Land suitability analysis of Alvar grassland vegetation in Estonia using Random Forest

Irada Ismayilova, Evelyn Uuemaa, Aveliina Helm, Christian Röger, Sabine Timpf

Angaben zur Veröffentlichung / Publication details:

Ismayilova, Irada, Evelyn Uuemaa, Aveliina Helm, Christian Röger, and Sabine Timpf. 2020. "Land suitability analysis of Alvar grassland vegetation in Estonia using Random Forest." *GI_Forum: Journal for Geographic Information Science* 8 (1): 63–72. https://doi.org/10.1553/giscience2020_01_s63.

GI_Forum 2020, Issue 1
Page: 63 - 72
Full Paper
Corresponding Author:
irada.ismayilova@geo.uni-augsburg.de
DOI: 10.1553/giscience2020_01_s63

Land Suitability Analysis of Alvar Grassland Vegetation in Estonia Using Random Forest

Irada Ismayilova[1], Evelyn Uuemaa[2], Aveliina Helm[2], Christian Röger[1] and Sabine Timpf[1]

[1]University of Augsburg, Germany
[2]University of Tartu, Estonia

Abstract

Calcareous alvar grasslands are one of the most species-rich habitats in Estonia. Land-use change and cessation of traditional agricultural practices have led to a decrease of the area of these valuable grasslands during the past century. Therefore, their conservation and restoration are becoming increasingly important. Efforts to restore these habitats have already been made in recent years. Land suitability analysis for potential restoration sites, using the machine learning technique Random Forest (RF), was performed for the first time in this study, which aimed to assess the use of RF for a suitability analysis of alvar grassland. RF predicted 610.91 km^2 of areas suitable for restoring alvar grasslands or for creating alvar-like habitats in Estonia. These areas include all existing alvar areas as well an additional 140.91 km^2 suitable for establishing new habitat similar to calcareous alvar grasslands. We discuss suitability analysis to help with restoration planning and find it to be a reasonable and efficient tool that has potential to provide relevant information. The quality of the prediction could be improved by including additional data relevant for alvar grasslands, such as soil depth, but such data was unfortunately unavailable.

Keywords:

alvar grasslands, restoration, land suitability, machine learning, Random Forest

1 Introduction

Alvars are calcareous grassland habitats with a limited distribution on Earth. They are found mostly in Estonia, Sweden and a few other smaller areas in the Northern hemisphere (Pärtel et al. 1999). These grasslands have high conservation value both in Estonia and in Europe as a whole due to their species richness, the variety of their important ecosystem services, and their high relevance in supporting natural and cultural heritage in European landscapes. Alvar grasslands are among Annex I of priority habitat types in the EU Habitats Directive (*6280 Nordic alvar and Precambrian calcareous flatrocks) (Rosén 1982). Over the millennia, Estonian alvar grasslands have been supported by moderate human influence, especially grazing. Land-use change, leading to the cessation of grazing, afforestation or direct destruction, have resulted in substantial decreases in area of alvars and the subsequent loss of

their species richness (Helm, Hanski & Pärtel 2006). The current distribution of alvar grasslands is shown in Figure 1.

Considering their high conservation value and position among priority habitat types in the Natura 2000 network, alvar grasslands as well as other types of dry calcareous grassland are in urgent need of restoration and more effective conservation (Helm, Urbas & Pärtel 2007). The most recent restoration work has focused on shrub clearance and the removal of other unwanted vegetation in long-abandoned or afforested alvar grasslands, with the aim of restoring original alvars (Holm 2019, Helm 2019). One managerial measure would be to look for suitable areas where new alvar-like habitats could be created.

