DOI: 10.24425/academiaPAS.2023.147459

Modeling Tree Species With Random Forests

Machine learning methods, such as the random forests algorithm, have revolutionized how we analyze growing volumes of data. The algorithm can be usefully applied in studying… real forests.

Łukasz Pawlik

Institute of Earth Sciences,
Faculty of Natural Sciences,
University of Silesia in Katowice

Marcin K. Dyderski

Institute of Dendrology,
Polish Academy of Sciences in Kórnik

Łukasz Pawlik, PhD, DSc is a geographer and geomorphologist, a professor at the University of Silesia working at the Institute of Earth Sciences, University of Silesia. His research explores the impact of biotic factors and natural disturbances on geomorphological processes, landforms, and the evolution of montane soils.

lukasz.pawlik@us.edu.pl

Marcin K. Dyderski, PhD, DSc is an Associate Professor at the Institute of Dendrology, Polish Academy of Sciences. His research focuses on plant responses to human activities, including mining, forestry, climate change, and biological invasions, especially invasive tree species.

mdyderski@man.poznan.pl

Machine learning is now finding more and more applications in various scientific fields, industries, and services. It represents humanity's dream of creating a learning-machine system capable of recognizing patterns, akin to the human brain's sophisticated abilities.

The quest to glean knowledge from data has been evident ever since humankind first began to systematically observe natural phenomena. Around 3,500 years ago, the ancient Egyptians identified the rhythm of the Nile floods, developing a simplified mathematical model of the phenomenon to use in planning planting and harvesting. Over time, the volume of data at humanity's disposal has grown rapidly, and conventional statistical methods, often coming with stringent assumptions, have often fallen short of providing the expected solutions. In today's era of multidimensional data streams and Big Data, advanced algorithms, supercomputing centers, and computing clusters assist in "data mining" – sifting through these mountains of data.

And yet our struggle to cope with data, it seems, is still just beginning, and the tension between the influx of new information and users' growing expectations could be a potential flashpoint in the development of machine learning. Some have spoken critically of these issues, sometimes garnering significant media attention. One significant voice in the ongoing discussion about the threats posed by artificial intelligence is Cathy O'Neil's book *Weapons of Math Destruction*, where she asserts: "To create a model, then, we make choices about what's important enough to include, simplifying the world into a toy version that can be easily understood and from which we can infer important facts and actions." Such simplifications can, in certain cases, have negative consequences, such as African Americans being discriminated against in

```
RF_andGLM_roc_auc_plots_training_set <- ggplot()+
  geom_roc(model33RFtrain,
           mapping = aes(m = dam, d = factor(obs, levels = c('no_dam', 'dam')),
                         color = 'RF model 33'), n.cuts=0,
           increasing =        )+
  geom_roc(model32RFtrain,
           mapping = aes(m = dam, d = factor(obs, levels = c('no_dam', 'dam')),
                         color = 'RF model 32'), n.cuts=0,
           increasing =        )+
  geom_roc(model43RFtrain,
           mapping = aes(m = dam, d = factor(obs, levels = c('no_dam', 'dam')),
                         color = 'RF model 43'), n.cuts=0,
           increasing =        )+
  ggplot2::annotate("text", size = 3, x = .75, y = .45,
                    color = 'brown',
                    label = paste("RF model 33 AUC = ", round(max(model33RFtrain$ro
  ggplot2::annotate("text", size = 3, x = .75, y = .40,
                    color = 'darkblue',
                    label = paste("RF model 32 AUC = ", round(max(model32RFtrain$ro
  ggplot2::annotate("text", size = 3, x = .75, y = .35,
                    color = 'orange',
                    label = paste("RF model 43 AUC = ", round(max(model43RFtrain$ro
  geom_roc(model28GLMtrain,
           mapping = aes(m = dam, d = factor(obs, levels = c('no_dam', 'dam')),
                         color = 'GLM model 28'), n.cuts=0,
           increasing =        )+
  geom_roc(model43GLMtrain,
           mapping = aes(m = dam, d = factor(obs, levels = c('no_dam', 'dam')),
                         color = 'GLM model 43'), n.cuts=0,
```

mortgage loan applications, or residents of Poland's Podhale region facing discrimination when applying for US visas.

