



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.

© 2012 Elsevier B.V. All rights reserved.

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).

(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 multi-dimensional 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).

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

wherein N is the number of training points, the label of the i th sample is y_i , b is a bias parameter, $K(x_i, x)$ 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).

$$K(x_i, x) = \exp[-\gamma \|x_i - x\|^2] \quad (2)$$

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).

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

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 km²) 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

Table 1
Characteristics of the species modeled on Moorea.

Species name	Biogeographical status	IUCN status	# of occurrences	Maximum height	Habitat (elevation range ^a)
<i>Lepinia taitensis</i>	Endemic to Tahiti and Moorea	CR	28	10 m	Wet valleys (193–495 m)
<i>Pouteria</i> (syn. <i>Planchonella</i>) <i>tahitensis</i>	Indigenous	N/A	20	20 m	Mesic to wet slopes (264–567 m)
<i>Santalum insulare</i> var. <i>raiateense</i>	Endemic to the Society archipelago	NT	81	5 m	Dry ridges (205–730 m)

IUCN status (2011): NT = near threatened; CR = critically endangered.

^a Elevation range extracted on the DEM from geo-localized field observations.

use example in Hawaii). The last abiotic descriptor is a compound topographic index (CTI), quantifying fluid drainage by micro-topography and explaining geomorphology (Gessler et al., 2000; Moore et al., 1993), with low CTI values representing convex positions like mountain crests and with high CTI values representing concave positions like coves or hill-slope bases. It is considered a secondary physiographic descriptor since it is computed from primary physiographic descriptors (elevation and slope) (Moore et al., 1991). Indeed, as shown by Eq. (4), CTI is a function of the slope angle β (in radians) and the specific catchment area (As) expressed as m^2 per unit width orthogonal to the flow direction.

$$CTI = \log(As/\tan\beta) \quad (4)$$

Abiotic descriptors were extracted from a 5 m-resolution digital elevation model (DEM) provided by the Urbanism Department of the Government of French Polynesia.

2.2.2. Biotic descriptor

We used a vegetation map as the biotic descriptor. Indeed forest overstory (canopy) can affect temperature, light, water, and nutrient availability (Riegel et al., 1992) which are essential factors explaining the presence or absence of plants, including rare species.

The vegetation map was obtained by classifying satellite imagery. The different vegetation classes are labeled according to the dominant plant species or communities that are visible in remotely sensed data. The use of both physiographic and remotely sensed data in classifications is commonly found in the literature (e.g. Hutchinson, 1982; Joshi et al., 2006; Linderman et al., 2004; Strahler, 1981). Here, satellite data is a mosaic of five 0.60 m-resolution Quickbird scenes acquired in 2006. Very high resolution imagery such as the one from the Quickbird satellite is useful for tree species identification (Turner et al., 2003; Xie et al., 2008). As suggested by Song et al. (2005), spectral and textural information was considered separately since they have a strongly different nature.

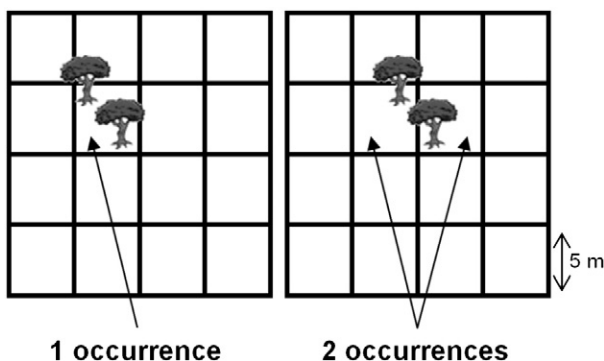


Fig. 1. Spatial sampling design used to define occurrences or “presence” pixels on a reference image.

Spectral information was from the blue (430–545 nm) to the near infrared (715–918 nm). Texture is generally referred to as the detailed spatial pattern of variability of the image average tone. Here, textural information includes eight gray-level co-occurrence matrix (GLCM) metrics introduced by Haralick et al. (1973), namely mean, variance, homogeneity, contrast, dissimilarity, entropy, second moment and correlation. These kinds of texture metrics are the most used and produce good output results (Augusteijn et al., 1995; Franklin and Peddle, 1989; Franklin et al., 2000; Gong et al. 1992; Marceau et al., 1990; Nyongui et al., 2002; Podest and Saatchi, 2002). Chen et al. (2004) and Franklin et al. (1996) showed that the accuracy of classification on texture metrics can be improved by using multiple extraction window sizes. Here, texture was extracted in windows of 3×3 pixels, 9×9 pixels and 15×15 pixels visually identified as representing intra-tree micro-texture (small branch structure), intra-tree macro-texture (large branch structure) and inter-tree texture (trunks, individuals structure) respectively.

Numerous algorithms have been proposed to classify two sources such as spectral bands and textural information. In the multi-source comparative studies found in the literature, “winning” algorithms include the Dempster orthogonal sum combination rule (Lee et al., 1987), artificial neural networks (Benediktsson et al., 1990; Serpico and Roli, 1995), the logarithmic opinion pool (Benediktsson and Kanellopoulos, 1999), the sequential maximum a posteriori (Michelson et al., 2000), the majority voting (Fauvel et al., 2006) and SVM (Chu and Ge, 2010; Huang et al., 2002; Song et al., 2005; Waske and Benediktsson, 2007). The latter was used in the most recent studies and has never been “beaten” in the previously mentioned comparative studies. SVM success in multi-source fusion is probably due to both its generally recognized performance in mono-source and its ability to weight numerous and heterogeneous sources (different types, different units, mixing of continuous and categorical data) according to their relevance.

The classification scheme we used is the one introduced by Waske and Benediktsson (2007): (i) spectral bands and textural information are classified by means of a single SVM applied on each source separately. The output of the classifier is an image containing the distance of each vector to the decision boundary (also called “rule image”) which expresses the probability of each pixel to belong to each class (one rule image is produced per class); and (ii) an additional SVM is applied on the set of rule images to perform the fusion.

Input pixels were assigned following some thirty ground truth missions between 22 July 2009 and 6 February 2011. Sampled surface represents 0.2 ha for each of the 17 classes identified on Moorea, namely a total of 3.4 ha i.e. 0.025% of the island (distribution of the sample points is shown in Fig. 3). The first half of this area was used for SVM training and the other half was put aside for classification assessment.

The resulting 0.60 m-resolution vegetation map was upsampled to a 5 m-resolution to correspond with abiotic descriptors using the nearest neighbor method, which is more reliable than other classic methods (e.g. bilinear interpolation and cubic convolution) to resample categorical data (Baboo and Devi, 2010). Fig. 4 summarizes the workflow of this study.

