Scientific
Research
Publishing

Predicting Cork Oak Suitability in Maâmora Forest Using Random Forest Algorithm*

Said Lahssini[1], Hicham Lahlaoi[2], Hicham Mharzi Alaoui[3], El Aid Hlal[1], Martino Bagaram[1], Quentin Ponette[4]

[1]National School of Forestry Engineering, Salé, Morocco
[2]Geosciences Laboratory, Faculty of Sciences Ain Chock, Hassan II University, Casablanca, Morocco
[3]Institut of Agronomy and Veterinary Hassan II, Rabat, Morocco
[4]Catholic University of Louvain, Louvain la Neuve, Belgium
Email: marghadi@gmail.com

Abstract

Maâmora is considered the most important cork-oak forest in the world with regard to surface. Therefore, anthropic pressure, including cork harvesting, grazing and soft acorn picking up by local communities, has harmful consequences on forest regeneration and the forest become older exceeding harvesting age. Thus, its sustainability depends on the managers' ability to succeed cork oak plantations. This work presents an assessment approach to evaluate *Quercus suber* suitability to its plantation which is based on a random forest algorithm (RF). In fact, this suitability has been assessed through analyzing management data related to previous plantation success rates (SR). Then a relationship between SR and a set of environmental and social factors has been investigated using the RF. Application of the fitted model to continuous maps of all involved factors enabled establishment of suitability maps which would help managers to make more rational decisions in terms of cork oak regeneration, ensuring Maâmora forest sustainability.

1. Introduction

Cork oak forests cover 350,000 ha in Morocco. This is about 15% of world's cork oak area. However, it contributes only for 4% to 6% in the world's cork production [1]. Many factors are involved for explaining this situa-

---

*Using random forest algorithm for assessing cork oak suitability in Maâmora-Morocco.

tion (social pressure, intensive harvesting, climate) and regressive trends of these forests have been described [2].

Maâmora as example is considered one of the largest cork oak forests in the world; it covers 133,000 ha. Unfortunately, cork oak covers only 70,383 ha. Since 1940, Maâmora has been managed according to four management plans. These plans revealed a cover regression trend which has been estimated at 30% of the cork oak area [2]. Therefore, since 1950, overgrazing and acorns picking has been considered a major obstacles to cork oak regeneration [3], since then regeneration problems still prevail and overgrazing rate is currently estimated at 81.57% [2].

To deal with this alarming situation of non-regeneration, forest managers have undertaken huge efforts for studying and regenerating cork oak. Results have been qualified as random [4]. Consequently, administration maintained and continued rehabilitation of cork oak. Thus, 20,000 ha have been planted under the scope of 2005-2014 program.

Cork oak regeneration studies in Maâmora pointed out that water balance (climate and soil factors) associated with the high anthropogenic pressure are the main factors which determine the success of planted areas. Moreover, previous experiments showed that initial density, soil type and waterlogging have a strong influence on the ability of the strains to reject. Further, Maâmora's soils consist on a sand layer resting above a clay layer. Thus, sand depth and slope of underground clay floor seems to be crucial for cork oak seedlings success [4] [5]. Besides, rains follow a decreasing gradient from west to east by the distance to ocean and the location of the stands.

Despite that soil factors have been well documented in the Maâmora forest, no soils clay layer depth/slope or a soil suitability to cork oak regeneration maps are available for all the Maâmora. This could explain same cases of regeneration failure.

Habitat suitability models aim predicting lands which fit well for a species according to specified requirements, preferences, or predictors [6]. As suitability analyzes the interactions among three types of factors: location, development actions, and environmental elements [6], they constitute a good tools for supporting decision making which improve the understanding of species habitat relationship in space and time and predict their likelihood of occurrence and abundance or performance based on habitat attributes affecting survival, growth and reproduction of considered species [7] [8].

Habitat suitability models can be classified according to the approach and to the classification function used [9]. Nowadays, predictive techniques have been improved with a direct consequence on models quality and credibility [10]. Further, machine learning (ML) methods become promising tools for assessing suitability.

ML is nonparametric techniques for regression or classification which present some advantages compared to other statistical methods. They are able to: 1) deal with complex relationships between predictors; 2) process nonlinear relationships between predictors; 3) deal with large amounts of data; 4) process complex and noisy data and 5) don't need any assumptions related to variable distribution [11].

