MODELING PRESENT AND PROSPECTIVE DISTRIBUTION OF *PHYTEUMA* GENUS IN CARPATHIAN REGION WITH MACHINE LEARNING TECHNIQUES USING OPEN CLIMATIC AND SOIL DATA

Assoc. Prof. Dr. Alexander Mkrtchian
Ivan Franko National University of Lviv, Ukraine

ABSTRACT

Species distribution modeling can be effectively carried out using open data and data analysis tools with machine learning techniques. Modeling of the distribution of *Phyteuma* genus in the Carpathian region has been carried out with data from the GBIF database, climatic data from the Worldclim database, and soil properties data from Soilgrids soil information system. Spatial distribution modeling was accomplished with machine learning techniques that have marked advantages over more traditional statistical methods, like the ability to fit complex nonlinear relationships common in ecology.

Four methods have been examined: Maxent, Random Forest, Artificial Neural Networks (ANN), and Boosted Regression Trees. AUC and TSS criteria calculated for testing data with cross-validation have been applied for assessing the performance of the models and to tune their parameters. ANN with a reduced set of predictor variables (6 from initial 21) appeared to fare the best and was applied for predictive modeling. Prospective data based on future climate projections from Worldclim were input to the model to get the prospective distribution of the plant taxon considering expected climate changes under different RCPs.

Keywords: *species distribution modeling, machine learning, Carpathians, open data*

INTRODUCTION

Accurate knowledge of species distribution is an important prerequisite for effective conservation practices, e.g. regarding the designation of protected areas. It concerns endangered species, as well as keystone species playing a critical role in maintaining the ecosystem integrity and umbrella species which protection indirectly protects many other species and the ecological community in general.

While counting and mapping species distribution in field is very laborious and cumbersome, species distribution modeling (SDM) becomes an indispensable tool, which application is facilitated nowadays by the availability of spatial data on factors determining species distribution, modern methods and techniques for data analysis, and processing capabilities of modern computers. Species distribution models estimate the relationship between species records at sites and the environmental and/or spatial characteristics of those sites [3]. There is a considerable amount of publications coming out recently devoted to the topic in general or some specific issues related to it (e.g. [3]).

Species distribution modeling can be effectively carried out using open data and open data analysis tools. It is especially of value for countries and projects with limiting research funding. Modern machine learning techniques are more suitable for the purpose comparing with more traditional statistical approaches due to the very nature of the problem: effects of predictive variables on target species distribution are usually non-linear, these variables are often highly interdependent and spatially autocorrelated, voids and errors in data are common, etc. Only quite recently did these techniques enter the mainstream of ecological modeling, mainly due to relatively high computational demands met only by relatively modern computers.

SDM results are not only helpful in delineating the presumable actual locations of target species, but can also be used as predictions of future distributions of species habitats, when data on prospective distributions of predictive variables are available. As climatic conditions are expected to change significantly in the course of the present century due to human-induced emissions of greenhouse gases, habitats of most species are expected to shift accordingly, as many climatic characteristics have direct physiological impact on plants and animals.

An objective of this study is to model the present and prospective distribution of *Phyteuma* (rampion) genus in the Carpathian region with open climatic and soil data, using a bunch of machine learning techniques. This genus, common for forested low- to middle altitude habitats in different parts of Carpathian region, could be regarded as umbrella taxon for the protection of most valuable Carpathian biological communities and ecosystems. The genus contains several species which were considered in aggregate due to insufficient number of records for single species and their similar ecological characteristics.

MATERIAL AND METHODS

Database maintained by the Global Biodiversity Information Facility (GBIF) was used as a data source for species observations [4]. Records were selected falling inside an arbitrary defined 600*800 km rectangle encompassing Carpathian mountain range as well as foothills and parts of neighbouring planes and hills (Fig. 1).