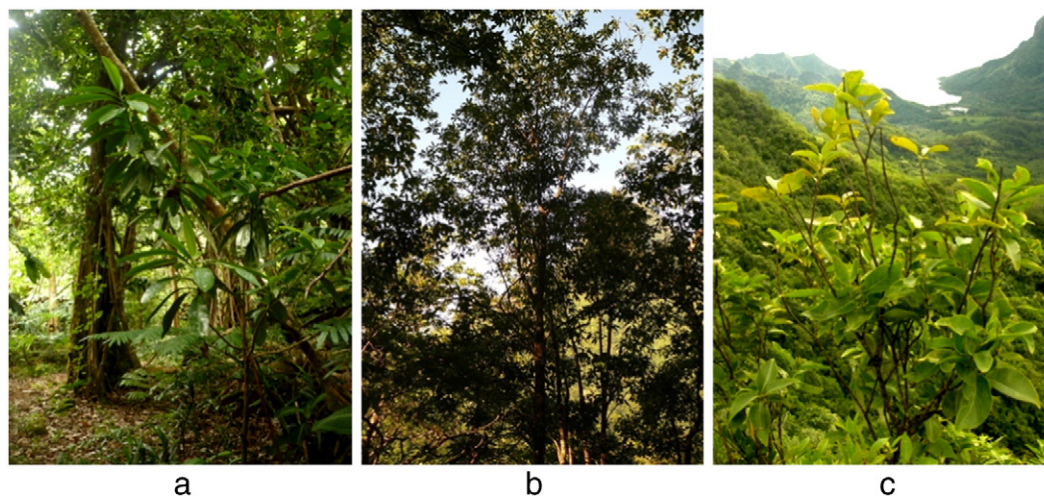


Fig. 2. (a) *Lepinia taitensis* in the understory of lowland wet valley forest at 250 m elevation (R. Pouteau). (b) A large population (20–30 individuals) of *Pouteria tahitensis* at 300 m elev. on slope mesic to wet forest (R. Taputuarai). (c) *Santalum insulare* on a dry and open ridge at 400 m elev. (J.-Y. Meyer).

2.3. Classifier comparison methodology

2.3.1. Implementation and training

The ability of RF and SVM to integrate the aforementioned biophysical descriptors was compared. We used the implementations found in the open source machine learning software from the University of Waikato (New Zealand) called Weka (<http://www.cs.waikato.ac.nz/ml/weka/>). Rare species classifications were trained on 67% of occurrences and evaluated on the remaining 33%. “Absence” set of pixels are actually made of “pseudo-absence” pixels i.e. pixels randomly sampled within the unoccupied space (function “Generate random sample using ground truth ROIs” of the ENVI software). The same number of pixels was used for the “presence” class than for the “absence”

class in order to avoid under- or over-estimation problems due to unbalanced training sets (Eitrich and Lang, 2005; Eitrich and Lang, 2006; Japkowicz and Stephen, 2002).

2.3.2. Optimal parameter determination

Classifier parameters, namely the number of trees and the number of variables to be randomly selected from the available set of variables for the RF and the regularization parameter C and the RBF kernel parameter γ for the SVM, were selected according to the cross-validation method (Hsu et al., 2009). The goal is to identify optimal parameters so that the classifier can accurately predict unknown data. This method is easy to use, quite fast and can be more reliable

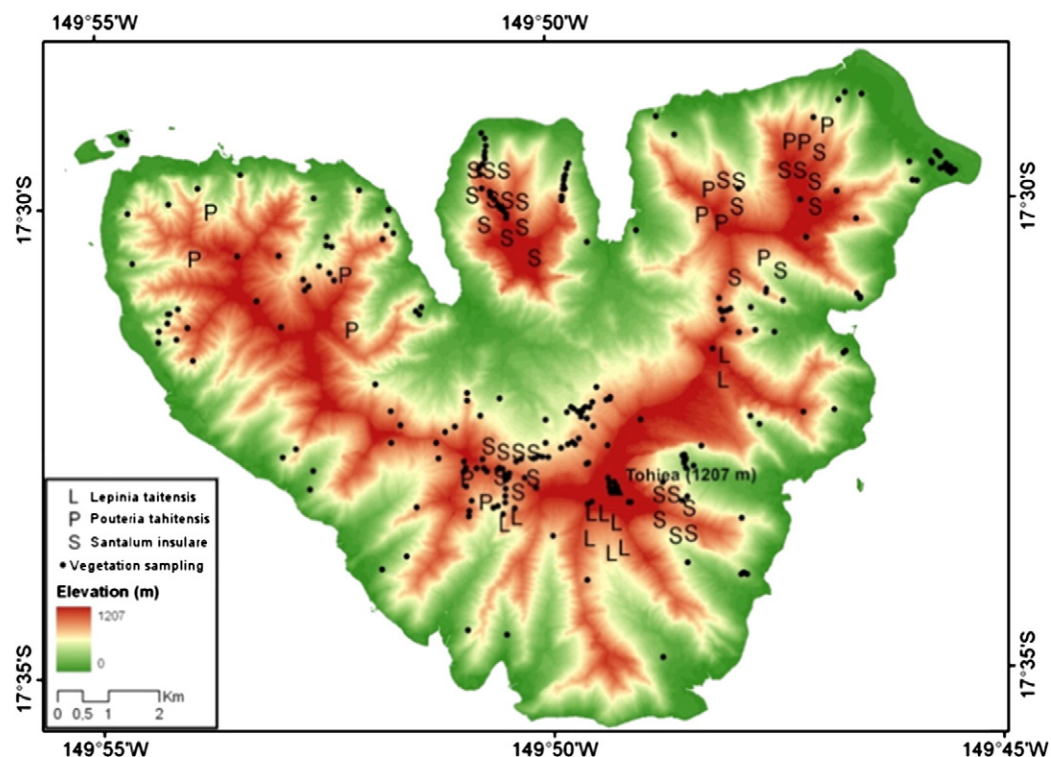


Fig. 3. Distribution of the GPS points sampled for vegetation mapping and location of rare species on Moorea. Each letter (L, P or S) may refer to one or several occurrences.

