37 (1), 2013: 64-68

http://ojs-cra.cilea.it/index.php/asr

Technical note

A spatial model for sporadic tree species distribution in support of tree oriented silviculture

Davide Melini[1*]

Received 09/12/2013 - Accepted 27/12/2013

Abstract - This technical note describes how a spatial model for sporadic tree species distribution in the territory of the Unione di Comuni Montana Colline Metallifere (UCMCM) was built using the Random Forest (RF) algorithm and 48 predictors, including reflectance values from ground cover – provided by satellite sensors – and ecological factors. The P.Pro.SPO.T. project - Policy and Protection of Sporadic tree species in Tuscany forest (LIFE 09 ENV/IT/000087) is currently carried out in this area with the purpose of initiating the implementation of tree oriented silviculture in the Tuscany forests. Tree oriented silviculture aims at obtaining both forest biodiversity protection and local production of valuable timber. After creating a map showing the probability of presence of sporadic tree species, it was possible to identify the most suitable areas for sporadic tree species which are under protection according to the regulation of the Tuscany Region.Using data and software provided free of charge, and applying the RF algorithm, distribution models could be developed in order to identify the most suitable areas for the application of tree oriented silviculture. This can provide a support to forestry planning that includes tree oriented silviculture, thus reducing its implementation cost.

Keywords - sporadic tree species, distribution, tree oriented silviculture, forest biodiversity, random forests, GRASS GIS

Introduction

Mori et al. (2007) and Pelleri et al. (2010) proposed and initiated the implementation of tree oriented silviculture in the woods of Tuscany. The P.Pro.Spo.T. project - Policy and Protection of Sporadic tree species in Tuscany forests (LIFE 09 ENV/IT/000087) - takes currently place in the "Colline Metallifere" area (southern Tuscany). This project includes demonstration areas about tree oriented silviculture, a technique in which target trees are identified and optimal conditions for their growth are created through different measures.

Investments in the application of tree oriented silviculture - favouring sporadic tree species - can realistically expect a financial return only in wooded areas whose ecological and microclimatic conditions are highly suitable for such species.

This technical note shows that it is possible to build cost-effective probabilistic models related to the presence of sporadic tree species. These models can be used to identify which forest areas are the most suitable for the application of tree oriented silviculture, thus supporting field sampling aimed at forest planning.

Materials and Methods

We developed the model using as reference the raster map of the forest areas in which the Unione di Comuni Montana Colline Metallifere (UCMCM) is responsible for forestry. The P.Pro.SPO.T. project takes place in this territory and includes demonstration areas. The 100 m resolution raster map - derived from Corine Land Cover (CLC) 2006 - covers 525.2 km^2 of forest area. The distribution model refers to the sporadic tree species which are under protection according to the article 12 of the *Regolamento Forestale della Toscana* (regional forestry regulation) 48/2003, which implements the Legge Forestale della Toscana (Tuscany regional law) 39/2000. These species can provide valuable timber and include for example *Acer* sp., *Fraxinus excelsior*, *Fraxinus oxycarpa*, *Tilia* sp., *Pyrus* sp., *Malus* sp., *Sorbus* sp., *Ilex aquifolium*, *Taxus baccata*, *Prunus avium*.

The model calculates - on the basis of 48 predictors - the degree of ecological suitability of the forest land for the sporadic tree species and it expresses their probability of presence in wooded areas. We used the Random Forest (RF) algorithm (Breiman 2001, Liaw and Wiener 2002), which is a variant of the Classification And Regression Trees (CART) algorithm. The RF algorithm classifies objects using predictors. Compared to other traditional methods, this system can be used to approximate any unknown function, even if it is nonlinear and involves complex interactions (Strobl et al. 2009). RF models split data into homogeneous groups using the infor-

