Ecological Informatics 9 (2012) 37–46

Contents lists available at SciVerse ScienceDirect

Ecological Informatics

journal homepage: www.elsevier.com/locate/ecolinf

Support vector machines to map rare and endangered native plants in Pacific islands forests

Robin Pouteau [a,d,*], Jean-Yves Meyer [c,d], Ravahere Taputuarai [b,d], Benoît Stoll [a]

[a] South Pacific Geosciences Laboratory, University of French Polynesia, Tahiti, French Polynesia
[b] Délégation à la Recherche, Government of French Polynesia, Tahiti, French Polynesia
[c] MaNature, Tahiti, French Polynesia
[d] Moorea Biocode Project, Moorea, French Polynesia

ARTICLE INFO

Article history:
Received 18 October 2011
Received in revised form 2 February 2012
Accepted 2 March 2012
Available online 10 March 2012

Keywords:
Rare species
Support vector machines (SVM)
Random forests (RF)
Biodiversity
Digital elevation model (DEM)
Vegetation mapping

ABSTRACT

It is critical to know accurately the ecological and geographic range of rare and endangered species for biodiversity conservation and management. In this study, we used support vector machines (SVM) for modeling rare species distribution and we compared it to another emerging machine learning classifier called random forests (RF). The comparison was performed using three native and endemic plants found at low- to mid-elevation in the island of Moorea (French Polynesia, South Pacific) and considered rare because of scarce occurrence records: *Lepinia taitensis* (28 observed occurrences), *Pouteria tahitensis* (20 occurrences) and *Santalum insulare* var. *raiateense* (81 occurrences). We selected a set of biophysical variables to describe plant habitats in tropical high volcanic islands, including topographic descriptors and an overstory vegetation map. The former were extracted from a digital elevation model (DEM) and the latter is a result of a SVM classification of spectral and textural bands from very high resolution Quickbird satellite imagery. Our results show that SVM slightly but constantly outperforms RF in predicting the distribution of rare species based on the kappa coefficient and the area under the curve (AUC) achieved by both classifiers. The predicted potential habitats of the three rare species are considerably wider than their currently observed distribution ranges. We hypothesize that the causes of this discrepancy are strong anthropogenic disturbances that have impacted low- to mid-elevation forests in the past and present. There is an urgent need to set up conservation strategies for the endangered plants found in these shrinking habitats on the Pacific islands.

1. Introduction

The detailed knowledge of rare species ecological range and geographic distribution is critical for biodiversity conservation and management (Ferrier, 2002; Rushton et al., 2004). Oceanic islands are famous for their unique biota with high endemism, but also their great vulnerability to anthropogenic disturbances (Caujapé-Castells et al. 2010; Loope et al. 1988) causing the decline of species abundance and distribution, leading sometimes to extinction (Whittaker and Fernandez-Palacios, 2007). As a result, a huge number of endangered species are currently found on island ecosystems (IUCN, 2011). Besides their conservation value, rare species may also play a key role for ecosystem functioning (Lyons and Schwartz, 2001; Lyons et al., 2005).

Occurrence records are scarce for rare species resulting in small training sample available for species distribution models (Pearson et al., 2007; Stockwell and Peterson, 2002; Wisz et al., 2008). A recent study of Williams et al. (2009) compared the ability of a range of models to predict distribution of six rare plant species (from 9 to 129 occurrences). These models included generalized linear models, artificial neural networks, the commonly used maximum entropy (Maxent) distribution and a classification and regression tree (CART) model called random forests (RF) (Breiman, 2001), the latter outperforming the former. RF, introduced by Breiman (2001), is an ensemble classifier developed to produce accurate predictions while limiting overfitting of the data. It consists of many decision trees and outputs the class that occurs most frequently in individual trees. Each input vector is used by each tree of the forest. Each tree gives a classification, and we say the tree “votes” for that class. The forest chooses the classification having the most votes over all the trees in the forest. RF has been recently and successfully used for species distribution modeling (Benito Garzon et al., 2008; Cutler et al., 2007; Prasad et al., 2006; Williams et al., 2009). RF is an easy to use classifier since it has only two parameters that the user has to determine. They are the number of trees to be used and the number of variables to be randomly selected from the available set of variables.