Among machine learning methods, many methods have been used for suitability assessment [12] such as classification and regression trees [13], artificial neural networks [14], heuristic methods such as genetic and simulated annealing algorithms [15] [16], and Random Forests [17]. Several studies showed that Random Forest (RF) models, often reach top predictive performances compared to other methodologies [10] [18] [19].

As summarized by Cutler *et al.* (2007), RF applications include: 1) regression; 2) survival analysis; 3) imputation of missing values; 4) clustering, multidimensional scaling and detecting general multivariate structure through the unsupervised learning; and 5) classification.

In this work, we used RF as a classification method and to explore relationship between observed cork oak plantation's success rate and some environmental factors judged important. Our paper is organized as follows: 1) part 1: materials and methods which consists on a brief description of the study area, factors considered and modelling framework including a general description of the modeling procedures, the random Forest model and predicted maps; 2) part 2: results will illustrate the application of RF to Maâmora forest to obtain estimation of potential plantation success rate and discuss the results.

2. Materiel and Methods

2.1. Study Area

Maâmora is located in the north of Morocco (**Figure 1**). The study area precipitations range from 400 mm in the
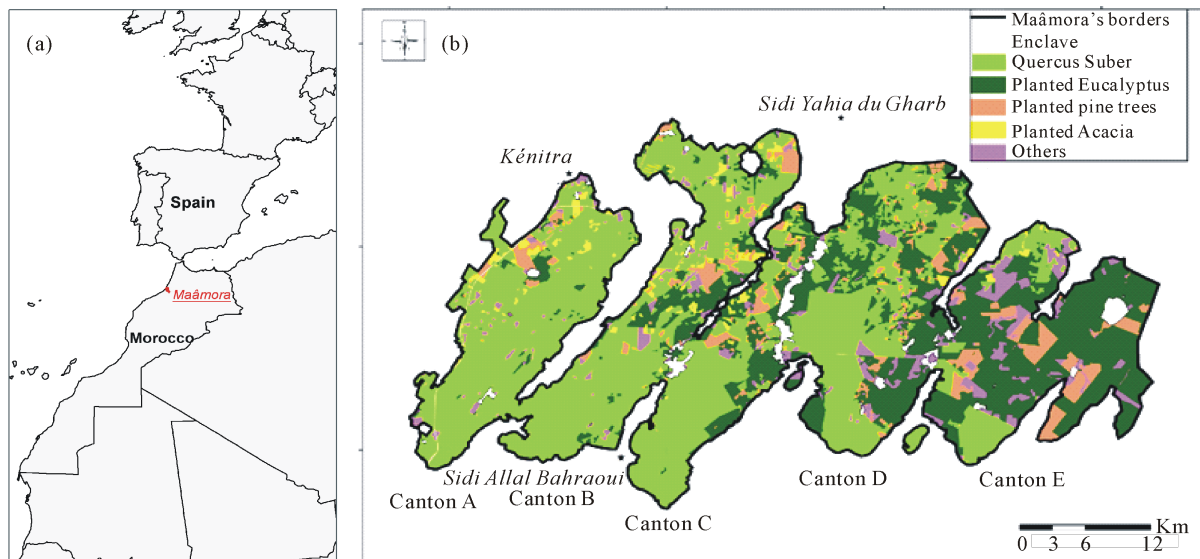
Figure 1. Localisation of the study area (a) in Morocco; (b) main tree species in the Maâmora forest.

East to 600 mm in the West 50 km far from the East. Bioclimatic conditions range from subhumid in the Western part to the semi-arid in the Central and Eastern part. Soils encountered consist of a sand layer resting on a clay layer; the transition between soil layers can be gross to net [5]. Further, sand thickness range between 30 cm and 6 m.

Maâmora offer a multitude of goods and services for administration and local communities. Thus, it produces per year 60,000 cubic meters of cork, 100,000 cubic meters of fuel wood from Eucalyptus, 21 million feed units, and incomes exceeding 70 million dirhams (1 USD = 10.4 dirhams). Also, many other services are offered to communities: recreation for 4 neighbor cities, hunting, soil protection and recharging the groundwater. Further, Maâmora's non wood incomes are estimated to US$440/ha/year [20].