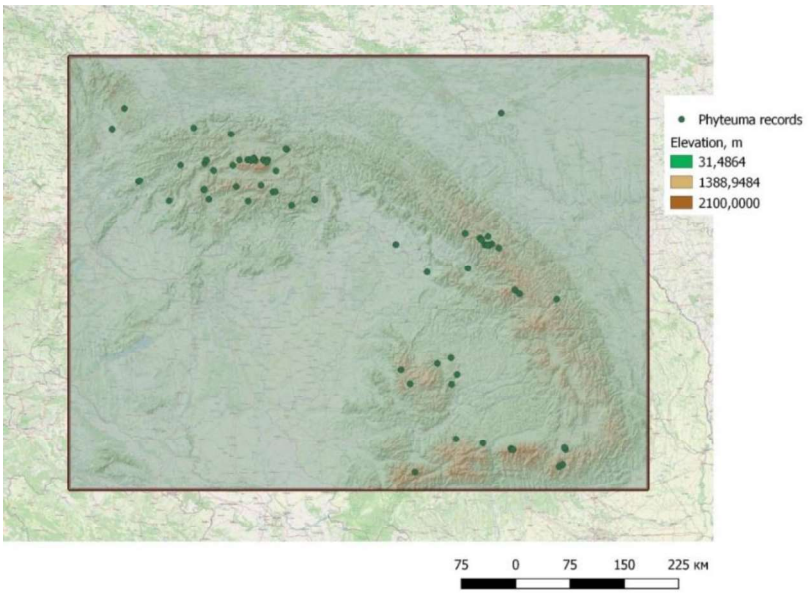

Fig. 1. *Study area and locations of Phyteuma records in GBIF database*

Data obtained from GBIF for the Carpathian region contained 148 records of *Phyteuma* genus in total. While the GBIF Secretariat claims to apply a set of semi-automatic steps to remove duplicates and false positives, this is still an issue, as became obvious after inspecting the records and founding many ones sharing the same species name and location. After removing duplicates, 80 records have been kept. Most of SDM algorithms require some kind of absence or background data to contrast the presence data to. A double of observation records (160 points) were thus designated as background, with random coverage of geographic space inside study area. Six data points (3 from observed data and another 3 from simulated background data) have subsequently been removed due to omissions in predictors data.

The choice of predictive variables for SDM was based on two considerations: 1) their relevance as ecologically meaningful characteristics related to ecological factors driving the distribution of species, and 2) the availability of respective open data in the form of global spatial layers.

The distribution of plant species is influenced by two types of ecological gradients: those related to climatic conditions (mainly to thermal and precipitation characteristics) and those related to properties of soils (nutrients availability, acidity, water retention capacity, aeration, etc.) Data on climatic conditions were derived from WorldClim database. It contains a set of global climate layers (grids) with a spatial resolution of about $1km^2$ [2]. Among others, there are layers of 19 bioclimatic variables (coded as BIO1 to BIO19), which are derived from monthly temperature and precipitation with a consideration to have biological significance. While two of them (BIO3 and BIO7) are totally excessive being functions of some

other ones, 17 out of 19 bioclimatic variables were taken out as predictive variables for SDM.

For data on soil conditions, SoilGrids digital maps of soil properties were used. These were produced for the entire globe at 250 m spatial resolution with state-of-the-art machine learning methods, taking as inputs soil observations data from about 240 000 locations worldwide and over 400 global environmental covariates [5]. From 11 available physical and chemical soil properties, 4 were chosen as the most suitable predictors: soil acidity, organic carbon stock, cation exchange capacity, and total nitrogen. Among six standard depth intervals available, 15–30 cm depth interval was chosen as the most appropriate for the purpose.

R free programming language and software environment for statistical computing and graphics provides for a number of packages with functions for spatial analysis and modeling, machine learning techniques, and specifically for SDM (e.g. sdm, dismo, etc.) A number of such functions were applied at different stages of data processing and analysis. The main analysis was carried out with SDMtune – a rather new R package that aims to facilitate training, tuning, and evaluation of species distribution models in a unified framework [6]. SDMtune package provides tools for tuning model hyperparameters with a novel genetic algorithm, and for data-driven variable selection to avoid model overfitting.

THEORY AND CALCULATION

Initially, *prepare SWD* function creates an SWD object, given the coordinates, the species' name and the environmental variables. *Train* function then applies to the SWD object one of a set of commonly used modeling methods, including Maxent (ME), Random forest (RF), Artificial neuron networks (ANN), and Boosted regression trees (BRT), which are derived from appropriate packages. A set of parameters specific to the method used can be added as arguments to the *predict* function. When *folds* parameter is specified after creating random folds, SDMmodel object is output that hosts all the models trained during the cross-validation. It can be used to subsequently make tests of the models to assess and compare their performance. With nonparametric machine learning algorithms, cross-validation is often the only means to assess the accuracy and reliability of their predictions.

Four commonly used SDM methods have been examined: Maxent (ME), Random forest (RF), Artificial neural networks (ANN), and Boosted regression trees (BRT), their performance being compared. A special R script has been written for the purpose that takes a SDM method and its hyperparameters as an input. First, presence/background data are randomly divided into 6 folds, one of which being designated as a validation data set. Model is then run with input method and its hyperparameters, and its performance metrics are calculated. The process is repeated 20 times (every time with different random folds and validation data sets), with the purpose of calculating metrics means and standard deviations. Metrics means thus calculated are more stable than metrics values obtained in any single run, while metric standard deviation characterizes the stability of metric estimates among the different runs.