The idea of teaching an automatic system to recognize patterns emerged early on, in connection with astronomical observations. A "learning system" was defined in 1997 by Tom Mitchell in the book *Machine Learning* as one that improves its performance with experience. In biology, the concept of a "perceptron," mimicking the functioning of neurons in the brain, was proposed in 1958. In 2001, Leo Breiman's classic publication *Random Forests* demonstrated how to isolate a classification model using "decision trees." Decision trees have become a popular method for various machine learning tasks; the "random forests" method is a way of averaging multiple deep decision trees.

Since most natural phenomena can be seen as bundles of interlinked factors, it is assumed that any particular phenomenon or environmental property (Y) can be explained by a set of predictors (x_1, ..., x_n) acting as independent variables. Not all predictors explain the given phenomenon or feature in the same way, and their strength is analyzed during the initial modeling stage. Historical data is a crucial component of this entire process, enabling the learning system to acquire knowledge about patterns based on such data. This knowledge is then applied in the form of a model to unsampled space (e.g., geographical space), a portion of the population or of dynamic phenomena, to predict their future states.

Multidimensional datasets consisting of a dependent variable plus a large number of predictors with often non-uniform formats (continuous or categorical data) and different probability distributions require a special approach. The task of the machine learning algorithm is to create a model that allows for prediction. In their book *Applied Predictive Modeling*, Max Kuhn and Kjell Johnson define modeling as follows: "the process of developing a mathematical tool or model that generates an accurate prediction." In *Hands-On Machine Learning with R*, Brad Boehmke and Brandon Greenwell emphasize that an essential feature of machine learning is that it is an iterative process, based on a heuristic approach.

Sometimes we know little about the phenomenon we are analyzing, and we tend to base our actions on incomplete data. We do not know which machine learning method will best reflect the actual pattern of the phenomenon under study or the state of the environment. Hence the need to apply, evaluate, modify the method or data, and regenerate (train) the model on the same data set. In many cases, this approach yields the best desired effect – a model with a certain degree of generalization (not overfit), applicable to many different test datasets.

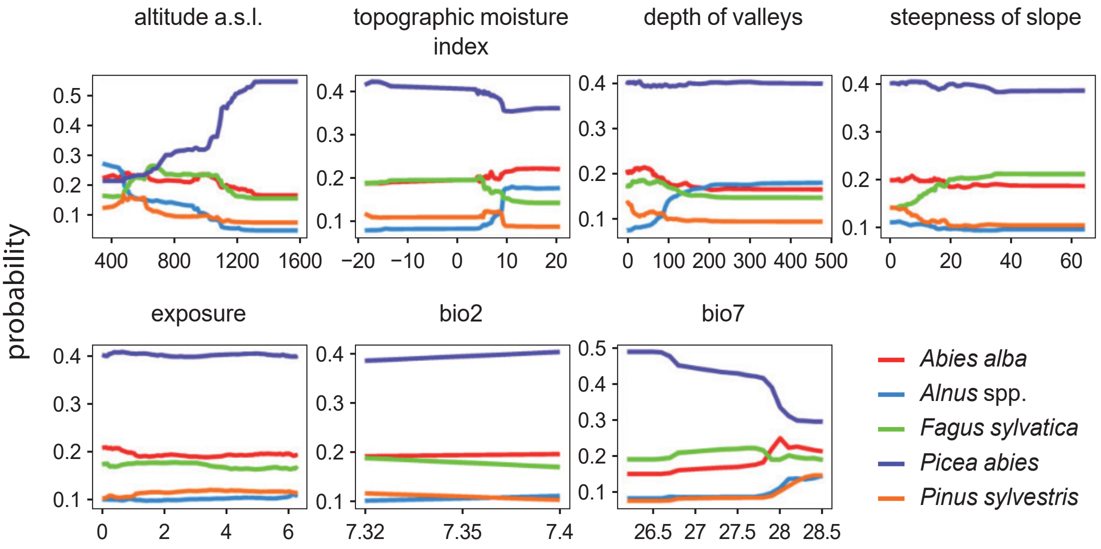
Data

The great abundance of different types of data makes initial data assessment and preparation crucial stages in developing any model. In fact, it is estimated that about 80% of the time spent on data analysis is actually dedicated to data preparation. The well-known principle of GIGO (garbage in, garbage out) always needs to be kept firmly in mind at this stage, as erroneous, incomplete, or insufficiently large datasets can lead to erroneous conclusions. Observational data collected by conventional methods – using measurement instruments – often yielded small-volume datasets plagued by errors and inaccuracies due to the measurement methods used (their spatial and temporal resolution), insufficient precision, or instrument malfunctions. Since not everything can be measured, any analysis always involves a certain level of simplification, which ultimately takes the form of a model. Perhaps the simplest analogy is the cartographic definition of a map – which states that a map is simply a model of reality.