[1] Independent researcher
[*] corresponding author: davide.melini@gmail.com

http://dx.doi.org/10.12899/ASR-751

mation provided for their training and can even deal with highly correlated predictor variables (Strobl et al. 2008). They use an ensemble of classification (or regression) trees that are calculated on random subsets of the data, using a subset of randomly restricted and selected predictors for each split in each classification tree (Strobl et al. 2008). RF models are implemented in the homonym library of R (R Development Core Team 2013), an open source software environment for statistical analysis. This procedure can be interfaced with the GRASS (Geographic Resources Analysis Support System) software (GRASS Development Team 2012) on Windows, Mac OS X and GNU/Linux systems using the spgrass6 library. The model for the sporadic tree species probability of presence was built using the data collected in 83 plots established in 1993, when the Tuscany Forest Inventory was conducted (Regione Toscana 1998). 80% of the data sets were used for training the RF model, 20% were used for model validation. The 80/20 proportion is widely used for splitting data sets in machine learning algorithm applications according to the Pareto principle, which states that, for many events, 80% of the effects come from 20% of the causes (Boslaugh and Watters 2008). That is why we considered 20% of the data as an appropriate proportion for the validation set.

We tested other methods - the k-nearest neighbors (k-NN) algorithm and the artificial neural networks (ANN) - with the same data set, in order to compare them with the RF method. The k-NN algorithm is a non-parametric method for classification and regression that predicts objects values or class memberships, based on the k closest training examples in the feature multidimensional space (Cunningham and Delaney 2007). ANN are computational models inspired by the brain of animals, which are capable of learning and pattern recognition. They are usually presented as systems of interconnected neurons which can compute values from inputs by feeding information through the network (Anil K. Jain et al. 1996).

We used the k-NN method as implemented in the KKNN (Schliep and Hechenbichler 2013) and the KNNFLEX (Dunlap Brooks 2007) R libraries. The data set was divided to compare the RF algorithm with the k-NN and the ANN methods. The RF method itself does not require a cross-validation: an unbiased estimate of the test set error is calculated internally during the run (Breiman 1996) as follows. Each tree is constructed using a different bootstrap sample from the original data. About one-third of the cases are left out of the bootstrap sample and not used in the construction of the k^{th} tree. Each case left out in the construction of the k^{th} tree is used in the k^{th} tree to get a classification. In this way, a test set classification is obtained for each case in about one-third of the trees. At the end of the run, if “j” is the class that got most of the votes every time case “n” was out-of-the-bag (OOB), the proportion of times that “j” is not equal to the true class of “n” averaged over all cases is the OOB error estimate. Through this process, we can also get an indirect estimation of the classification accuracy.

The RF algorithm was selected because it performed the best classification accuracy, calculated through the confusion matrix (a table where each column represents the instances in a predicted class, while each row represents the instances in an actual class).

In order to assess the directional variability of the predictors, we calculated 48 raster maps using the statistics of minimum, maximum and standard deviation, which are calculated using a 3 x 3 cell neighborhood in each input layer.

The RF algorithm was applied to a database that was obtained by extracting values from the digital maps of predictors in the locations of the forest inventory plots. These values were then added to the data regarding the presence of sporadic tree species, derived from regional forest inventory. These data were used to build a model which describe the presence of sporadic tree species.

We built the model using 500 classification trees. According to the RF method, each classification tree provides a classification and it is said that the tree "votes" for that class. The algorithm chooses the classification that gets the majority of the votes (Emmert-Streib and Dehmer 2010). Different numbers of trees were tested as well, showing no statistically significant differences in the classification accuracy. After that, we applied the model to the predictors' raster maps, obtaining a map showing the probability of presence of the sporadic tree species. The resulting raster map was processed by reclassifying the values from 0 to 0.5 as 0 (1^{st} class) and the values from 0.51 to 1 as 1 (2^{nd} class). The areas with a higher probability of presence (2^{nd} class) were considered the most suitable for the application of tree oriented silviculture. We chose a reclassification with 0.5 as threshold because we wanted to adopt very selective criteria in the identification of the areas that are suitable for the tree oriented forestry application.