The number of hyperparameters amenable to tuning varies from 2 for ME to 5 for BRT. All of them except the size of hidden layer parameter for ANN method have got default values, though these are not always guaranteed to yield an optimal performance for the purpose. While some of the hyperparameters were chosen to be kept at default values, others were tuned with a view to achieve better performance, as indicated by appropriate metrics. Tuning models hyperparameters was performed with grid search method implemented in the function *gridSearch*. This function creates all possible combinations from an input range of possible values for hyperparameters and returns the values of the chosen evaluation metric for every possible combination so that the user can see the effect of varying the hyperparameters on the model performance and choose those values for hyperparameters that maximize the metric chosen.

Metrics employed to evaluate model performance were 1) area under the receiver operating characteristic (ROC) curve (AUC), 2) the true skill statistic (TSS). AUC is regarded as a threshold independent measure that assesses the discriminatory power of the model in separating presences from absences. TSS is defined as the sum of sensitivity and specificity of the discriminating capacity minus one. It was introduced to the assessment of SDM results in [1], where it is recommended as a simple and intuitive measure for the performance of species distribution models. In comparison with more widely used kappa statistic TSS measure is insensitive to prevalence while still keeping all the advantages of the former.

Initial models include 21 predictive variables (17 of which are related to climate and another 4 – to soil properties), many of which are significantly correlated. It looks desirable to reduce this number without compromising model performance, as more parsimonious models are usually characterized by smaller variance of the parameter estimates and are less prone to overfitting. The importance of separate variables for the model performance can be assessed with *varImp* function from SDMtune package. This function randomly permutes one variable at a time (using training and absence/background datasets) and computes the decrease in training AUC. Here such a “pruning” of predictive variables was achieved with *reduceVar* function. It removes variables with an importance lower than a given threshold in a stepwise fashion, starting from the variable with the lowest importance; however variables are removed only if the model performance after this does not decrease compared to the initial model, according to a given evaluation metric.

After a model with optimal hyperparameters values and a set of predictive variables has been built, it can be input to *predict* function to obtain the prediction maps of species occurrences. Model predictions can be regarded as the relative probabilities of species occurrence in the area. However, in conservation and environmental management practices the information presented as predicted species presence/absence may be more practical. To transform relative probabilities into presence/absence maps, complementary-log-log (cloglog) link function was applied to detect a value that maximizes the sum of sensitivity and specificity. This value is then applied as a threshold to relative probabilities maps.

When data on prospective distributions of predictive variables are available, future distributions of species habitats can also be forecasted based on models built on present-time data. Climate projections from 14 CMIP5 global climate models (GCMs) for three representative concentration pathways (RCPs) derived from WorldClim database were used as a data source for future climatic conditions, while soil conditions were supposed to be relatively stable, thus present-time values were directly used in forecasts.

A special R script has been written that takes as input one of 3 RCPs and one of two prediction periods (2050 or 2070) for which data are available. For each of 14 GCMs it downloads a raster stack of bioclimatic variables for the respective year and RCP, reprojects and crops it to the study area extent, drops unnecessary variables leaving only those present in the final model, renames layers, adds to them relevant soil properties layer(s), runs the model with these layers as an input, and adds the model prediction into a raster stack. When the predictions for all of the 14 GCMs have been accumulated in a stack, the median value of the stack is calculated and output as a final prediction for the given RCP and year. Median was chosen instead of mean because it is less subject to possible outliers in some model predictions.

Final predictions can be presented as relative occurrence probabilities maps or as predictive presence/absence maps after applying a threshold to the former. Based on these maps, prospective habitats areas can be calculated.

RESULTS AND DISCUSSION

Tuning of model hyperparameters with *gridSearch* was the first stage in model-building process. For the ME model, default value 1 for the regularization multiplier appeared suboptimal, and 0.75 was used as the one producing better output. For the RF model, the optimal values of the number of trees lie in the range 200–1000, with the default value of 500 being close to optimal. The best results for ANN were achieved with 12 units in the hidden layer and weight decay = 6. As to BRT model, the default number of trees = 100 seems to be close to optimal, the shrinkage parameter gave best results in the range from 0.01 to default 0.1, while the bagging fraction default value 0.5 seemed suboptimal, with those in the range 0.6–1 producing slightly better results. Thus, values chosen for the tuned model were 100 for the number of trees, 0.05 for shrinkage, and 0.8 for bagging fraction. It was found in general that most models are not especially sensitive to moderate variations in hyperparameters values. The exception is ANN that produced nonsensical results with default value of weight decay = 0 while quite good results appear when increasing this parameters to 2 and above.