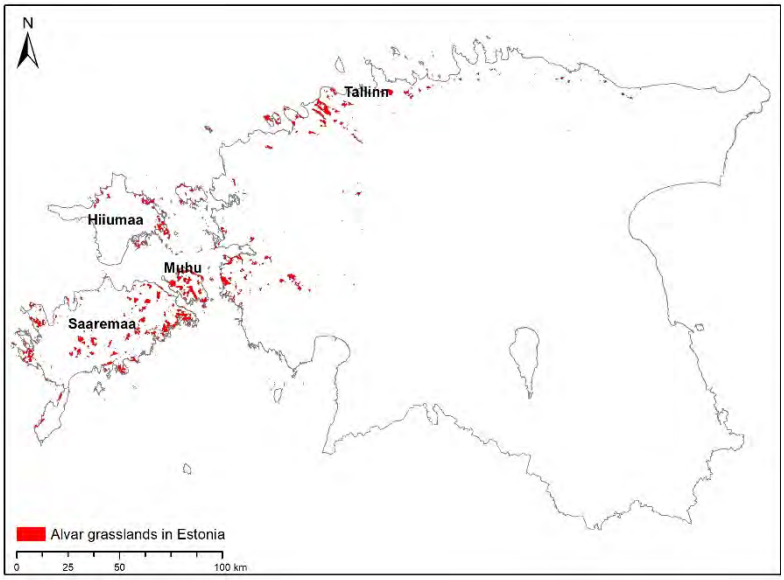

Figure 1: Distribution of alvar grasslands in Estonia

From 2014 to 2019, ca 2,500 ha of overgrown alvar grasslands were restored in western Estonia as part of the LIFE to Alvars project (Holm 2019). This as well as other grassland restoration projects have focused solely on historical grassland areas that have been degraded by becoming overgrown with trees and shrubs. However, for future restoration planning it would be helpful to identify all potential regions where grassland restoration or the creation of new grassland areas would be environmentally feasible.

Land suitability analysis is one of the most frequently used techniques in environmental management and the planning of habitat restoration. For example, Novak and Short (2000) performed a suitability analysis for eelgrass meadows on Plum Island. Hunter et al. (2016) carried out a restoration suitability assessment for swamps in order to safeguard and improve the provision of important ecosystem services. However, land suitability analysis for alvar grasslands in Estonia has not been performed so far. Therefore, this study aims to identify potentially suitable areas of alvar grassland for restoration or for the creation of alvar-like

habitats. We use a method from machine learning, called Random Forest (RF), because of the limitations of available datasets relating specifically to alvar grasslands and yet a large amount of data to process. The literature on RF confirms that it is timesaving for handling large amounts of data and capable of highly accurate predictions. Our analysis covers the whole of Estonia.

2 Random Forest method

Machine learning methods are becoming increasingly popular in land suitability analysis thanks to their ability to deal with complex relationships between predictor variables, robustness in managing big and noisy data, and being economical in terms of time required (Lahssini et al. 2015). RF, as proposed by Breiman (2001, p. 6), is "a classifier consisting of a collection of tree-structured classifiers {h(x,Θk), k=1, ...} where the {Θk} are independent identically distributed random vectors and each tree casts a unit vote for the most popular class at input x". This method is an extension of classification and regression trees and uses the Classification and Regression tree algorithms (CART). The classification algorithm is used when aiming to predict categorical values or labels; the regression algorithm is used when aiming for numerical predictions (Strecht et al. 2015). RF can be described as trees where branches are formed by answers to yes/no questions and are not pruned (but can be). Each tree in the forest is constructed using bootstrap samples from the original dataset. It uses a random selection of explanatory variables or factors to split the tree at nodes.

The goal of RF is to identify the best model to analyse the relationship between dependent and independent variables (Friedman and Meulmann, 2003). In order to evaluate how good the RF model is, the data needs to be split into two parts: training and testing data. This helps to evaluate the performance of the algorithm for the chosen problem by training one sample of data and validating it using the test sample. RF, in both classification and regression models, also provides a measure of the importance of a variable based on the contribution of the variable to the model at each node and each tree where it appears. Another estimated value that can be obtained from the model is the Out Of Bag (OOB) score, an average error of prediction of out of bag samples (samples that do not appear in bootstrap samples (Breiman 2001). RF has also proved to be a suitable method when there is a correlation between the variables involved in the analysis (Georgian et al. 2019).

Several machine learning techniques have already been incorporated into land suitability analysis. For instance, Wen et. al. (2009) used classification and regression trees to investigate hydrological requirements of the river Red, while Park et al. (2003) applied an artificial neural network to predict aquatic insect species. Landscape configuration and habitat suitability were analysed by Holzkaemper et al. (2006) using genetic and simulated annealing algorithms. However, many studies have shown that RF most often attains the best predictive performance (Garzon et al. 2006). Lahssini et al. (2015) and Vincenzi et al. (2011) used RF to detect cork oak suitability and *Ruditapes philippinarum*'s potential spatial distribution assessment respectively. The probability of correct predictions in both studies was more than 90%.