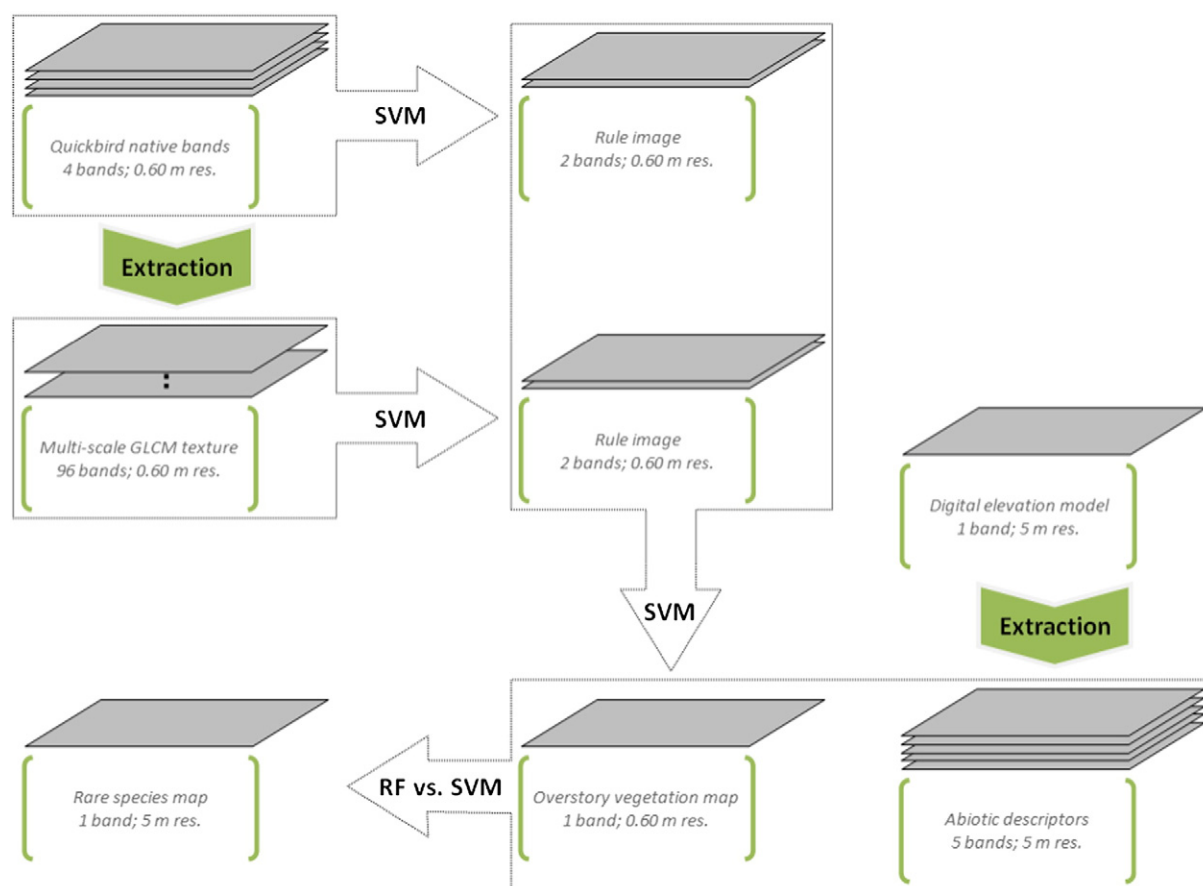


Fig. 4. Workflow of this study.

than more advanced iterative techniques that do not always consider parameters as independent.

The principle consists in partitioning the available set of pixels in n subsets. Then the classifier is trained on $n-1$ subsets and we assess accuracy with the remaining subset. This method can prevent the overfitting problem. Finally, this operation is repeated for each pair of parameter that we make vary exponentially (e.g. 1.10^1 , 1.10^2 , 1.10^3 , ...), then linearly (e.g. 10, 11, 12, ...).

2.3.3. Accuracy assessment

Both machine learning algorithms were compared on the basis of the accuracy they produce when trained on the same set of training pixels where the same biophysical descriptors were extracted. In Congalton and Green (2009), several methods for assessing classification accuracy are introduced. In our study case, we used the same two metrics as in Williams et al. (2009).

On the one hand, AUC score refers to the area under the ROC (Receiver Operating Characteristic) curve: ROC metric is often represented by a curve corresponding to corrected assigned pixels rate according to the misclassification rate. The best possible prediction method would yield a point in the upper left corner or coordinate (0,1) of the ROC space, representing 100% correctly assigned pixels. On the other hand, Cohen's kappa expresses whether correctly assigned pixels may have been assigned by chance or not based on the classification decision rule. A value of 1 indicates perfect agreement and 0.5 indicates a pattern arising by chance.

Based on the accuracy metrics obtained for each classifier, a statistical analysis can be performed to test if the difference is significantly equal or different from zero. We compared the AUC of the two ROC curves yielded by the RF and the SVM by using the Delong test (Delong et al.,

1988) implemented in the 'pROC' R package (Robin et al., 2011). The Cohen's kappa values were compared with a Z-test. In both cases, the null hypothesis states that there is no significant difference between accuracy metrics yielded by the RF and the SVM. The null hypothesis was rejected or failed to be rejected at a 5% significance level.

3. Results

3.1. Vegetation map

The SVM classification of the Quickbird imagery (Fig. 5) gives fairly good results with a kappa of 0.842 and an AUC of 0.965. Texture is arguably the most contributing information since the classification based on the single textural information (without spectral bands) gives a kappa of 0.821 and an AUC of 0.955 (data not shown).

3.2. Contribution of biophysical descriptors

Calculation of the descriptors relative contribution presented in Fig. 6 was based on the difference of AUC (Δ AUC) and the difference of kappa (Δ kappa) yielded with all descriptors and without the regarded descriptor. In our case, the relative contribution is always positive which confirms their adaptation to the context of Pacific high volcanic islands and to our target species but varies among target species. Although AUC and Cohen's kappa scores are not closely correlated, the three most important contributing descriptors are (by order of contribution): elevation, CTI and slope. Elevation is arguably the most contributing descriptor for the three target species with a Δ AUC of 4.2% for *L. taitensis*, 19.8% for *P. tahitensis* and 8.9% for *S. insulare* and a Δ kappa of 19.5% for *L. taitensis*, 27.8% for *P. tahitensis* and 25.6% for