2.2. Factors Involved

Variables used as predictors were climatic, topographic, soil and social. The topographic variable considered is altitude which was derived from digital elevation model SRTM DEM. Soil variables considered are kind, sand depth and slope of the clay layer. Maâmora's soil map has been used [21]. Maps of sand thickness and clay layer have been used form a previous work [22].

Climatic variables are: annual precipitation, minimum average temperature of the coldest month, maximum average temperature of the warmest month and Emberger quotient Q2 [23]. Climatic variables have been extracted from World climatic database [24].

Social factor has been modeled as distance from rural area (douar) to the forest stands. Since, anthropogenic pressure is higher around local communities that abuse for uses rights.

As response variable, success rates grades have been considered. Thus, all planted perimeters since 1970 have been mapped (**Figure 2**).

2.3. Modelling Framework

2.3.1. Modelling Procedure

Land suitability modelling process consists of four phases: model selection, training, prediction and final map production [10]. In the scope of our work, modelling procedure is described in **Figure 3**. Further, model selection and validation are based on training RF on the available data (historical regeneration maps on **Figure 2**). Thus, 2500 sample cell within regeneration areas have been chosen randomly. This data set has been randomly divided to 2 data sets with a respective ratio of 4/5 and 1/5: first part is the training set (Train DS) and the second one is the evaluation set (validation DS). On each cell of the two data sets, environmental and social factors considered as predictors have been extracted with their corresponding observed regeneration success rate (five
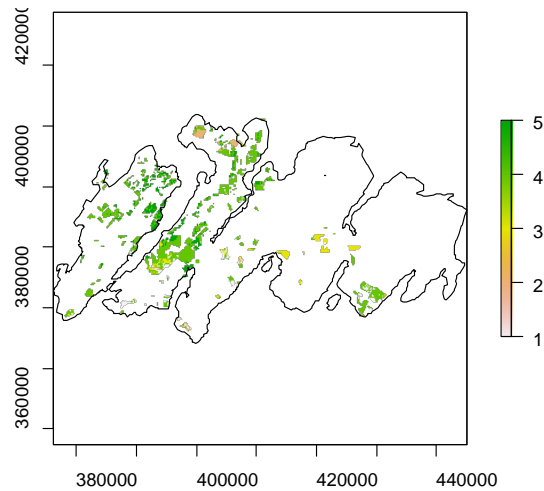
Figure 2. Map of areas previously planted by cork oak. Numbers from 1 to 5 refers to observed success rates: 0% - 20%, 20% - 40%, 40% - 60%, 60% - 80%, and 80% - 100%.
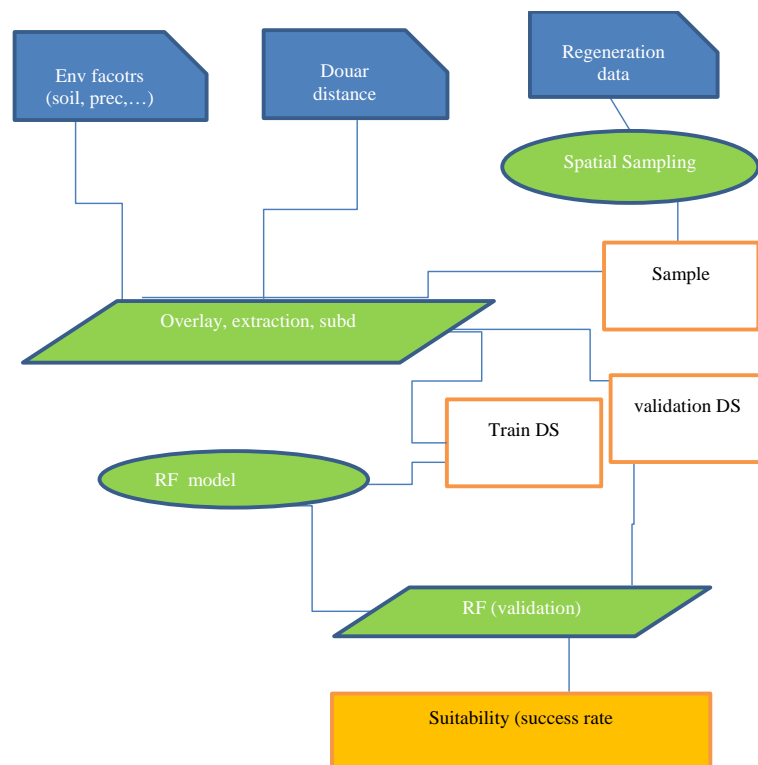

Figure 3. Modelling procedure.

classes ranging from failure to success have been identified and numbered from 1 to 5: 0% - 20%, 20% - 40%, 40% - 60%, 60% - 80%, and 80% - 100%). Then, the first data set was used to develop the RF model. The second data set was used to evaluate the quality of model and to compare predicted and observed regeneration success rate.

