of prediction accuracy in the remote estimation of LAI over a tropical rainforest in Sulawesi, Indonesia. Remote Sens Environ 113(10):2234–2242. https://doi.org/10.1016/j.rse.2009.06.007

Roberts DR, Bahn V, Ciuti S, Boyce MS, Elith J, Guillera-Arroita G, Dormann CF (2017) Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. Ecography 40(8):913–929. https://doi.org/10.1111/ecog.02881

Sánchez-González M, de-Miguel S, Martin-Pinto P, Martínez-Peña F, Pasalodos-Tato M, Oria-de-Rueda JA, Martínez de Aragón J, Canellas I, Bonet JA (2019) Yield models for predicting aboveground ectomycorrhizal fungal productivity in *Pinus sylvestris* and *Pinus pinaster* stands of northern Spain. Forest Ecosyst 6(1):52. https://doi.org/10.1186/s40663-019-0211-1

Schratz P, Becker M (2021) mlr3spatiotempcv: spatiotemporal resampling methods for 'mlr3'. R package versión 0.1.1. https://CRAN.Rproject.org/package=mlr3spatiotempcv

Schratz P, Muenchow J, Iturritxa E, Richter J, Brenning A (2019) Hyperparameter tuning and performance assessment of statistical and machine-learning algorithms using spatial data. Ecol Model 406:109–120. https://doi.org/10.1016/j.ecolmodel.2019.06.002

Shmueli G (2010) To explain or to predict? Stat Sci 25(3):289–310. https://doi.org/10.1214/10-STS330

Sing T, Sander O, Beerenwinkel N, Lengauer T (2005) ROCR: visualizing classifier performance in R. Bioinformatics 21(20):7881. http://rocr.bioinf.mpi-sb.mpg.de. Accessed 10 Nov 2020

Smoliński S, Radtke K (2016) Spatial prediction of demersal fish diversity in the Baltic Sea: comparison of machine learning and regression-based techniques. ICES J Marine Sci. https://doi.org/10.1093/icesjms/fsw136

Snowdon P (1991) A ratio estimator for bias correction in logarithmic regressions. Can J For Res 21(5):720–724. https://doi.org/10.1139/x91-101

Stojanova D, Panov P, Gjorgjioski V, Kobler A, Džeroski S (2010) Estimating vegetation height and canopy cover from remotely sensed data with machine learning. Ecol Inf 5(4):256–266. https://doi.org/10.1016/j.ecoinf.2010.03.004

Stokland JN, Siitonen J, Jonsson BG (2012) Biodiversity in dead Wood, biodiversity in dead Wood. Cambridge University Press, Cambridge. https://doi.org/10.1017/CBO9781139025843

Taye ZM, Martínez-Peña F, Bonet JA, Martínez de Aragón J, de-Miguel S (2016) Meteorological conditions and site characteristics driving edible mushroom production in *Pinus pinaster* forests of Central Spain. Fungal Ecol 23:30–41. https://doi.org/10.1016/j.funeco.2016.05.008

Thessen A (2016) Adoption of machine learning techniques in ecology and earth science. One Ecosyst 1:e8621. https://doi.org/10.3897/oneeco.1.e8621

Thornton PE, Running SW (1999) An improved algorithm for estimating incident daily solar radiation from measurements of temperature, humidity, and precipitation. Agric Forest Meteorol 93(4):211–228. https://doi.org/10.1016/s0168-1923(98)00126-9

Thornton PE, Running SW, White MA (1997) Generating surfaces of daily meteorological variables over large regions of complex terrain. J Hydrol 190(3–4):214–251. https://doi.org/10.1016/s0022-1694(96)03128-9

Thuiller W (2003) BIOMOD - optimizing predictions of species distributions and projecting potential future shifts under global change. Glob Change Biol 9(10):1353–1362. https://doi.org/10.1046/j.1365-2486.2003.00666.x

Toloşi L, Lengauer T (2011) Classification with correlated features: unreliability of feature ranking and solutions. Bioinformatics 27(14):1986–1994. https://doi.org/10.1093/bioinformatics/btr300

Valavi R, Elith J, Lahoz-Monfort JJ, Guillera-Arroita G (2018) blockCV: an R package for generating spatially or environmentally separated folds for k-fold cross-validation of species distribution models. Method Ecol Evol. doi:https://doi.org/10.1111/2041-210x.13107

Venables WN, Ripley BD (2002) Modern applied statistics with S, 4th edn. Springer, New York. ISBN 0-387-95457-0

Wheeler DC, Páez A (2009) Geographically weighted regression. In: Fischer M, Getis A (eds) Handbook of Applied Spatial Analysis. Springer, Berlin. https://doi.org/10.1007/978-3-642-03647-7_22

Wood SN, Thomas MB (1999) Super-sensitivity to structure in biological models. Proc R Soc Lond B Biol Sci 266(1419):565–570. https://doi.org/10.1098/rspb.1999.0673

Wright MN, Ziegler A (2017) Ranger: a fast implementation of random forests for high dimensional data in C++ and R. J Stat Softw 77(1), 1-17. Doi:https://doi.org/10.18637/jss.v077.i01

Ye H, Beamish RJ, Glaser SM, Grant SC, Hsieh C, Richards LJ, Schnute JT, Sugihara G (2015) Equation-free mechanistic ecosystem forecasting using empirical dynamic modeling. PNAS 112:E1569–E1576. https://doi.org/10.1073/pnas.1417063112