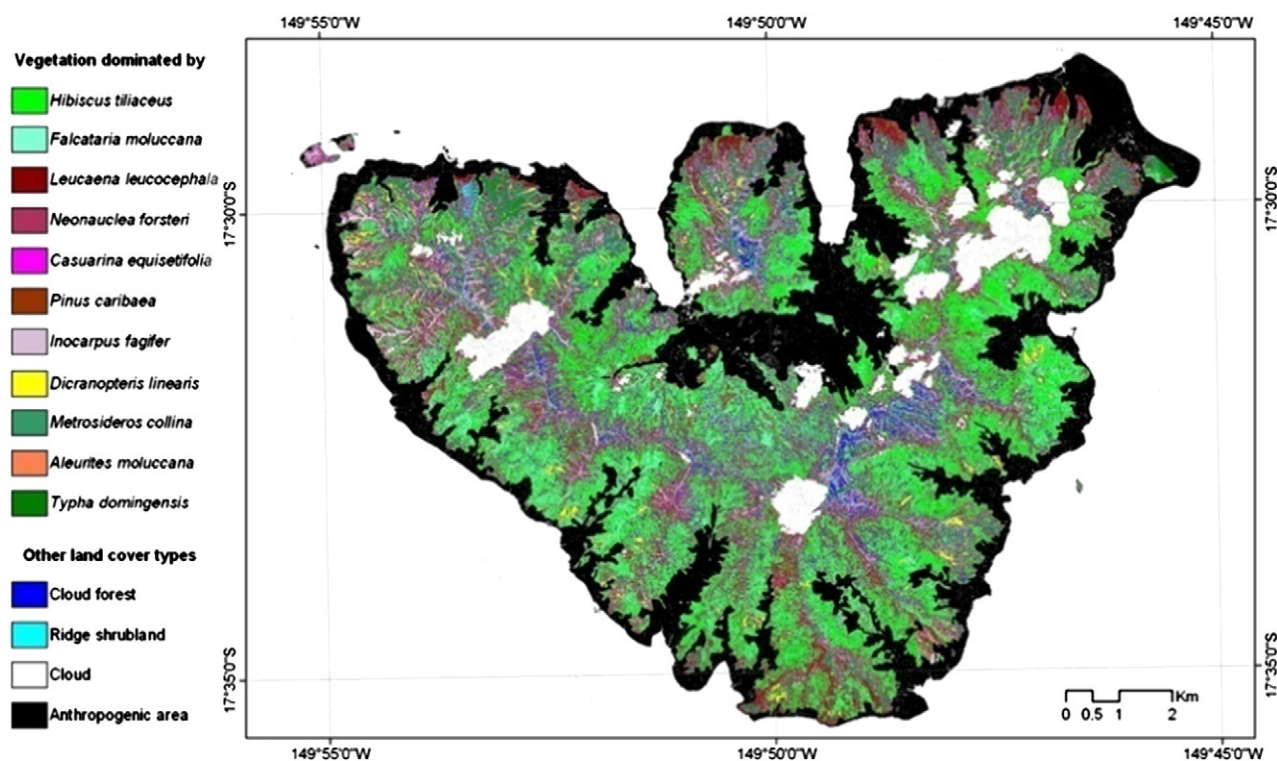


Fig. 5. Vegetation map of the Moorea island resulting from a SVM classification of a mosaic of five Quickbird satellite images (2006).

S. insulare. CTI is the second most contributing descriptor for *L. taitensis* ($\Delta \text{AUC} = 4.2\%$ and $\Delta \text{kappa} = 7.8\%$) and for *P. tahitensis* ($\Delta \text{AUC} = 7.7\%$ and $\Delta \text{kappa} = 13.0\%$) but only the fourth most contributing descriptor for *S. insulare* ($\Delta \text{AUC} = 3.0\%$ and $\Delta \text{kappa} = 7.7\%$). Slope is the third most contributing descriptor for the three target species with a ΔAUC of 1.6% for *L. taitensis*, 5.5% for *P. tahitensis* and 2.7% for *S. insulare* and

a Δkappa of 0.8% for *L. taitensis*, 9.9% for *P. tahitensis* and 16.0% for *S. insulare*.

Nevertheless, aspect is also substantially influencing spatial distribution of *S. insulare* ($\Delta \text{AUC} = 1.1\%$ and $\Delta \text{kappa} = 17.9\%$). The same applies to overstory vegetation that conspicuously contribute to the spatial distribution of *L. taitensis* ($\Delta \text{AUC} = 3.2\%$ and $\Delta \text{kappa} = 0.7\%$).

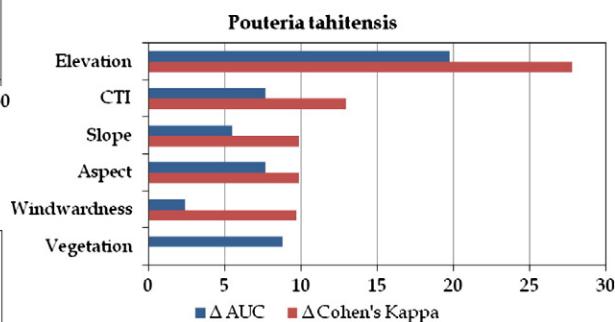
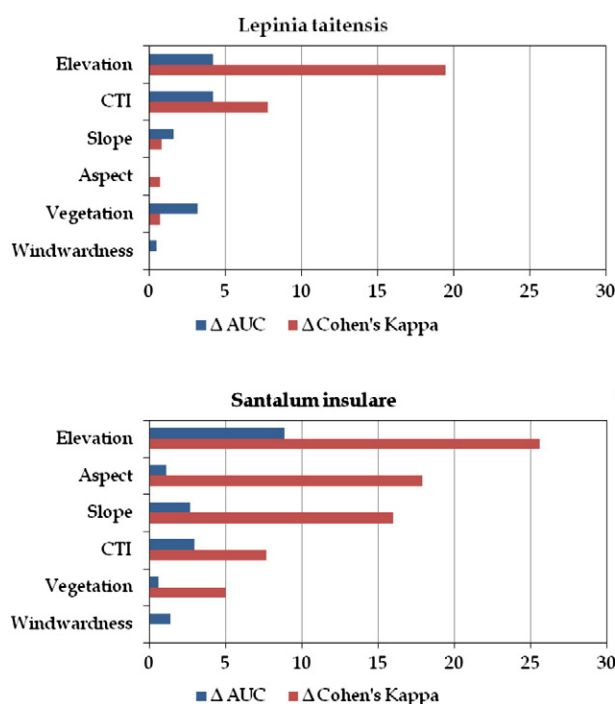


Fig. 6. Relative contribution of each biophysical descriptor in the SVM classification. Relative contribution is calculated as the difference of accuracy (%) yielded with all descriptors and without the regarded descriptor.

3.3. Classifier outputs

Spatial patterns predicted by both machine learning algorithms are visually quite similar but differ greatly in surface area (Fig. 7). The potential distribution ranges of *L. taitensis*, *P. tahitensis* and *S. insulare* are indeed 28 ha, 42 ha and 39 ha respectively according to the RF model and 15 ha, 38 ha and 32 ha respectively according to the SVM model. The latter model also seems more sensible to micro-topography than the former which outputs shapes with sharp contours. In both cases, there is a huge difference between predicted potential habitat and current habitat based on observed occurrences which are 700 m² (4.7‰ of the potential habitat modeled by SVM), 500 m² (1.3‰) and 2025 m² (6.3‰) respectively.

As shown by comparative results presented in Table 2, both classifiers yield very good numerical results with $87.4\% \leq \text{AUC} \leq 97.9\%$ and $70.0\% \leq \text{kappa} \leq 87.0\%$. Both classifiers yield the same kappa for *L. taitensis* (85.4%) and the same AUC for *S. insulare* (97.2%) but the AUC of *L. taitensis* classification is significantly higher when performed by SVM (97.9%) than when performed by RF (97.4%) and the kappa of *S. insulare* classification is significantly higher when performed by SVM (87.0%) than when performed by RF (84.5%). In the case of *P. tahitensis*, the target species with the most scarce occurrence records (only 20), SVM significantly outperforms RF regarding both AUC (89.0% and 87.4% respectively) and kappa (78.0% and 70.0% respectively).