2.3.2. Used Software
The process was mainly implemented within the R open source software environment for data analysis and sta-

tistical computing [25]. Many R packages have been used: "rgdal" for reading GIS data files, "sp" for spatial data handling, "raster" for raster and terrain analysis, "gstat" for kriging, "randomForest" [26] which implement the RF algorithm.

2.3.3. Random Forest Algorithm

RF implements an automatic combination of tree predictors and the base model is achieved through training on different bootstrap replicates of the data [27]. Each tree is trained by selecting a random set of variables and a random sample from the training dataset. As consequence to this feature, the problem of correlated variables will alleviate because they may be extracted in turn [10].

RF needs three parameters to be defined: "ntree" (number of bootstrap samples for the original data), "mtry" (number of different predictors tested at each node), "nodesize" (minimal size of the terminal nodes of the trees). According to many RF algorithm users, it is often convenient to improve the quality of the model by choosing an optimal value for "mtry". Liaw and Wiener (2002) proposed a default value of one third of the predictors number.

Moreover, the "ranfomForest" package within R has a tuneRF function that is specifically for searching for the "optimal" value for "mtry". Also, it produces two useful plots: variable importance (measured as the deterioration of the predictive ability of the model when each predictor is replaced by a random noise) and partial plots (measures the marginal effect of predictor in RF estimates of success rate). In addition, it computes the OOB ("out of bag") error estimate, which is computed for each tree over the data remaining out of the corresponding bootstrap sample, and then averaged [27].

2.3.4. Predictive Maps

Once the RF model has been calibrated and validated, it was applied to the entire Maâmora to appreciate the suitability to cork oak and to estimate the predicted success rate of cork oak plantation.

During the validation process, confusion matrix has been used to measure the quality of classification results obtained for the independent dataset (validation dataset). Further, the Kappa coefficient assessed the quality of classification results observed during previous plantation and those predicted by the RF by using a per-pixel approach.

3. Results

The out-of-bag estimates of the error rate (ERROOB) were used to select the optimum Random Forest parameters. For the calibration dataset (CD), the Random Forest was able to classify the samples with an out-of-bag estimate of error rate less than 10% (OOB error = 8.85%).

Figure 4 shows the ranking of predictors by their importance measured as the increased mean square error (%IncNodePurity) representing the deterioration of the predictive ability of the model when each predictor is replaced in turn by a random noise. Higher %IncNodePurity indicates greater variable importance. Only four of the predictors contributed noticeably to the estimation of suitability grade, namely precipitation (Prec), elevation (Alt), temperature (m & M) and Emberger's quotient (Q2).

Moreover, **Figure 5** shows partial plots which represent the marginal effect of a single factor included in the RF model on estimates of cork oak plantation suitability. In these partial plots of marginal effects, only the range of values can be compared between plots of different variables.

The RF model provided a good prediction of success rate grade for regeneration in the validation dataset. Per class OOB errors ranges from 7.3% for the fourth classes to 16% for the third class. Further, using a per-pixel approach for classification quality assessment (observed success rates from previous plantation vs. predicted ones) through a kappa index gives a value of 0.827 which mean an almost perfect agreement [28].

The RF model application to the whole Maâmora is shown in **Figure 6**. The most suitable area (4th and 5th classes: with success rates greater than 60%) are located on the western part of the forest. While the eastern part represent the lowest predictions scores of suitability. This can be explained by continentally and the rainfall gradient.