Reflectance values (Landsat 7 ETM+ 1, 2, 3, 4, 5, 6 bands)

We used the reflectance values provided by a Landsat 7 ETM+ scene for the spectral bands 1, 2, 3 and 4 (red and near IR), 5, and 6. Particularly red and infrared values express the absorption of radiation in the wavelengths used by the photosynthesis

and describe the functionality of ecosystems. The Landsat scenes of 2006 are available at the URL http://glcfapp.glcf.umd.edu:8080/esdi/index.jsp. A principal component analysis was conducted using the i.pca module of GRASS software.

Normalized Difference Vegetation Index (NDVI)

This index - calculated through the equation [1] using the reflectance values of bands 3 and 4 provided with Landsat 7 ETM+ - assesses the absorption of photosynthetically active solar radiation and describes the ecological functionality of forest ecosystems (Lloyd 1990).

$$NDVI = (band\ 4 - band\ 3) / (band\ 4 + band\ 3) \quad [1]$$

Altitude, slope, aspect

The terrain morphology influences many ecological factors. Altitude, exposition and slope are the driving factors determining the sliding velocity of water on the ground surface, the prevailing winds and the available solar radiation. Maps of slope and aspect were derived from a DTM (Digital Terrain Model). The DTM has a 90 m resolution and is provided by the Consortium for Spatial Information http://srtm.csi.cgiar.org. Elevation data were collected during the SRTM (Shuttle Radar Topography Mission), which provided data on a global scale to generate a high-resolution digital topographic database of the Earth. The SRTM consisted of a radar system that flew aboard the Space Shuttle Endeavour during an 11-day mission in February 2000. Despite the availability of DTMs with a higher resolution, we chose to use this one because we considered a 100 m resolution to be adequately precise for the purposes of the present work. We resampled the 90 m resolution SRTM DTM to 100 m resolution using the GRASS GIS software (nearest neighbor resampling).

Surface water accumulation

The calculation of the accumulation of water flows was performed using the GRASS software. This software provides an output map in which the absolute value of each cell represents the amount of overland flow that moves across the cell coming from cells at higher altitudes. The areas with greater water availability were identified calculating the logarithm of accumulation. Due to the wide extension of the study area, the flow accumulation values have a very wide range. Therefore we chose to calculate the logarithm of the flow accumulation (Catani et al. 2013).

Average annual precipitation

The average annual precipitation map was obtained by interpolation using the rainfall data of the Regional Hydrological Service (www.sir.toscana.it). These data were recorded in 59 stations in the provinces of Grosseto and Siena and refers to the years between 1919 and 2005. Unfortunately, the availability of these data was not homogeneous and we therefore we selected only the stations reporting data from at least 10 consecutive recent years. We derived the values of average annual precipitation and provided the coordinates of the stations in order to create a georeferenced database in vector format.

The interpolation was conducted using the Universal Kriging method with elevation (provided by SRTM DTM) as external drift, implemented in the gstat library (Pebesma and Edzer 2004) of R software. The Universal Kriging method was used because it performed better compared to other methods that have been proved to be useful at regional scale for bioclimatic data interpolation (Attorre et al. 2007). We verified the normal distribution and the absence of autocorrelation of the errors. The variogram model we chose was exponential and the interpolation was performed using a range of 30 km.

Results

The ACC (accuracy) of the different models was calculated using the cross-validation confusion matrix. The RF method predicted with an ACC of 0.72; the best result using the k-NN algorithm was an ACC of 0.58. The ANN predicted with a maximum ACC of 0.64.