3 Using Random Forest to Predict Suitability for Alvar grasslands

We predicted the probability of the suitability of areas throughout Estonia for restoration of alvar grassland or for the creation of alvar-like habitats. In stage one, we focused on defining environmental variables and extracting predictor variables. In the second stage, we used the Scikit-Learn library in Python for the prediction of suitability and defined the most suitable probability threshold for the given question. Results were visualized, and were examined for reliability by experts in the department botany at the University of Tartu (UT).

3.1 Choosing environmental predictors

The main data sources used in this work were the Estonian Soil Database (ESD), a LiDAR-based Digital Elevation Model (DEM), and the Estonian Geological Database. From these datasets, we chose eight predictor variables for further use.

The occurrence of alvar grasslands is limited to three main bedrocks: Silurian, Ordovician, and to a lesser extent Cambrian (Pärtel et al. 1999). Bedrock is considered the most important predictor variable in identifying suitable areas for alvar grassland. Alvar-type vegetation occurs only on thin and calcareous soils. Usually the soil depth over the bedrock is less than 20 cm, and in some alvars it is even less than 5cm (Pärtel et al. 1999). Therefore, we extracted soil-type and soil-texture information from the ESD. Since soil silt, soil sand and soil rock content describe soil state and condition, they were also used as predictor variables. No concrete example of a correlation between the slope of terrain, Topographic Wetness Index (TWI) and alvar-like vegetation in Estonia has been established. However, prior statistical analysis of available datasets showed that slope and TWI values under alvar grasslands are always within a certain range. Therefore, we considered using slope and TWI as further predictor variables to ensure higher information gain for the RF models. A DEM with 1 m resolution was used to calculate slope and TWI. The datasets, their sources, and the predictor variables involved in the suitability analysis are shown in Table 1.

Table 1: Environmental datasets and variables used in land-use suitability predictions

Datasets	Source	Predictor variables (data type)
Soil Database	Estonian Land Board & Kmoch et. al., (2019)	Soil type (categorical)
		Soil texture (categorical)
		Soil silt content (numerical)
		Soil sand content (numerical)
		Soil rock content (numerical)
DEM	Estonian Land Board	Slope (numerical)
		TWI (numerical)
Geological Database	Estonian Land Board	Bedrocks (categorical)

Data for the location of Alvar grasslands in Estonia was provided by the Botany Department of UT, in the form of two datasets. One of the two contained the most recent alvar grassland distribution information available for Estonia. This dataset is a product of the survey of the Estonian Semi-Natural Community Conservation (2000–2010) and alvar distribution mapping based on the Estonian state-run database EELIS. The second dataset is a result of the Estonian vegetation mapping from 1930 to 1950 and was helpful to understand the historical distribution of alvars. We merged these datasets to create a single one in which we assigned "1" to all areas indicating presence of alvar grasslands. For absence data, we generated random points in the areas outside of alvar grasslands with suitable bedrocks. The absence of alvar grasslands was indicated by "0".

In order to assure identical extent, cell size and coordinate systems for the suitability analysis, pre-processing of layers was carried out in ArcGIS. This step resulted in one big joint database, with 41,657 objects indicating presence ("1") and absence ("0") of alvar grasslands and containing predictor variables.

3.2 Suitability modelling using Random Forest

In order to identify suitable areas using RF, a list of variables (predictors), summarized in Table 1, was used. As a first step, we did one-hot encoding for categorical variables. We then created five models with different combinations of predictor variables in order to find the most suitable combination. As part of the general procedure, we split the data into training and test sets. There is no information available on which proportions for splitting the data work best, but in many similar studies, datasets have been divided into 60/40 or 70/30 ratios. We therefore used both these ratios and chose the better option (Table 2). The target variable in the training phase was the alvar grassland presence or absence data. In order to assess the performance of the trained model (how well it can recognize alvars), test sets were used. Using the "RandomizedSearcCV" function from the scikit-learn library in Python, we aimed to define the best set of parameters (e.g. n_estimators, max_depth).