The RF algorithm used does not require any assumptions related to normality of the variables and can deal with nonlinear relationships. These features may have a substantial utility within the context of forest management and could be used to model distribution shifts resulting from climate change. Moreover it constitutes a
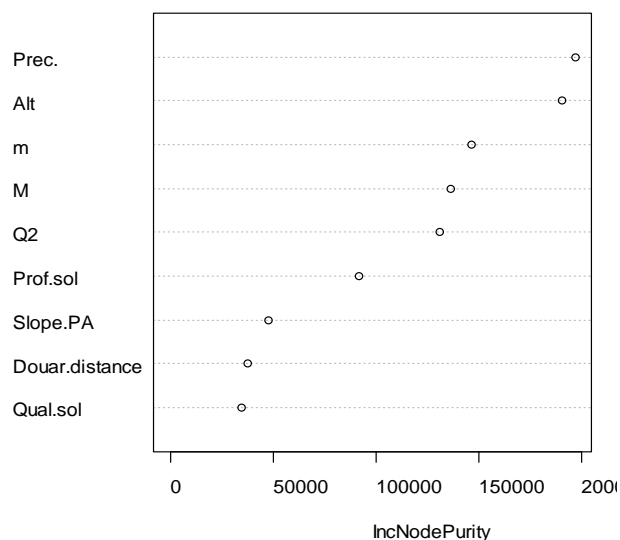
Figure 4. Variable importance plot generated by the random forest algorithm included within the "randomForest" package for R software.
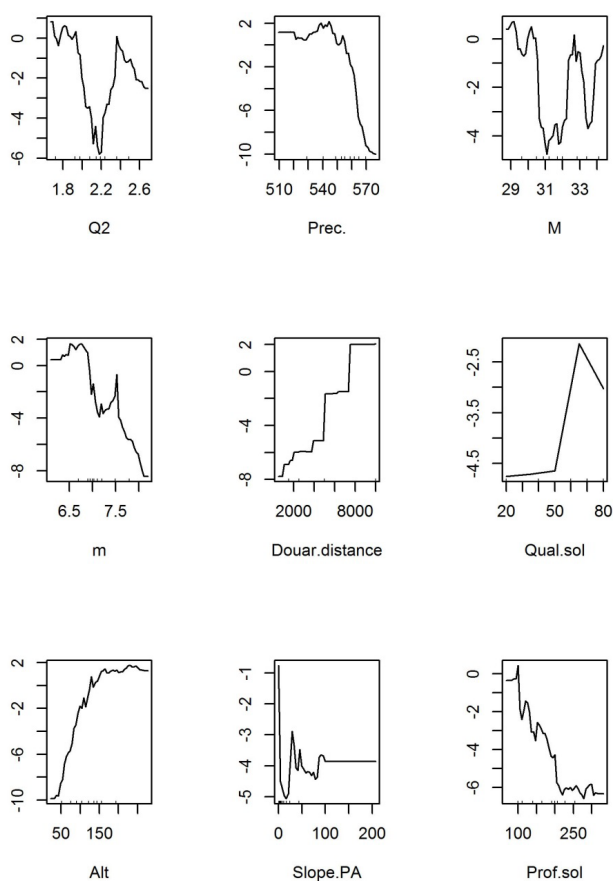

Figure 5. Partial plots generated by the RF included within the "randomForest" package for R software. It represents the marginal effect of single variables included in the model on estimates of suitability level while averaging out the effect of all the other variables.
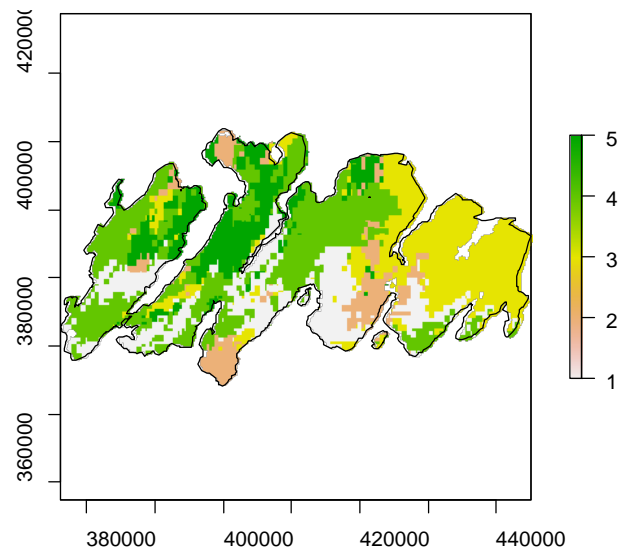
Figure 6. Predicted cork oak plantation suitability map for Maâmora.

promoting approach [7] [10]. Results obtained with the random forest method for predicting suitability maps are very encouraging; they can predict the success rate grades of regeneration. Further, compared to other approaches: LepoutreAbac [5] or suitability using AHP [22], RF presents a highest predictive accuracy.

