A Comparison of Machine Learning Algorithms for Classification of Tropical Ecosystems Observed by Multiple Sensors at Multiple Scales

R. Pouteau[a,*], A. Collin[b], B. Stoll[a]

[a] South Pacific Geosciences Laboratory, University of French Polynesia, BP 6570, 98702 Faa'a, French Polynesia - (robin.pouteau, benoit.stoll)@upf.pf
[b] Insular Research Center and Environment Observatory (CRIOBE), BP 1013 Papetoai, French Polynesia - antoinecollin1@gmail.com

Abstract – A substantial number of studies compare conventional classifiers (*e.g.* Maximum Likelihood, Decision Trees, Neural Networks or Support Vector Machines (SVM)) in a single location. We propose here an in-depth comparison of classifications by assessing the potential of SVM (often the "winner" in the previously mentioned studies) versus a range of the machine learning algorithms developed during the last decade: Naïve Bayes, C4.5 algorithm, Random Forest, Regression Tree and *k*-Nearest Neighbor. They were tested over different ecosystems across Moorea Island (French Polynesia) using various sensors. Our results show that SVM outperforms other classifiers in 75% of the situations. We point out that SVM has a successful ability to deal with complex pixel-by-pixel classification problems (with high level of details, speckle noise and few bands). This ability is intrinsic to the paradigm of SVM classification as (i) it is based on few meaningful pixels, *i.e.* support vectors; and (ii) it occurs in a high dimensional feature space even when bandwidth is narrow.

Keywords: Ecosystem; land cover; vegetation; marine; mapping; support vector machines (SVM); machine learning algorithms

1. INTRODUCTION

One of the most widespread applications in the field of remote sensing is certainly classification. It consists in grouping pixels of a scene into classes of objects to create a thematic representation. Today, an increasing number of sensors of greater diversity and higher resolution are available to the remote sensing community. We have thus the means to consider increasingly complex objects at finer and finer scales for classification.

For a long time remote sensing have focused on anthropogenic structures (urban areas, crop lands, forest plantations, *etc*), but an increasing number of studies are revolving around tropical ecosystems such as rainforests or coral reefs which are structurally complex objects with global stakes. Complexity of tropical ecosystems does not lie in its spatial organization at the landscape scale but at the finer scales of the community or the population. Complex classes are made of several species (generally more than in non-tropical areas) with individual variation of growth and development, phenology, disturbance and stress influencing the spectral response of the ecosystems. The transition area between two adjacent but different communities ("ecotone") in marine as well as in terrestrial ecosystems is also progressive and the limit between two classes is intrinsically thematician-dependant (Andréfouët and Roux, 1998; Pouteau *et al*., 2010).

Another reason of the development of applications on tropical ecosystems is the emergence of more and more efficient classification methods. The success of a classifier lies in its performance (in terms of accuracy) and its constancy (in various situations). However, a substantial number of studies compare conventional classifiers (*e.g.* Maximum Likelihood, Decision Trees, Neural Networks and Support Vector Machines (SVM)) in a single geographic location. We propose here an in-depth comparison of classifications by assessing the potential of SVM (often the "winner" in the previously mentioned studies) versus a range of the machine learning algorithms developed during the last decade. They were tested over various geographic locations at multiple scales and on multiple remotely sensed data.

2. MATERIAL AND METHODS

2.1 Study site

Moorea is the fourth highest island in French Polynesia (South Pacific) with a highest point, mont Tohiea reaching 1207 m. It is 134 km² with a shape vaguely resembling a triangle with two nearly symmetrical bays opening to the north side: the Cook's and Opunohu Bays. It was selected because of the diversity of land covers and remotely sensed data from multiple sensors available on this island.

A range of areas was selected across the Moorea Island including anthropogenic terrestrial ecosystem (ATE), natural terrestrial ecosystem (NTE) (native rainforest) and marine ecosystem (ME) (coral reef) to test whether our results are generic (Figure 1). For imagery with a decametric resolution, the full island was considered; for imagery with metric resolution, the red square as ATE and the green rectangle as NTE were considered; and for imagery with sub-metric resolution, the yellow square as ATE, the blue square as ME and the magenta square as NTE were considered. These subsets were chosen in order to consider comparable number of pixels for each source.

* Corresponding author.
** This work was supported in part by the French Marine Protected Areas Agency and the Research Department of the Government of French Polynesia.