4. Discussion

4.1. Random forests vs. support vector machines

RF and SVM were compared on their ability to predict rare and endangered species distributions. RF was found to be optimal for predicting rare species occurrences among a wide panel of algorithms in Williams et al. (2009). To our knowledge, SVM has never been used for predicting rare species distribution. However, it generally outperforms RF in our study case, especially when the number of occurrence is small. The main reason is most likely the result of the paradigm of SVM based on a small pixel sample (i.e. support vectors) (Foody and Mathur, 2006). Consequently, SVM is able to be trained with few meaningful pixels and to fit limited information. The results presented in this study suggest that, in our context, SVM is able to

Table 2

Comparison of RF and SVM to predict occurrences for rare plant species.

	RF(%)	SVM (%)
(a) AUC scores		
<i>Lepinia taitensis</i>	97.4 ^a	97.9 ^a
<i>Pouteria tahitensis</i>	87.4 ^a	89.0 ^a
<i>Santalum insulare</i> var. <i>raiateense</i>	97.2	97.2
(b) Cohen's kappa		
<i>Lepinia taitensis</i>	85.4	85.4
<i>Pouteria tahitensis</i>	70.0 ^a	78.0 ^a
<i>Santalum insulare</i> var. <i>raiateense</i>	84.5 ^a	87.0 ^a

^a The difference between classifiers (RF and SVM) is not statistically significantly equal to zero at a 5% threshold (Delong test for AUC scores and Z-test for Cohen's kappa).

predict rare species distribution with good accuracy from only 13 training pixels (67% of the 20 recorded occurrences of *P. tahitensis*) of each class ("presence" and "absence"). In addition, in a similar context, SVM should be used rather than RF when the number of available training pixels range from 13 to 54 (67% of the 81 recorded occurrences of *S. insulare*).

Reciprocally, SVM may not be very impacted by insignificant pixels. Noisy pixels are more frequent in high resolution imagery than in coarse resolution imagery due to information aggregation in large pixels (Hatton et al., 1997; Turner et al., 2003). SVM is thus probably more adapted than CART approaches such as RF, typically showing difficulties at a fine scale (Thuiller et al., 2003).

Moreover, RF classification process occurs in a six dimension space (one dimension per biophysical descriptor). By using the RBF kernel which works in an infinite dimensional feature space, it is easier for SVM to separate potential habitat from inappropriate habitat.

In this study, classifiers were trained on "presence" and "pseudo-absence" pixels but regarding the wide potential habitat predicted by both classifiers, it is probable that some pixels within the unoccupied potential habitat has been randomly sampled by this technique and labeled as "absence" pixels. This is a classical limitation of presence/absence models compared to presence-only models such as Maxent and the genetic algorithm for rule-set prediction (GARP). However, SVM has the advantage that presence-only extensions exist such as the one-class SVM, the biased SVM or the more recent positive and unlabeled learning (PUL) algorithm which requires only a small set

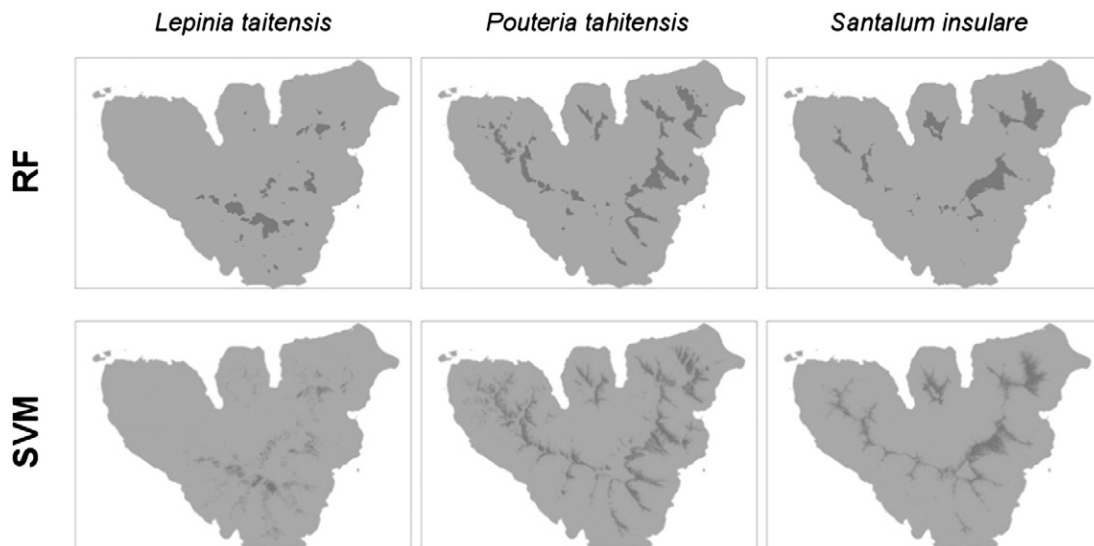


Fig. 7. Comparison of the distribution maps of three rare species produced by RF and SVM classification. Light gray color refers to area where the species is predicted to be absent and dark gray to its potential habitat.

of data to be labeled (Li et al., 2011). Consequently, further works should test them.

4.2. Potential habitat vs. current habitat

Our previous analysis showed that the potential habitats predicted by both classifiers were excessively wider than the observed current habitat of our target species.

First, we assume that this low density is not related to an overestimation of the potential habitats made by the models as they give fairly good analytical results to discriminate occurrence records and randomly sampled areas. Results provided by SVM are also very consistent with our field observations, including the prominent role of elevation, slope, humidity and micro-topography on plant distribution in island forests; the strong influence of slope aspect on the distribution of *S. insulare* which essentially occurs on dry ridges and slopes facing north; and the importance of the overstorey vegetation for *L. taitensis*, a small tree found in the understorey of dense rainforest with a probable strong dependence of incident light transmitted by the canopy.

Secondly, the assumption that the native and endemic flora of Moorea is currently poorly known can be excluded as the island has been surveyed by naturalists and botanists in the past (Florence et al., 2007; Grant et al. 1974) and in a more extensive way for the last 5 years during the “Moorea Biocode Project” (Taputuarai and Meyer, unpub. data).