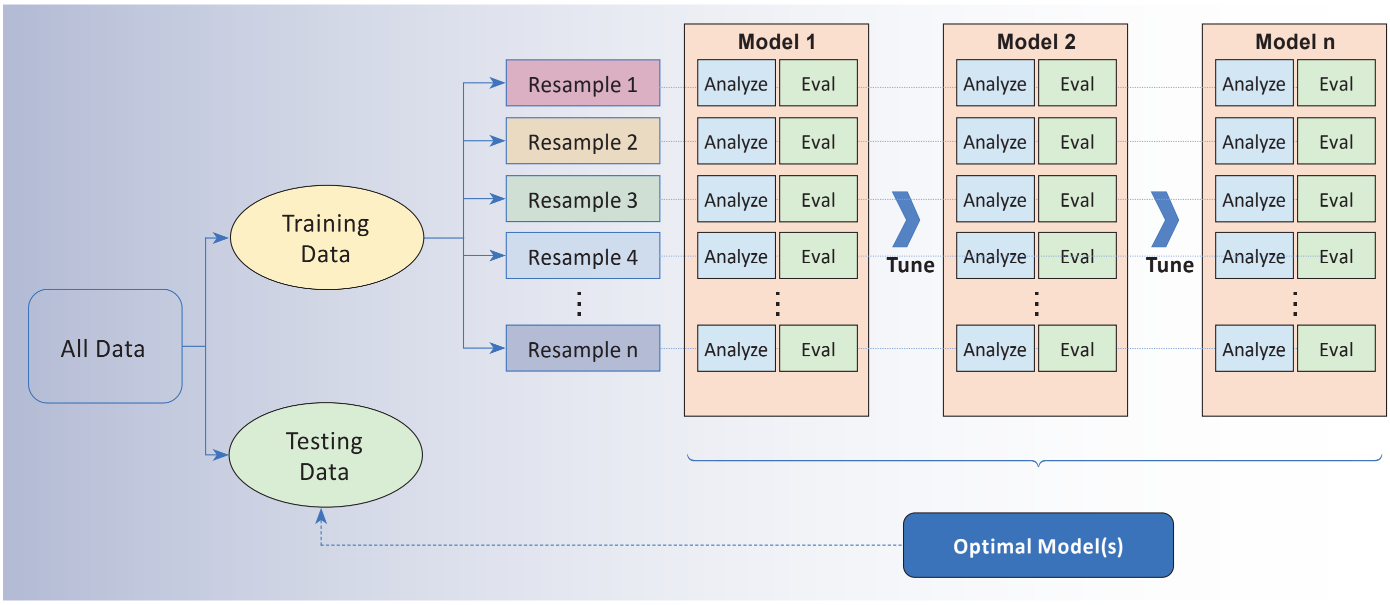
With the advent of systematic laboratory measurements, satellite imaging, and automated meteorological measurements, the resulting large volumes of data presented an opportunity to gain a better understanding of the complexity of the physical and biological world. However, extracting information from such thick jumbles of data required new computational techniques, larger databases, and faster processors. Nowadays, Internet users themselves generate data (leaving digital footprints), which is then brought to bear in a model determining, for example, which ads will be displayed on their computer screens. Email systems are trained to automatically recognize spam messages, and highway toll booths recognize our car's license plates, deciding whether to allow us to drive through – or not.