The RF variable importance measure indicated precipitation and altitude as important factors in the modelling (**Figure 5**). Even if precipitation is expected because of the Mediterranean climate of the Maâmora Forest, it was not expected that the altitude is also important as other factors.

When comparing the observed success rates from previous plantations with the predicted ones by RF model, the kappa statistic used indicate a good accuracy of the RF model.

As stated by Benito Garzón *et al.* (2006), biological validation and interpretation of results obtained by modelling process are very important. Thus, our results are encouraging because they agree with the main bibliographic conclusions [4] [5] [22].

As compared to the larger previous cork oak distribution in the Maâmora Forest, what can be observed now is the result of historical and social dynamic on which area have been reduced mainly by intense anthropic activity (overgrazing, shifting species, etc.). Furthermore, Plantation success rates are not only affected by climatic, topographic, soil and social factors but with other factors. Moreover, seedling production and installation (seeds origin/quality, seedling production process, soil preparation and improvement, etc…), anthropogenic pressure, overgrazing and global warming issues, among others, have probably determined the current observed success rates. Models results should therefore be interpreted with full knowledge of these inevitable limitations. Assuming these limitations, the modelling framework may offer a good tool for ensuring sustainability of Maâmora forest, hence giving to the managers possibilities for planning regeneration and rehabilitation activities.

4. Conclusions

Succeeding *Quercus suber* plantation is a challenging operation for Maâmora forest managers. Indeed, suitability map remains a valuable tool which is very helpful to forest managers for planning forest regeneration and ensuring sustainability. The main purpose of this article is to produce this map.

The use of random forest algorithm made it possible to achieve the overall objective of this work. The produced map enables managers to choose the most suitable area to be planted and to guarantee the success of their regenerations efforts. Furthermore, it ensures suitability of the forest. The resulting map helps managers to make more rational decisions related to cork oak regeneration and to ensure the sustainability of Maâmora forest.

Acknowledgements

the project "Comportement des plantations du chêne liège de la forêt de Maâmora, Maroc".

References

[1]   Hammoudi, A. (2002) La Subéraie: Biodiversité et Paysage. Viveexpo 2002.

[2]   HCEFLCD (2011) Procès-verbal d'Aménagement de la Forêt de la Maâmora (Version Prel.), HCEFLCD-Rabat, Maroc.

[3]   Marion, J. (1951) La Régénération Naturelle du Chêne-liège en Maâmora. *Annales de la recherche forestière au Maroc*, 25-57.

[4]   Belghazi, B., Badouzi, M., Belghazi, T. and Moujjani, S. (2013) Semis et Plantations dans la Forêt de Chêne-liège de la Maâmora (Maroc). *Forêt Méditerranéenne*, **XXXII**, 301-314.

[5]   Lepoutre, B. (1965) Régénération Artificielle du Chêne-liège et Equilibre Climacique de la Subéraie en Forêt de la Maâmora. *Annales de la recherche forestière au Maroc*, **9**, 1-188.

[6]   Collins, M.G., Steiner, F.R. and Rushman, M.J. (2001) Land-Use Suitability Analysis in the United States: Historical Development and Promising Technological Achievements. *Environmental Management*, **28**, 611-621. http://dx.doi.org/10.1007/s002670010247

[7]   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. http://dx.doi.org/10.1016/j.ecolmodel.2011.02.007

[8]   Santos, X., Brito, J.C., Sillero, N., Pleguezuelos, J.M., Llorente, G.A., Fahd, S. and Parellada, X. (2006) Inferring Habitat-Suitability Areas with Ecological Modelling Techniques and GIS: A Contribution to Assess the Conservation Status of *Vipera latastei*. *Biological Conservation*, **130**, 416-425. http://dx.doi.org/10.1016/j.biocon.2006.01.003