It is more likely that the discrepancy between potential habitat surface area and current habitat surface area is due to anthropogenic disturbances which are particularly severe at low and mid-elevation on Moorea: (i) habitat loss and fragmentation caused by deforestation, intentional and accidental fires, and grazing by feral ungulates (mainly goats and pigs) since the pre-European period (ca. 1000 years ago) and in the modern times; (ii) invasions by alien plant species with more than 180 alien species currently naturalized on the island (Fourdrigniez, M., pers. comm., 2011). *L. taitensis* and to a lesser extent *P. tahitensis* are threatened by the small tree *Miconia calvenscens* which overtops rainforest native flora (Meyer, 2004) and covers about 25% of the island (Pouteau et al., 2011b). Native plant species are also threatened by other major invasive plants on Moorea which are listed in Table 3; (iii) proliferation of introduced predatory animals. Our three target plant species are severely depredated by rats (*Rattus* spp.) feeding on their large seeds (Lhuillier et al., 2006; Meyer and Butaud, 2009); and (iv) extirpation or extinction of native and endemic avian frugivores by over-hunting, predation (rats, cats, Swamp Harrier *Circus approximans*) and competition with alien birds (Red-vented Bulbul *Pycnonotus cafer*, Common Myna *Acridotheres tristis*) which lead to a decrease or lack of seed dispersal (Spotswood and Meyer, 2009).

Table 3
List of invasive plants among the most widely spread on Moorea.

Family	Genus	Species	Habit
Asteraceae	<i>Mikania</i>	<i>micrantha</i>	Vine
Bignoniaceae	<i>Spathodea</i>	<i>campanulata</i>	Tree
–	<i>Tecoma</i>	<i>stans</i>	Tree
Convolvulaceae	<i>Merremia</i>	<i>peltata</i>	Vine
Melastomataceae	<i>Miconia</i>	<i>calvenscens</i>	Tree
Mimosaceae	<i>Falcataria</i>	<i>moluccana</i>	Tree
Myrtaceae	<i>Psidium</i>	<i>cattleianum</i>	Tree
–	<i>Psidium</i>	<i>guajava</i>	Tree
–	<i>Syzygium</i>	<i>cumini</i>	Tree
Poaceae	<i>Melinis</i>	<i>minutiflora</i>	Grass
–	<i>Miscanthus</i>	<i>floridulus</i>	Grass
Rosaceae	<i>Rubus</i>	<i>rosifolius</i>	Shrub
Verbenaceae	<i>Lantana</i>	<i>camara</i>	Shrub

Moreover, the three target species can be considered as *K*-strategists, characterized by a large individual size, slow life cycle and the production of a few number of large fruits (MacArthur and Wilson, 1967). Their life history traits made them thus more vulnerable to rapid environmental changes.

In a future work, it will be interesting to repeat this research for other rare plant species including those occurring at higher elevation and along the coastline on other French Polynesian and Pacific islands. This would enable one to better decide conservation status by comparing plant species potential habitat with their current habitat, and more generally speaking to better understand forest dynamics at the landscape scale in a wider range of insular ecosystems.

5. Conclusion

We compared two ecological niche models, random forests (RF) and support vector machines (SVM), in order to predict the distribution of rare species in island forest ecosystems. Our analysis focused on three endangered native and endemic plants on the tropical oceanic island of Moorea (French Polynesia) with small occurrence records. It was based on six fine scale environmental descriptors, namely elevation, slope steepness, slope aspect, windwardness, a compound topographic index (CTI) quantifying fluid drainage, and a vegetation map obtained by classifying a set of Quickbird satellite scenes. Results revealed that SVM significantly outperforms RF especially when the number of observed occurrences is scarce. By producing more accurate maps of rare and endangered species, SVM is thus a tool of great practical value for conservation macroecologists and resource managers, and should be more considered in future research.

The high accuracy of distribution maps provided by SVM also allowed us to better understand the causes of the rarity and/or endangerment of some island native and endemic species by comparing their predicted potential habitat with their current observed range. Our data showed that the three target species have a wide geographic distribution but small population size probably caused by strong anthropogenic disturbances aggravated by their “outmoded” life history traits, i.e. not adapted to rapidly changing ecosystems. By contrast, other rare species found on the island of Moorea in relatively well-preserved habitats (e.g. the subalpine vegetation or montane cloud forests) have a narrower geographic distribution but a relatively higher population density. We assume that the rare species found in the low- and mid-elevation forests of the Pacific islands are therefore much more prone to extinction and should be of high conservation priority.

Acknowledgments

The authors are grateful to Jean-François Butaud for sharing his GPS points of the target plants, Marie Fourdrigniez for her help during field surveys, the Service de l'Urbanisme of the Government of French Polynesia for providing the DEM, the Délégation à la Recherche of the Government of French Polynesia and the “Moorea Biocode Project” for financial support. We deeply thank Thomas W. Gillespie (Department of Geography, University of California, Los Angeles) for revising the English on an early draft of this paper, and the anonymous reviewers for their relevant comments.

References

- Augusteijn, M.F., Clemens, L.E., Shaw, K.A., 1995. Performance evaluation of texture measures for ground cover identification in satellite images by means of a neural network classifier. *IEEE Transactions on Geoscience and Remote Sensing* 33, 616–625.
- Baboo, S.S., Devi, R., 2010. An analysis of different resampling methods in Coimbatore, District. *Global Journal of Computer Science and Technology* 10, 61–66.
- Baruch, Z., Goldstein, G., 1999. Leaf construction cost, nutrient concentration, and net CO₂ assimilation of native and invasive species in Hawaii. *Oecologia* 121, 183–192.