We estimated the accuracy of the models using the R-squared value and OOB Error estimate produced by k-fold cross-validation with 3-fold. Using the best model parameters from the hyperparameter tuning process and the dataset covering the whole of Estonia (alvar grassland vegetation presence or absence was excluded), we fitted the RF model and obtained continuous values between 0 and 1 representing the probability of suitability. In order to convert the values into binary maps, we had to identify the threshold above which the areas were most suitable, and below which areas had low suitability or were unsuitable. Because of the nature of the analysis, we set the suitability threshold after examining the results.

Table 2: RF Models with different combinations of predictor variables and split options

Base set of predictor variables used in training/test models	Split options
[Model 1] Soil type, Soil texture, Soil clay, Soil silt, Soil sand, Soil rock, Slope, TWI, Bedrock [Model 2] Soil type, Soil clay, Soil silt, Soil sand, Soil rock, Slope, TWI, Bedrock [Model 3] Soil type, Soil texture, Slope, TWI, Bedrock [Model 4] Soil texture, Soil clay, Soil silt, Soil sand, Soil rock, Slope, TWI, Bedrock [Model 5] Soil type, Soil texture, Soil clay, Soil silt, Soil sand, Soil rock, Slope, TWI	[Split option **1**] 60/40 & [Split option **2**] 70/30

4 Results of the suitability analysis

Once all the models had been run, their prediction statistics were checked and R^2 values and OOB scores were examined. The results are shown in Table 3.

Table 3: R-squared and OOB scores for individual RF models

Variant	Split option	R^2 score	OOB score
Model 1	1	0.7617	0.75438
Model 2	1	0.7457	0.74309
Model 3	1	0.7888	0.75424
Model 4	1	0.7783	0.77606
Model 5	1	0.7631	0.76548
Model 1	2	0.7756	0.77503
Model 2	2	0.7884	0.79678
Model 3	2	0.7919	0.79670
Model 4	2	0.7861	0.79668
Model 5	2	0.6692	0.68487

Table 3 shows R^2 and OOB scores for different model variations. There were almost no differences between these figures across all models. Only one model (Model 5) stood out slightly from the rest. It contained soil type, soil texture, slope, TWI and bedrock as predictor variables. Further, split option 2 (70/30) was chosen as suitable for alvar grassland

identification. The accuracy of the selected model was 0.79 and 0.8 for the R^2 and OOB scores respectively. This means that RF has a useful prediction capability. We also checked which predictor variables contribute the most to the information gain of the RF model, and to what extent the accuracy will decrease if a certain variable is removed from the model. We used a Mean Decrease in Accuracy metric which was calculated via the permutation feature importance algorithm (rfpimp; Breiman 2001). The results are shown in Figure 2. In the chosen model, bedrock was the most important predictor while TWI was the least important.

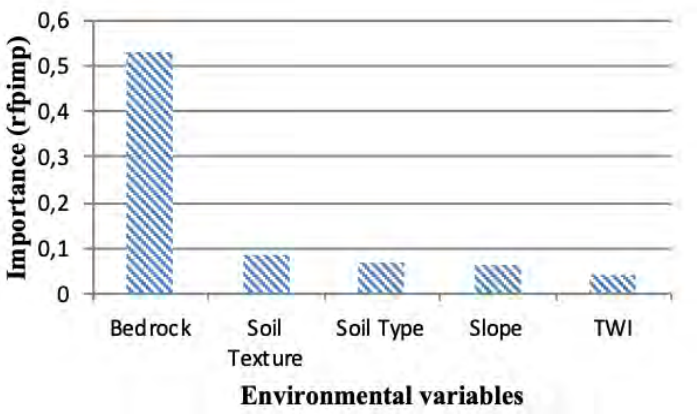

Figure 2: Importance of each variable in the final RF model

The predicted occurrence ranges from 0.04 to 0.884%. Therefore, probability of more than 80% was considered highly suitable; probability of less than 80% was considered unsuitable or of low suitability for alvar grasslands. The aim was to find suitable areas with very high probability, and therefore a high threshold was selected.