[9]   Ortigosa, G.R., De Leo, G.A. and Gatto, M. (2000) VVF: Integrating Modelling and GIS in a Software Tool for Habitat Suitability Assessment. *Environmental Modelling & Software*, **15**, 1-12. http://dx.doi.org/10.1016/s1364-8152(99)00029-8

[10]   Benito Garzón, M.B., Blazek, R., Neteler, M., Dios, R.S. de, Ollero, H.S. and Furlanello, C. (2006) Predicting Habitat Suitability with Machine Learning Models: The Potential Area of *Pinus sylvestris* L. in the Iberian Peninsula. *Ecological Modelling*, **197**, 383-393. http://dx.doi.org/10.1016/j.ecolmodel.2006.03.015

[11]   Recknagel, F. (2001) Applications of Machine Learning to Ecological Modelling. *Ecological Modelling*, **146**, 303-310. http://dx.doi.org/10.1016/S0304-3800(01)00316-7

[12]   Olden, J.D., Lawler, J.J. and Poff, N.L. (2008) Machine Learning Methods without Tears: A Primer for Ecologists. *The Quarterly Review of Biology*, **83**, 171-193. http://dx.doi.org/10.1086/587826

[13]   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. http://dx.doi.org/10.1002/eco.46

[14]   Park, Y.-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. http://dx.doi.org/10.1016/S0304-3800(02)00258-2

[15]   Termansen, M., McClean, C.J. and Preston, C.D. (2006) The Use of Genetic Algorithms and Bayesian Classification to Model Species Distributions. *Ecological Modelling*, **192**, 410-424. http://dx.doi.org/10.1016/j.ecolmodel.2005.07.009

[16]   Holzkämper, 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. http://dx.doi.org/10.1016/j.ecolmodel.2006.05.001

[17]   Garzón, M.B., de Dios, R.S. and Ollero, H.S. (2008) Effects of Climate Change on the Distribution of Iberian Tree Species. *Applied Vegetation Science*, **11**, 169-178. http://dx.doi.org/10.3170/2008-7-18348

[18]   Prasad, A.M., Iverson, L.R. and Liaw, A. (2006) Newer Classification and Regression Tree Techniques: Bagging and Random Forests for Ecological Prediction. *Ecosystems*, **9**, 181-199. http://dx.doi.org/10.1007/s10021-005-0054-1

[19]   Cutler, D.R., Edwards Jr., T.C., Beard, K.H. Cutler, A., Hess, K.T., Gibson, J. and Lawler, J.J. (2007) Random Forests for Classification in Ecology. *Ecology*, **88**, 2783-2792. http://dx.doi.org/10.1890/07-0539.1

[20]   Croituru, L. and Liagre, L. (2013) Contribution des Forets à une Economie Verte dans la Région MENA. *Forêt Méditerranéenne*, **34**, 283-291.

[21]   Heusch, B. and Billaux, P. (1966) Carte pédologique du Gharb, de la Maâmora Septentrionale et de leur Bordure Orientale. *Congrès de Pédologie Méditérranèerme*, INRA, Rabat-Maroc.

[22]   Bagaram, M. (2014) Elaboration d'une Base de Données Géographiques et d'un Catalogue des Stations de la Subéraie

de la Maâmora. *Mémoire de 3ème cycle*, ENFI, Salé-Maroc.

[23] Emberger, L. (1955) Une classification Biogéographique des Climats. *Recueil des Travaux des Laboratoires de Botanique*, *Géologie et Zoologie de la Faculté des Sciences de L'Université de Montpellier*, *Série Botanique*, **7**, 3-43.

[24] Hijmans, R.J., Cameron, S.E., Parra, J.L., Jones, P.G. and Jarvis, A. (2005) Very High Resolution Interpolated Climate Surfaces for Global Land Areas. *International Journal of Climatology*, **25**, 1965-1978. http://dx.doi.org/10.1002/joc.1276

[25] R Development Core Team (2014) R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna.

[26] Liaw, A. and Wiener, M. (2002) Classification and Regression by Random Forest. *R News*, **2**, 18-22.

[27] Breiman, L. (2001) Random Forests. *Machine Learning*, **45**, 5-32. http://dx.doi.org/10.1023/A:1010933404324

[28] Landis, J.R. and Koch, G.G. (1977) The Measurement of Observer Agreement for Categorical Data. *Biometrics*, **33**, 159-174. http://dx.doi.org/10.2307/2529310