A result map showing the probability of presence of sporadic tree species (Fig. 1) was created applying the model built with the RF algorithm to the predictors' raster maps. The probability of presence of sporadic trees species in the result map (obtained applying the model built with the RF method) varies between 0.12 and 0.88. The areas with a probability of presence from 51% to 100% (2nd class) cover a surface of 3,255 ha: 1,560 ha belong to the Tuscany Region properties, the remaining areas are private

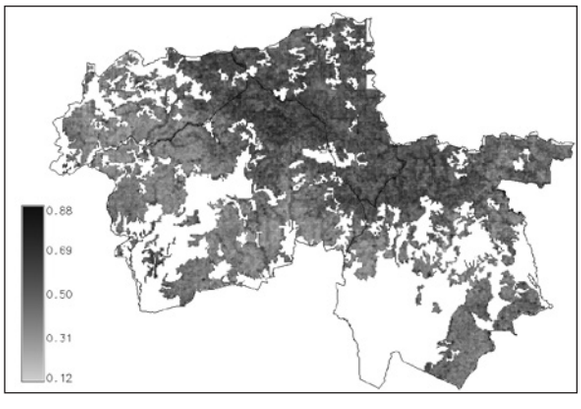

Figure 1 - Raster map on the probability of presence of sporadic tree species in the UCMCM territory, built using the RF algorithm.

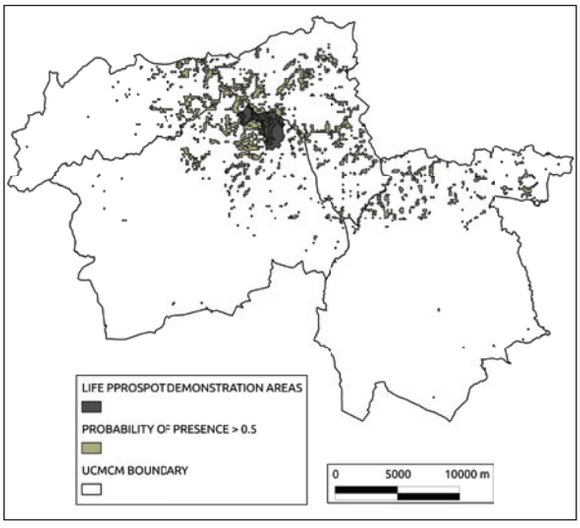

Figure 2 - Distribution of wooded areas where the probability of presence of sporadic tree species is from 51% to 100%. These surfaces include a large part of LIFE+ P.Pro.Spo.T. demonstration areas.

properties, mainly located in mesic deciduous hardwood forests. In addition, most of the forests chosen as demonstration areas for the P.Pro.Spo.T. project fall within the "high probability of presence" surfaces identified through the model (Fig. 2).

The terrains situated in low altitude areas and presenting a high probability of presence of sporadic tree species are located along the main watercourses and/or are exposed to the North.

Discussion

In the construction of the model of distribution of sporadic tree species, the RF algorithm performed better in terms of classification accuracy compared to the other non-parametric methods used (k-NN and ANN). The problem we examined - the modelling of the distribution of a set of sporadic tree species with different ecological characteristics - appeared to be a very complex matter. For this reason, we consider the results obtained with the k-NN and ANN methods satisfactory as well and useful for the development of the model. It is our opinion that the better performance of the RF algorithm depends on the logical and mathematical features of this method, which do not perform only a single classification during the model building, but rather a large number of classifications (500 in the basic configuration). Furthermore, the classification is conducted with the automatic selection of a random sample containing about 70% of the data, using a subset of randomly restricted and selected predictors for each split in each classification tree. This allows the comparison of all the 500 or more different and independent classifications, assuring the selection of the most correct one in each case.

Conclusions

Building a spatial model using the RF algorithm we identified a surface of 3,255 ha - within the UCMCM wooded area - where the probability of presence of sporadic tree species is greater than 50%. This model provides useful information for forest planning both on private and public properties. For instance, it is possible to identify suitable areas for the application of tree oriented silviculture corresponding to over 6% of the forest surfaces where the UCMCM is the competent authority. This model is built using data and software provided free of charge and it can support low cost forest planning including tree oriented silviculture. It allows to overcome the need for preliminary field investigations aimed at verifying the presence and the possibility of diffusion of sporadic tree species, thus reducing the cost of forest planning aimed at the application of tree oriented silviculture.