Of all the suitable areas for alvar grasslands, 45% are currently forests, 34% are croplands, and 11% are grasslands, as shown in Table 4.

Table 4: Actual land use of the areas predicted to be suitable for alvar grassland restoration using RF

Land use	Areal percentage (%)
Forest	44.94
Cropland	33.49
Grassland	11.02
Other	6.66
Shrubland	2.32
Urban	1.14
Wetland	0.34
Water	0.09

The final result of our analysis is illustrated in Figure 3. RF predicts a total of 610.91 km² where currently no alvar grasslands exist. Out of those, 470 km² were once alvar grasslands. The most suitable areas for alvar grassland restoration in Estonia are in the western islands (Saaremaa, Muhu, Hiiumaa), and north-western and northern inland areas. Low suitability or unsuitable areas fall in southern Estonia.

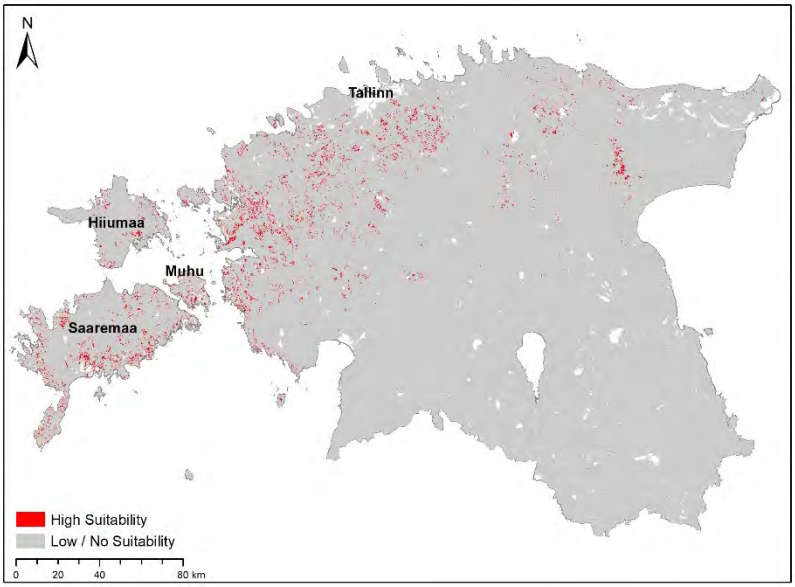

Figure 3: Results of the RF model for highly suitable and less suitable areas for alvar grassland restoration or creation of alvar-like habitats

5 Discussion and Conclusions

In this study, we predicted the potential occurrence of alvar grassland vegetation in Estonia. Limited distribution of these grasslands in Estonia and the small number of environmental variables that characterize alvar grasslands in Estonia makes the suitability analysis very challenging. Our case study shows that RF is a suitable method for finding these areas. It has the ability to learn by itself from the available data (in this case, current locations of alvar grasslands and their properties) and make predictions based solely on data – i.e. without human constraints such as the need for expert knowledge. Compared to other well-known methods such as Multi Criteria Decision Making (MCDM), this reduces massively the time spent preparing the data prior to performing land suitability analysis. Thus, RF is used in many studies dealing with land suitability analysis. Our model performs well, with an accuracy of 80%. Experts in the Botany Department at the University of Tartu, Estonia, confirmed our results. Similar studies applied to other areas and land suitability types such as Garzon et al. (2006) reach an accuracy of 90% or even higher. We believe that the difference between our accuracy and that reached by similar studies is due to limitations caused mainly by our choice of datasets for the analysis.

We expected bedrock, soil texture and soil type to provide the highest contributions to the final model. The data we use does not contain soil depth information, as this is not (yet) available for the whole of Estonia. Albert (1998) states that alvar grasslands occur on thin soils of no more than 20 centimetres. In our dataset, some areas have an actual soil depth record while others show the depth of a soil profile up to one meter. Additionally, soil pH could not be included in our suitability analysis. Including this information should increase the prediction accuracy.

We conclude that RF is a reliable method for performing land suitability analysis for alvar grasslands. An accuracy of 80% is acceptable considering the limitations of our data. Although other information such as soil depth and soil pH is missing, the datasets used perform well with RF.