Supervised learning

One of the essential features of a dataset that determines the applicability of certain machine learning methods is whether the observations provide information about a dependent variable – in other words, about the model's response (target variable) for a specific set of predictors. If the dataset contains labels (in a classification model, where the dependent variable can be binary, this may be yes or no, 0 or 1, destruction or no destruction) or specific numerical values (in a regression model, e.g. estimating forest biomass), supervised learning can be applied, where we provide information about the variable Y as input. This way, we can "supervise" an expected outcome of the model when a specific algorithm is applied – for instance, the *random forests* technique. We should note a certain limitation of this method, which may arise from the nature of the phenomenon itself. Let's assume we are modeling the growth of real trees in a real forest. If we analyze how the rate of tree growth is a dependent on temperature and rainfall based only on data from temperate climate zones, that model cannot be applied to analyze forests in a different climatic zone. It has been trained based on observations within a certain value range and for certain tree categories that do not occur in another zone where the user plans to apply the model. This points out an important feature of regional models. For example, in their 2021



Graphs showing the averaged probability for analyzed tree species: bio2 stands for the daily temperature range averaged over the same year, bio7 for the difference between the maximum temperature of the warmest month and the minimum temperature of the coolest month of the year



Using training data in an iterative modeling and evaluation process, until an optimal model is developed (based on Boehmke and Greenwell, 2020)

article "Predicting into unknown space? Estimating the area of applicability of spatial prediction models," Meyer and Pebesma proposed that models can be validated using the area of applicability (AOA) method. In a nutshell, this is the area in which the model can learn the relationship between variables based on training data and the model quality estimated during cross-validation is maintained at a certain acceptable level.

An essential aspect of supervised model validation is that a portion of the data from the main dataset, e.g. 25% of observations, is set aside to be used as a test set, not used for developing the model itself (training). This way, we maintain control over the model's quality. Note that a model can be overfitted, meaning it demonstrates great predictive power for a data subset of similar characteristics but becomes useless when applied to entirely new data.

Applying algorithms

The random forest algorithm is one of the most popular machine learning methods based on multiple decision or classification trees. When building a tree, a random sample of *m* predictors (independent variables) is chosen. If a classification model is being developed, this value equals the square root of the number of predictors. Using a limited number of predictors helps bypass the problem of collinearity (correlation) because they are randomly selected when building, for example, 1000 classification trees. Thanks to this procedure, the random forest algorithm handles strongly correlated data well because the result is averaged for many trees.

The random forests algorithm has found a wide variety of applications – including, interestingly enough, to modeling real forests, which develop under a complex mixture of biotic, abiotic, and anthropogenic factors, not easily explained by simpler models. For instance, the method has been used to model the spatial distribution of major tree species and their biomass in national parks in southern Poland, as well as to model the damage caused by Cyclone Klaus in 2009 in forests in southwestern France.

To determine the occurrence patterns of individual tree species in alpine forests, we used data from forest stand descriptions from five national parks and maps illustrating the climate and geomorphometric characteristics of these areas. To identify the dominant tree species, a classification model was applied, assigning each observation (stand) to one of five species. In this model, "decision trees" were used to determine, based on the data, the probability that a specific tree species would dominate in a given location. We demonstrated that different factors were decisive for the occurrence of different species. For example, the probability of spruce occurrence was mainly related to elevation above sea level, while pine occurrence was found to be linked to exposure and slope. Applying this model allowed us to explain which factors are most important for the occurrence of particular species. Additionally, visualizing the model's response when assuming a constant level for all variables except one enabled us to simulate changes in environmental conditions, showing how modifying specific factors will affect the studied species. Using a regression model, in turn, allowed us to infer how a unit change in a specific predictor (e.g., exposure) will affect stand biomass. Thus, our models can be used to predict how changes in climate or topography could influence trees' ability to accumulate carbon and mitigate climate change. ■

Further reading:

Dyderski M.K., Pawlik Ł., Spatial distribution of tree species in mountain national parks depends on geomorphology and climate, *Forest Ecology and Management* 2020, doi.org/10.1016/j.foreco.2020.118366

Pawlik Ł., Godziek J., Zawolik Ł., Forest damage by extra-tropical cyclone Klaus – modelling and prediction, *Forests* 13 (12)/2022, doi: 10.3390/f13121991

Pawlik Ł., Harrison S.P., Modelling and prediction of wind damage in forest ecosystems of the Sudety Mountains, *Science of the Total Environment* 2022, doi.org/10.1016/j.scitotenv.2021.151972