Nonetheless, in the field of remotely sensed data classification, a machine learning algorithm called the support vector machines

* Corresponding author at: Laboratoire GePaSud, Université de la Polynésie française, BP 6570, 98702 Faa'a, French Polynesia. Tel.: +689 866438; fax: +689 803842.
E-mail address: r.pouteau@yahoo.fr (R. Pouteau).

1574-9541/$ – see front matter
doi:10.1016/j.ecoinf.2012.03.003

(SVM) (Vapnik, 1998) may be an important technique for modeling rare species distributions. Algorithms used in remotely sensed data classification for classifying object reflectance are substantially the same than those used in species distribution models for classifying environmental layers (Franklin, 1995). Thus, SVM was successfully used for common species distribution modeling in few recent studies (Drake et al., 2006; Guo et al., 2005; Pouteau et al., 2011a).

SVM was originally introduced as a binary classifier (Vapnik, 1998) and is extensively described by Burges (1998), Hsu et al. (2009) and Schölkopf and Smola (2002). In its classical implementation, it uses two classes (e.g. presence/absence) of training samples within a multidimensional feature space to fit an optimal separating hyperplane (in each dimension, vector component is image gray-level). In this way, SVM tries to maximize the margin that is the distance between the closest training samples, or support vectors, and the hyperplane itself.

SVM consists of projecting vectors into a high dimensional feature space by means of a kernel trick then fitting the optimal hyperplane that separates classes using an optimization function. For a generic pattern x, the corresponding estimated label $\hat{y}$ is given by Eq. (1).

$$\hat{y} = \mathrm{sign}[\mathrm{f}(x)] = \mathrm{sign}[\mathrm{sum}(i \text{ from } 1 \text{ to } N) y_i.\alpha_i.\mathrm{K}(x_i, x) + b] \quad (1)$$

wherein N is the number of training points, the label of the ith sample is y_i, b is a bias parameter, $\mathrm{K}(x_i,x)$ is the chosen kernel and α_i denotes the Lagrangian multipliers.

Several kernels are used in the literature. According to Hsu et al. (2009) and supported by many other authors, the Gaussian radial basis function (RBF) has both advantages (i) of being very successful since it works in an infinite dimensional feature space; and (ii) having a single parameter $\gamma > 0$, contrary to the other well working kernels (e.g. polynomial). The equation is Eq. (2).

$$\mathrm{K}(x_i, x) = \exp\left[-\gamma \|x_i - x\|^2\right] \quad (2)$$

Noise in the data can be accounted for by defining a distance tolerating the data scattering, thus relaxing the decision constraint. This regularization parameter is called C.

Only α_i belonging to support vectors s_i has no null value so the classification function is actually Eq. (3).

$$\hat{y} = \mathrm{sign}[\mathrm{f}(x)] = \mathrm{sign}[\mathrm{sum}(i \text{ from } 1 \text{ to} P_s) y_i.\alpha_i.\mathrm{K}(s_i, x) + b] \quad (3)$$

wherein P_s is the number of support vectors. Thus, the decision boundary is solely based on few meaningful pixels. This is why SVM may be much appropriated for predicting distribution of species with scarce occurrence records. Nevertheless, to our knowledge, it has never been used for rare species distribution modeling.

The aim of this study is twofold: (i) to determine which model among RF and SVM is the most relevant to map rare species in a study case focusing on endangered native and endemic plants on Pacific islands; and (ii) comparing their predicted potential habitat with their current observed range, to understand the causes of their rarity and endangerment.

2. Material and methods

2.1. Target rare and endangered species

The present study was conducted on the oceanic tropical island of Moorea (Society archipelago, French Polynesia), located at 17°33′ South and 149°50′ West in the South Pacific Ocean. It is a small (ca. 140 km^2) and young volcanic island (1.5–2.5 million years old) with a rough topography and the highest summit reaching 1207 m elevation.