Future research will focus on applying other land suitability analysis methods such as MCDM in order to allow a comparison of the results, and on optimizing the choice of datasets. Taking into account factors like soil depth and soil pH will most probably enhance our results. Our goal is to reach an accuracy of 90% or higher.

References

Albert, D. A., & Kost, M. A. (1998). Natural community abstract for lakeplain wet prairie. *Michigan Natural Features Inventory, Lansing, MI.*

Breiman, L. (2001). Random forests. *Machine Learning* 45: 5–32

Friedman, J.H., Meulman, J.J. (2003). Multiple additive regression trees with application in epidemiology. *Statistics in medicine*. 22 (9): 1365–1381

Garzon, M. B., Blazek, R., Neteler, M., De Dios, R. S., Ollero, H. S., & Furlanello, C. (2006). Predicting habitat suitability with machine learning models: the potential area of Pinus sylvestris L. in the Iberian Peninsula. *Ecological modelling*, 197(3-4), 383-393.

Georgian, S., Anderson, O., Rowden, A. (2019). Ensemble habitat suitability modelling of vulnerable marine ecosystem indicator taxa to inform deep-sea fisheries management in the South Pacific Ocean. *Fisheries Research* 211:256–274

Helm, A. (2019). Large-scale restoration of Estonian alvar grasslands: im-pact on biodiversity and ecosystem services: Final Report of Action D.1

Helm, A., Hanski, I., & Pärtel, M. (2006). Slow response of plant species richness to habitat loss and fragmentation. *Ecology Letters* 9: 72–77

Helm, A., Urbas, P., & Pärtel, M. (2007). Plant diversity and species characteristics of alvar grasslands in Estonia and Sweden. *Acta Pythogeographica Suecica*. 88: 33-42

Holzkaemper, A., Lausch, A. and Seppelt, R. (2006). Optimizing Landscape Configuration to Enhance Habitat Suitability for Species with Contrasting Habitat Requirements. *Ecological Modelling* 198: 277-292

Holm, A. (2019). Life to alvars project: Report of Action C.1

Hunter, R., Day, J., Shaffer, G., Lane, R.,Englande, A., Reimers, R., Kandalepas, D., Wood, William B., Day, J., Hillmann, E., Bank, E. (2016). Restoration and Management of a Degraded Baldcypress Swamp and Freshwater Marsh in Coastal Louisiana. *Water* 8: 79-101

Lahssini, S., Lahlaoi, H., Mharzi, H., Bagaram, M., Ponette, Q. (2015). Predicting Cork Oak Suitability in Maˆamora Forest Using Random Forest Algorithm. *Journal of Geographic Information Systems* 7: 202-210

Novak, B., Short, T. (2000). Creating the Basis for Successful Restoration: An Eelgrass Habitat. *Ecological Engineering* 15: 239-252

Park, S., Céréghino, R., Compin, A. and Lek, S. (2003). Applications of Artificial Neural Networks for Patterning and Predicting Aquatic Insect Species Richness in Running Waters. *Ecological Modelling* 160: 265-280

Pärtel M.,Mändla R. & Zobel M. (1999). Landscape history of a calcareous (alvar) grassland in Hanila, western Estonia, during the last three hundred years. *Landscape Ecology* 14: 187-196

Rosén, E. (1982). Vegetation development and sheep grazing in limestone grasslands of south Öland, Sweden. *Acta Phytogeographica Suecica*. 72: 1-104

Strecht, P., Cruz, L., Soares, C., Moreira, J., Abreu, R. (2015). A Comparative Study of Classification and Regression Algorithms for Modelling Students' Academic Performance. *International Educational Data Mining Society*.

Vincenzi, S., Zucchetta, M., Franzoi, P., Pellizzato, M., Pranovi, F., De Leo, G.A. and Torricelli, P. (2011). Application of a Random Forest algorithm to Predict Spatial Distribution of the Potential Yield of Ruditapes philippinarum in the Venice Lagoon, Italy. *Ecological Modelling* 222: 1471-1478

Wen, L., Ling, J., Saintilan, N. and Rogers, K. (2009). An Investigation of the Hydrological Requirements of River Red Gum (Eucalyptus camaldulensis) Forest, Using Classification and Regression Tree Modelling. *Ecohydrology* 2: 143-155