- Benediktsson, J.A., Kanellopoulos, I., 1999. Classification of multisource and hyperspectral data based on decision fusion. *IEEE Transactions on Geoscience and Remote Sensing* 37, 1367–1377.
- Benediktsson, J.A., Swain, P.H., Ersoy, O.K., 1990. Neural network approaches versus statistical methods in classification of multisource remote sensing data. *IEEE Transactions on Geoscience and Remote Sensing* 28, 540–552.
- Benito Garzon, M., Sanchez De Dios, R., Sainz Ollero, H., 2008. Effects of climate change on the distribution of Iberian tree species. *Applied Vegetation Science* 11, 169–178.
- Breiman, L., 2001. Random forests. *Machine Learning* 45, 5–32.
- Burges, C.J.C., 1998. A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery* 2, 121–167.
- Carlquist, S., 1974. *Island Biology*. Columbia University Press, New York.
- Caujapé-Castells, J., Tye, A., Crawford, D.J., Santos-Guerra, A., Sakai, A., Beaver, K., Lobin, W., Florens, F.B.V., Moura, M., Jardim, R., Gómes, I., Kueffer, C., 2010. Conservation of oceanic island floras: present and future global challenges. *Perspectives in Plant Ecology, Evolution and Systematics* 12, 107–129.
- Chen, J., Saunders, S.C., Crow, T.R., Naiman, R.J., Brososke, K.D., Mroz, G.D., Brookshire, B.L., Franklin, J.F., 1999. Microclimate in forest ecosystem and landscape ecology. *Bioscience* 49, 288–297.
- Chen, D., Stow, D.A., Gong, P., 2004. Examining the effect of spatial resolution and texture window size on classification accuracy: an urban environment case. *International Journal of Remote Sensing* 25, 2177–2192.
- Chu, H.T., Ge, L., 2010. Synergistic use of multi-temporal ALOS/PALSAR with SPOT multispectral satellite imagery for land cover mapping in the Ho Chi Minh city area, Vietnam. *Proc. IEEE Geoscience and Remote Sensing Symposium*, Hawaii.
- Congalton, R.G., Green, K., 2009. *Assessing the Accuracy of Remotely Sensed Data: Principles and Practices*, second ed. CRC Press, Boca Raton.
- Cutler, D.R., Edwards, T.C., Beard, K.H., Cutler, A., Hess, K.T., 2007. Random forests for classification in ecology. *Ecology* 88, 2783–2792.
- Delong, E.R., Delong, D.M., Clarke-Pearson, D.L., 1988. Comparing the areas under two or more correlated receiver operating characteristic curve: a nonparametric approach. *Biometrics* 44, 837–845.
- Drake, J.M., Randin, C., Guisan, A., 2006. Modelling ecological niches with support vector machines. *Journal of Applied Ecology* 43, 424–432.
- Eitrich, T., Lang, B., 2005. Parallel tuning of support vector machine learning parameters for large and unbalanced data sets. In: Berthold, M.R., Glen, R., Diederichs, K., Kohlbacher, O., Fischer, I. (Eds.), *Computational Life Sciences*. Lecture Notes in Computer Science, 3695. Springer, Konstanz, pp. 253–264.
- Eitrich, T., Lang, B., 2006. Efficient optimization of support vector machine learning parameters for unbalanced datasets. *Journal of Computational and Applied Mathematics* 196, 425–436.
- Fauvel, M., Chanussot, J., Benediktsson, J.A., 2006. A combined support vector machines classification based on decision fusion. *Proc. IEEE Geoscience and Remote Sensing Symposium*, Denver, pp. 2494–2497.
- Ferrier, S., 2002. Mapping spatial pattern in biodiversity for regional conservation planning: where to from here? *Systematic Biology* 51, 331–363.
- Florence, J., Chevillotte, H., Ollier, C., Meyer, J.-Y., 2007. Base de données botaniques Nadeaud de l'Herbier de la Polynésie française. <http://www.herbier-tahiti.pf2007>.
- Foody, G.M., Mathur, A., 2006. The use of small training sets containing mixed pixels for accurate hard image classification: training on mixed spectral responses for classification by a SVM. *Remote Sensing of Environment* 103, 179–189.
- Franklin, J., 1995. Predictive vegetation mapping: geographic modeling of biospatial patterns in relation to environmental gradients. *Progress in Physical Geography* 19, 474–499.
- Franklin, S.E., Peddle, D.R., 1989. Spectral texture for improved class discrimination in complex terrain. *International Journal of Remote Sensing* 10, 1437–1443.
- Franklin, S.E., Wulder, M.A., Lavigne, M.B., 1996. Automated derivation of geographic window sizes for remote sensing digital image texture analysis. *Computers and Geosciences* 22, 665–673.
- Franklin, S.E., Hall, R.J., Moskal, L.M., Maudie, A.J., Lavigne, M.B., 2000. Incorporating texture into classification of forest species composition from airborne multispectral images. *International Journal of Remote Sensing* 21, 61–79.
- Gessler, P.E., Chadwick, O.A., Chamran, F., Althouse, L., Holmes, K., 2000. Modeling soil-landscape and ecosystem properties using terrain attributes. *Soil Science Society of America Journal* 64, 2046–2056.
- Gong, P., Marceau, D.J., Howarth, P.J., 1992. A comparison of spatial feature extraction algorithms for land-use classification with SPOT HRV data. *Remote Sensing of Environment* 40, 137–151.
- Grant, M.L., Fosberg, F.R., Smith, H.M., 1974. Partial flora of the Society Islands: Ericaceae to Apocynaceae. *Smithsonian Contributions to Botany*, No. 17. Smithsonian Institution Press, Washington.
- Guo, Q., Kelly, M., Graham, C.H., 2005. Support vector machines for predicting distribution of Sudden Oak Death in California. *Ecological Modelling* 182, 75–90.
- Haralick, R.M., Shanmugam, K., Dinstein, I., 1973. Textural features for image classification. *IEEE Transactions on Systems, Man, and Cybernetics* 3, 610–620.
- Hatton, T.J., Salvucci, G.D., Wu, H.L., 1997. Eagleson's optimality theory of an ecohydrological equilibrium: quo vadis? *Functional Ecology* 11, 665–674.
- Hsu, C.W., Chang, C.C., Lin, C.J., 2009. A practical guide to support vector classification. Technical Note. Department of Computer Science and Information Engineering, National Taiwan Univ., Taiwan.
- Huang, C., Davis, L.S., Townshend, J.R.G., 2002. An assessment of support vector machines for land cover classification. *International Journal of Remote Sensing* 23, 725–749.
- Hutchinson, C.F., 1982. Techniques for combining Landsat and ancillary data for digital classification improvement. *Photogrammetric Engineering and Remote Sensing* 48, 123–130.
- IUCN, 2011. IUCN Red List of Threatened Species, Version 2011.2. <http://www.iucnredlist.org> 2011.
- Japkowicz, N., Stephen, S., 2002. The class imbalance problem: a systematic study. *Intelligent Data Analysis* 6, 429–450.
- Joshi, C., De Leeuw, J., van Andel, J., Skidmore, K.A., Lekhak, H.D., van Duren, I.C., Norbu, N., 2006. Indirect remote sensing of a cryptic forest understorey invasive species. *Forest Ecology and Management* 225, 245–256.
- LaRosa, A.M., Purrell, M., Franklin, J., Denslow, J., 2007. Designing a control strategy for *Miconia calvescens* in Hawaii using spatial modelling. 9th International Conference on the Ecology and Management of Alien Plant Invasions (poster), Perth.
- Lee, T., Richards, J.A., Swain, P.H., 1987. Probabilistic and evidential approaches for multisource data analysis. *IEEE Transactions on Geoscience and Remote Sensing* 25, 283–293.
- Lhuillier, E., Butaud, J.-F., Bouvet, J.-M., 2006. Extensive clonality and strong differentiation in the insular Pacific tree *Santalum insulare*: implications for its conservation. *Annals of Botany* 98, 1061–1072.
- Li, W., Guo, Q., Elkan, C., 2011. A positive and unlabeled learning algorithm for one-class classification of remote-sensing data. *IEEE Transactions on Geoscience and Remote Sensing* 49, 717–725.
- Linderman, M.A., Liu, J., Qi, J., An, L., Ouyang, Z., Yang, J., Tan, T., 2004. Using artificial neural networks to map the spatial distribution of understorey bamboo from remotely sensed data. *International Journal of Remote Sensing* 25, 1685–1700.
- Loope, L.L., Hamann, O., Stone, C.P., 1988. Comparative conservation biology of oceanic archipelagoes. *BioScience* 38, 272–282.
- Lyons, K.G., Schwartz, M.W., 2001. Rare species loss alters ecosystem function – invasion resistance. *Ecology Letters* 4, 358–365.
- Lyons, K.G., Brigham, C.A., Traut, B.H., Schwartz, M.W., 2005. Rare species and ecosystem functioning. *Conservation Biology* 19, 1019–1024.
- MacArthur, R., Wilson, E.O., 1967. *The Theory of Island Biogeography*. Princeton University Press, Princeton.
- Marceau, D.J., Howarth, P.J., Dubois, J.M., Gratton, D.J., 1990. Evaluation of the grey-level co-occurrence matrix method for land-cover classification using SPOT imagery. *IEEE Transactions on Geoscience and Remote Sensing* 28, 513–519.
- Meyer, J.-Y., 2004. Threat of invasive alien plants to native flora and forest vegetation of Eastern Polynesia. *Pacific Science* 58, 357–375.
- Meyer, J.-Y., Butaud, J.-F., 2009. The impacts of rats on the endangered native flora of French Polynesia (Pacific Islands): driver of plant extinction or coup de grace species? *Biological Invasions* 11, 1569–1585.
- Michelson, D.B., Liljeberg, B.M., Pilesjö, P., 2000. Comparison of algorithms for classifying Swedish landcover using Landsat TM and ERS-1 SAR data. *Remote Sensing of Environment* 71, 1–15.
- Moore, I.D., Grayson, R.B., Ladson, A.R., 1991. Digital terrain modelling: a review of hydrological, geomorphological and biological applications. *Hydrological Processes* 5, 3–30.
- Moore, I.D., Gessler, P.E., Nielsen, G.A., Peterson, G.A., 1993. Soil attribute predicting using terrain analysis. *Soil Science Society of America Journal* 57, 443–452.
- Mueller-Dombois, D., Fosberg, F.R., 1998. *Vegetation of the Tropical Pacific Islands*. Springer Press, New York.
- Nagy, L., Grabherr, G., Körner, C., Thompson, D.B.A. (Eds.), 2003. *Alpine Biodiversity in Europe*. Springer Verlag, Berlin.
- Nyongui, A., Tonye, E., Akono, A., 2002. Evaluation of speckle filtering and texture analysis methods for land cover classification from SAR images. *International Journal of Remote Sensing* 23, 1895–1925.
- Pearson, R.G., Raxworthy, C.J., Nakamura, M., Peterson, A.T., 2007. Predicting species distributions from small numbers of occurrence records: a case using cryptic geckos in Madagascar. *Journal of Biogeography* 34, 102–117.
- Podest, E., Saatchi, S., 2002. Application of multiscale texture in classifying JERS-1 radar data over tropical vegetation. *International Journal of Remote Sensing* 23, 1487–1506.
- Pouteau, R., Meyer, J.-Y., Stoll, B., 2011a. A SVM-based model for predicting distribution of the invasive tree *Miconia calvescens* in tropical rainforests. *Ecological Modelling* 222, 2631–2641.
- Pouteau, R., Meyer, J.-Y., Taputuarai, R., Stoll, B., 2011b. A comparison between GARP model and SVM regression to predict species potential distribution: mapping the invasive *Miconia calvescens* on Moorea, French Polynesia. *Proc. International Symposium for Remote Sensing of Environment*, Sydney.
- Prasad, A.M., Iverson, L.R., Liaw, A., 2006. Newer classification and regression tree techniques: bagging and random forests for ecological prediction. *Ecosystems* 9, 181–199.
- Richardson, A.D., 2004. Foliar chemistry of balsam fir and red spruce in relation to elevation and the canopy light gradient in the mountains of the northeastern United States. *Plant and Soil* 260, 291–299.
- Riegel, G.M., Miller, R.F., Krueger, W.C., 1992. Competition for resources between understorey vegetation and overstorey *Pinus ponderosa* in Northeastern Oregon. *Ecological Applications* 2, 71–85.
- Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.-C., Müller, M., 2011. pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics* 12, 77.
- Rushton, S.P., Ormerod, S.J., Kerby, G., 2004. New paradigms for modelling species distribution? *Journal of Applied Ecology* 41, 193–200.
- Schölkopf, B., Smola, A., 2002. *Learning with Kernels*. MIT Press, Cambridge.
- Serpico, S.B., Roli, F., 1995. Classification of multisensor remote-sensing images by structured neural networks. *IEEE Transactions on Geoscience and Remote Sensing* 33, 562–578.
- Song, X., Fan, G., Rao, M., 2005. Automatic CRP mapping using nonparametric machine learning approaches. *IEEE Transactions on Geoscience and Remote Sensing* 43, 888–897.