Acknowledgments

The author would like to thank Francesco Pelleri (CRA-SEL Arezzo) for his expert advice on ecological characteristics of sporadic tree species, as well as the anonymous reviewers for their helpful observations.

References

Attorre F., Alfò M., De Sanctis M., Francesconi F., Bruno F. 2007 - *Comparison of interpolation methods for mapping climatic and bioclimatic variables at regional scale.* International Journal of Climatology 27 (13): 1825-1843.

Boslaugh S., Watters P.A. 2008 - *Statistics in a Nutshell.* O' Reilly Media, Inc. 480 p.

Breiman L. 1996 - *Technical report. Out-of-bag estimation.* Department of Statistics, University of California, 13 p.

Breiman L. 2001 - *Random forests.* Machine Learning 45 (1): 5-32.

Catani F., Lagomarsino D., Segoni S., Tofani V. 2013 - *Landslide susceptibility estimation by random forests technique: sensitivity and scaling issues.* Natural hazards and Earth System Sciences 13: 2815-2831

Cunningham P., Delaney S.J. 2007 - *k-Nearest Neighbor Classifiers. Technical Report UCD-CSI-2007-4.* School of Computer Science and Informatics, University College Dublin, 17 p.

Dunlap Brooks A. 2007 - *Knnflex: a more flexible KNN.* [Online]. Available: http://cran.cermin.lipi.go.id/web/packages/knnflex/index.htm [2013]

Emmert-Streib F., Dehmer M. 2010 - *Medical biostatistics for complex diseases.* Wiley-Blackwell Publishing, 30 p. GRASS Development Team 2012 - Geographic Resources Analysis Support System (GRASS) Software. Open Source Geospatial Foundation Project. [Online] Available: http://grass.osgeo.org [2013].

Jain A.K., Jianchang M., Mohiuddin K.M. 1996 - *Artificial neural networks: A tutorial.* IEEE Computer 29 (3): 31-44.

Liaw A., Wiener M. 2002 - *Classification and regression by Random Forest.* R News 2 (3): 18-22. [Online]. Available: http://CRAN.R-project.org/doc/Rnews/ [2013].

Lloyd D. 1990 - *A phenological classification of terrestrial vegetation cover using shortwave vegetation index imagery.* International Journal of Remote Sensing, 11: 2269 - 2279.

Mori P., Bruschini S., Buresti Lattes E., Giulietti V., Grifoni F., Pelleri F., Ravagni S., Berti S., Crivellaro A. 2007 - *La selvicoltura delle specie sporadiche in Toscana.* Manuale n. 3 "Supporti tecnici alla Legge Regionale Forestale della Toscana". Editori ARSIA e Regione Toscana, 366 p.

Pebesma E., Edzer J. 2004 - *Multivariable geostatistics in S: the gstat package.* Computers & Geosciences 30: 683-691.

Pelleri F., Giulietti V., Sansone D., Samola A., Nitti D. 2010 - *Valorizzazione delle rosacee arboree. Esperienze nei cedui delle Colline Metallifere (GR).* Sherwood - Foreste ed Alberi Oggi 160 (2): 5-11.

R Development Core Team 2013 - R: *A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0. [Online]. Available: http://www.R-project.org. [2013].

Regione Toscana 1998 - *Boschi e macchie di Toscana, volume 3, Inventario Forestale.* Edizioni Regione Toscana, 219 p.

Schliep K., Hechenbichler K. 2013 - *Knn: Weighted k-Nearest Neighbors Classification, Regression and Clustering.* [Online]. Available: http://cran.r-project.org/web/packages/kknn/index.html [2013]

Strobl C., Boulesteix A.L., Kneib T., Augustin T., Zeileis A. 2008 - *Conditional variable importance for random forests.* BMC Bioinformatics 9: 307.

Strobl C., Malley J., Tutz G. 2009 - *An introduction to recursive partitioning: rationale, application, and characteristics of classification and regression trees, bagging, and random forests.* Psychological methods 14 (4): 323-348.