This work was part of the “Moorea Biocode Project”, an international research program seeking to collect DNA sequence, distribution, morphological and ecological data of all non-microbial terrestrial and marine life in an island ecosystem (http://www.mooreabiocode.org/).

Three target species were selected based on their rarity and endangerment on Moorea according to their IUCN (International Union for Conservation of Nature) conservation status (IUCN, 2011) (Table 1).

We compiled the available data on the location and abundance of the target species. The term “occurrence” used hereinafter refers to a 5 m × 5 m area where an isolated individual or a population of individuals is present. It means that if two or more geographically close individuals located by GPS (Global Positioning System) occur in the same 5 m × 5 m pixel of a geo-referenced image, this pixel is considered as a single occurrence but if two or more geographically close individuals occur in two adjacent 5 m × 5 m pixels, both pixels are considered as occurrences (Fig. 1).

Lepinia taitensis (Apocynaceae) is a small tree commonly 2–5 m that grows up to 10 m in height on Moorea (pers. obs., Fig. 2.a). An endemic to the islands of Tahiti and Moorea (Society archipelago), it occurs in low- to mid-elevation wet valley forests. It is listed as “critically endangered” (CR) by the IUCN Red List of Threatened Species (IUCN, 2011). We recorded a total of 28 occurrences on Moorea.

Pouteria (syn. *Planchonella*) *tahitensis* (Sapotaceae) is a large tree, often between 10 and 20 m in height on Moorea (pers. obs., Fig. 2.b). It was previously described as an endemic to the Society (Florence et al., 2007), but is probably native to South Pacific islands (Swenson, U., pers. comm., 2011). It is mainly found on slopes in mid-elevation mesic to wet forests. It is not classified by IUCN (2011) but we considered it rare and endangered on Moorea since only 20 occurrences are recorded on the island for a total of about 50 mature trees (pers. obs.).

Santalum insulare var. *raiateense* (Santalaceae) is a shrub up to 3 m tall, endemic to the islands of Moorea and Raiatea (Society archipelago) where it occurs on low- to mid-elevation dry and mesic ridges and slopes (pers. obs., Fig. 2.c). It is considered “near threatened” (NR) on Moorea and CR on Raiatea (IUCN, 2011). A total of 81 occurrences were recorded on Moorea.

2.2. Biophysical descriptors

Vegetation patterns of the Pacific high volcanic islands depend on (i) abiotic factors such as climate, geology, geomorphology, soil substrate and disturbance regime; and (ii) biotic components such as the floristic region, plant dispersal capacities and ecological plant type and function (Carlquist, 1974; Mueller-Dombois and Fosberg, 1998). RF and SVM were compared on their ability to model the ecological niche of our three target species. To describe these ecological niches, our analysis was based on six fine scale environmental descriptors.

2.2.1. Abiotic descriptors

To map rare plant species, we used the five following topo-climatic proxies. The first proxy is elevation that affects air temperature. Considering an environmental lapse rate of 0.0058 °C/m as observed in Hawaii (Baruch and Goldstein, 1999), there is a shift of 7 °C between sea-level and the highest summit of Moorea (Mt Tohiea, 1207 m). Air temperature is one of the most important factors controlling vegetation zonation and key processes such as evapotranspiration, carbon fixation and decomposition, plant productivity and mortality in mountain ecosystems (Chen et al., 1999; Nagy et al., 2003; Richardson, 2004). Slope steepness (called “slope” hereafter) can be considered as a proxy of overland and subsurface flow velocity and runoff rate, effect of micro-topography on precipitation, geomorphology, soil water content (Wilson and Gallant, 2000), mechanical effect on plant rooting and seed dispersion. Slope exposure (called “aspect” hereafter) was used as a proxy of solar insolation and evapotranspiration (Wilson and Gallant, 2000). Windwardness was used to express exposure to trade wind (see LaRosa et al. (2007) for a