- Spotswood, E.N., Meyer, J.-Y., 2009. Interactions between plants and avian frugivores in the Society Archipelago, French Polynesia. *Proc. Pacific Science Inter-Congress*, Papeete.
- Stockwell, D.R.B., Peterson, A.T., 2002. Effect of sample size on accuracy of species distribution models. *Ecological Modelling* 143, 1–13.
- Strahler, A.H., 1981. Stratification of natural vegetation for forest and rangeland inventory using Landsat digital imagery and collateral data. *International Journal of Remote Sensing* 2, 15–41.
- Thuiller, W., Araujo, M.B., Lavorel, S., 2003. Generalized models vs. classification tree analysis: predicting spatial distributions of plant species at different scales. *Journal of Vegetation Science* 14, 669–680.
- Turner, W., Spector, S., Gardiner, N., Fladeland, M., Sterling, E., Steininger, M., 2003. Remote sensing for biodiversity science and conservation. *Trends in Ecology & Evolution* 18, 306–314.
- Vapnik, V., 1998. *Statistical learning theory. Support Vector Machines for Pattern Recognition*. John Wiley & Sons, New York.
- Waske, B., Benediktsson, J.A., 2007. Fusion of support vector machines for classification of multisensor data. *IEEE Transactions on Geoscience and Remote Sensing* 45, 3858–3866.
- Whittaker, R.J., Fernandez-Palacios, J.M., 2007. *Island Biogeography: Ecology, Evolution and Conservation*, second ed. Oxford University Press, Oxford.
- Williams, J.N., Seo, C., Thornes, J., Nelson, J.K., Erwin, S., O'Brien, J.M., Schwartz, M.W., 2009. Using species distribution models to predict new occurrences for rare plants. *Diversity and Distributions* 15, 565–576.
- Wilson, J.P., Gallant, J.C., 2000. *Terrain Analysis: Principles and Applications*. John Wiley & Sons, New York.
- Wisz, M.S., Hijmans, R.J., Li, J., Peterson, A.T., Graham, C.H., Guisan, A., 2008. Effects of sample size on the performance of species distribution models. *Diversity and Distributions* 14, 763–773.
- Xie, Y., Sha, Z., Yu, M., 2008. Remote sensing imagery in vegetation mapping: a review. *Plant Ecology* 1, 9–23.