The initial results of applying four mentioned SDM modeling methods using a full set of 21 predictive variables are shown on Table 1. Maxent method appears as inferior, while three other methods gave results of comparable accuracy, as seen in *testing* columns. Random forest method appeared to be prone to overfitting, as implied by 1 values of *AUC* and *TSS* metrics calculated for training dataset. It was impossible to statistically prove performance differences between ANN and BRT methods (their metrics means plus-minus their standard deviations overlap). ANN

still shows slightly better values for both of the performance metrics and slightly bigger differences in these values between testing and training columns, implying smaller variance (and smaller proclivity to overfitting).

Table 1. *Performance metrics of different SDM methods calculated on training dataset and with the aforementioned testing procedure. For testing case, standard deviations are given in parentheses.*

Method	***AUC (training)***	***AUC(testing)***	***TSS (training)***	***TSS(testing)***
Maxent (ME)	0.8162	0.796 (0.0138)	0.6114	0.5938 (0.0252)
Artificial neural networks (ANN)	0.9437	0.9369 (0.005)	0.7508	0.7891 (0.018)
Random forest (RF)	1	0.9321 (0.0076)	1	0.7851 (0.0261)
Boosted regression trees (BRT)	0.964	0.9304 (0.0062)	0.8145	0.775 (0.0162)

Applying *reduceVar* function to ANN model allowed to significantly reduce the number of predictive variables without compromising model performance metrics, thus making a model more parsimonious. Picking predictive variables was based on inspection of *reduceVar* function graphic output: variables were chosen that either were retained up to the later stages of the pruning algorithm run or which withdrawal led to relatively high decrease in values of performance metrics. Six variables out of initial set of 21 were thus chosen out for the final model; five of them relate to climate and another one to soil characteristics, that is: Annual precipitation (BIO12); Precipitation of wettest quarter (BIO16); Temperature seasonality (BIO4); Precipitation of coldest quarter (BIO19); Precipitation seasonality (BIO15); Soil acidity (phh).

Table 2 shows performance metrics of ANN model with these six predictive variables in comparison with a model with a full set of 21 variables. It can be seen that reducing the number of predictive variables to the six most important ones didn`t cause the decrease in performance metrics calculated with cross-validation on testing data. The direct result of applying *predict* function is a map of the relative probabilities of species occurrence in the area (Fig. 2, left). A threshold value 0.229 was used to convert it to habitat presence/absence map that looks more customary for practitioners(Fig. 2, right).

Table 2*. Performance metrics of ANN SDM method calculated with a full and reduced sets of predictive variables.*

Model	***AUC (training)***	***AUC(testing)***	***TSS (training)***	***TSS(testing)***
Full set of 21 variables	0.9437	0.9369 (0.005)	0.7508	0.7891 (0.018)
Reduced set of 6 variables	0.94	0.9381 (0.0036)	0.7384	0.7893 (0.0132)

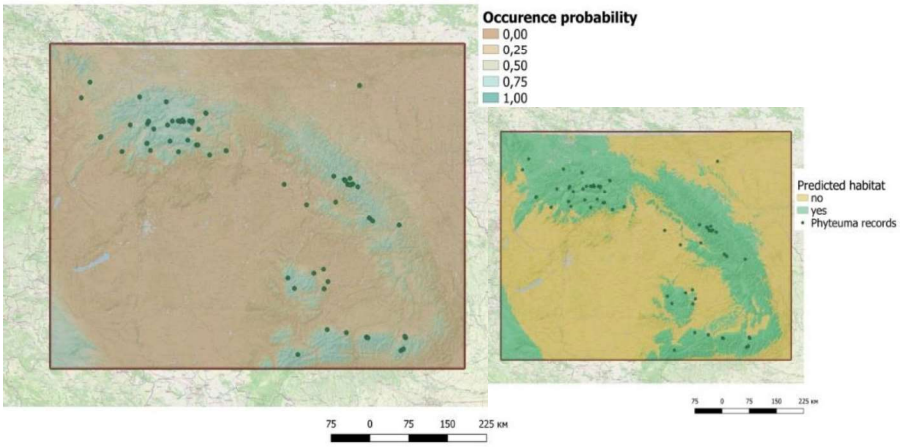

Fig. 2. *Relative occurrence probabilities (left) and predicted habitat (right) of Phyteuma.*

Predictions of future *Phyteuma* genus habitat under three different RCPs for years 2050 and 2070 were made with the same model, to which modified values of climatic variables reflecting assumed future climatic conditions were input. Results for RCP 50 and year 2050 are shown on Fig. 3. Comparing to Fig. 2, there are some spatial shifts: e.g. habitat area is predicted to somewhat increase in Eastern Carpathians while a decrease is expected in Southern Carpathians. Table 3 shows the expected habitat area changes for 2050 and 2070 under different RCPs. It shows that while habitat area for *Phyteuma* genus is expected to be stable or somewhat increase under moderate climate changes scenarios (RCP 26 and 45), the considerable decrease is expected for the most severe scenario RCP 85.

CONCLUSION

SDMs represent a valuable cost-effective tool to identify current important areas for threatened species that require attention from conservationists, and to forecast ecosystem impacts of rapid human-induced environmental changes. Machine learning approaches are becoming increasingly popular, facilitated by the recent availability of high computational power, and due to their ability to fit complex nonlinear relationships without requiring an a priori definition of a data model. Another important advance is the increased availability of open data on species observations and ecological factors. In the given case RF, BRT and ANN methods achieved results of similar accuracy, and reducing the number of predictive variables from 21 to 6 seemed feasible. An important prerequisite to successive modeling is the choice of predictive variables that are ecologically meaningful for the target species; the combination of ecological knowledge and statistical skill is thus needed to obtain reliable results.

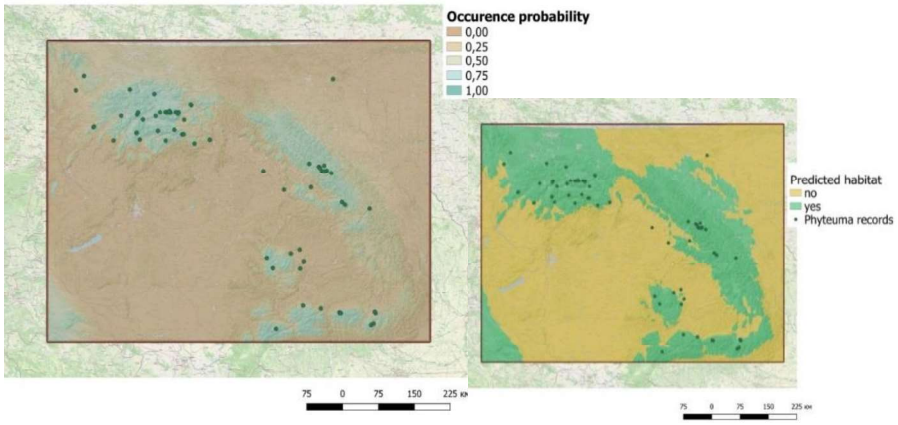

Fig. 2. *Prospected occurrence probabilities (left) and habitat area (right) of Phyteuma for year 2050 based on RCP45 climate projections.*

Table 3. *Predicted Phyteuma habitat area changes for years 2050 and 2070.*

RCP / Year	26	45	85
2050	178.9	166.9	149.2
2070	192.8	172.3	127.7

REFERENCES

[1] Allouche O., Tsoar A., Kadmon R., Assessing the accuracy of species distribution models: Prevalence, kappa and the true skill statistic (TSS), Journal of Applied Ecology, vol. 43/issue 6, pp 1223-1232, 2006.

[2] Fick S.E., Hijmans R.J., WorldClim 2: new 1km spatial resolution climate surfaces for global land areas, International Journal of Climatology, vol. 37/ issue 12, pp 4302-4315, 2017.

[3] Franklin J., Mapping species distributions: spatial inference and prediction, Cambridge University Press, Cambridge, UK, 2009.

[4] GBIF: The Global Biodiversity Information Facility, What is GBIF?, 2020. Available from https://www.gbif.org/what-is-gbif.

[5] Hengl T., Mendes de Jesus J., Heuvelink G.B.M., Ruiperez Gonzalez M., Kilibarda M., Blagotić A., SoilGrids250m: Global gridded soil information based on machine learning, PLoS ONE, vol. 12(2), 2017.

[6] Vignali S., Barras A.G., Arlettaz R., Braunisch V. SDMtune: An R package to tune and evaluate species distribution models, Ecology and Evolution, vol. 10/issue 20, pp 11488-11506, 2020.