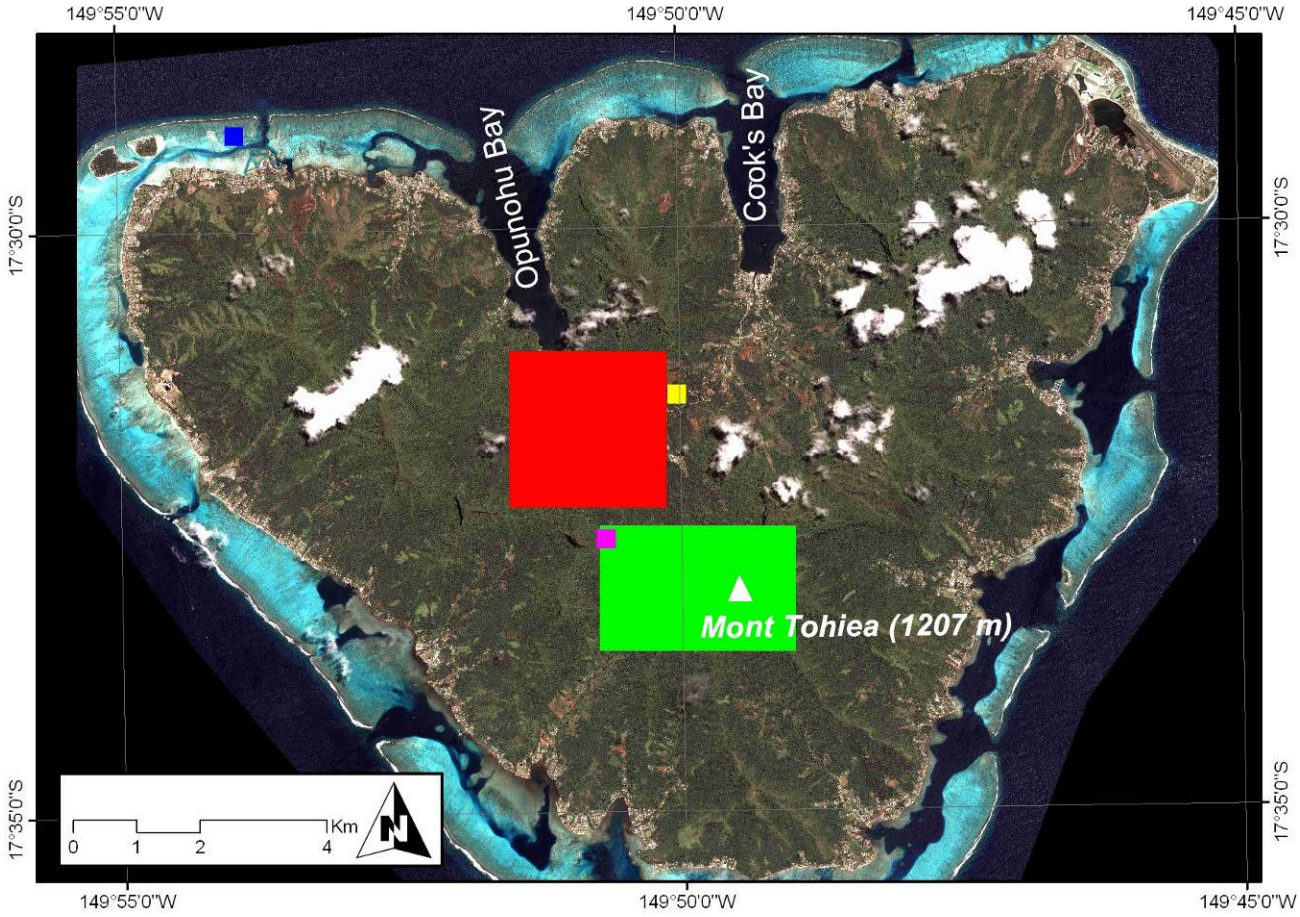

Figure 1. Presentation of the Moorea Island and localization of the subsets considered for classification: for imagery with a decametric resolution, the full island was considered; for imagery with metric resolution, the big rectangles were considered; and for imagery with sub-metric resolution, the small squares were considered. For color meaning, please see the second paragraph of part 2.1.

Table A. Main characteristics of image products from the different sensors

Sensor	Resolution (m)	Date	# bands
Landsat-7 ETM+	30.0	2000	3
SPOT	20.0	1986	3
AirSAR	5.00	2000	4
TerraSAR-X	2.75	2009	2
Quickbird	0.60	2006	4
WorldView-2	0.50	2010	8

2.2 Remotely sensed data

The analysis was based on both Synthetic Aperture Radar (SAR) and multispectral data (Table A). SAR data was used for terrestrial areas only. SAR data available on Moorea include a 5 m-resolution AirSAR scene from 2000 with 4 bands: Cvv, Lhh, Lhv, Lvv and a 2.5 m-resolution TerraSAR-X scene from 2009 with also 4 bands: Xvv, Xvh, Xhv, Xhh. On the other hand, multispectral data used in this study include a 30 m-resolution Landsat-7 ETM+ scene from 2000 with 3 bands, a 20 m-resolution SPOT scene from 1986 with 3 bands, a 0.6 m-resolution pan-sharpened Quickbird scene from 2006 with 4 bands and a 0.5 m-resolution pan-sharpened WorldView-2 scene from 2010 with 8 bands.

2.3 Compared machine learning algorithms

SVM, introduced by Vapnik (1998), was compared with a range of machine learning algorithms developed during the last decade and, to our knowledge, rarely compared: Naïve Bayes (Rish, 2001), C4.5 algorithm (Quinlan, 1996), Random Forest (RF; Breiman, 2001), Boosted Regression Tree (BRT; Lawrence *et al*., 2004) and *k*-Nearest Neighbors (*k*NN; Franco-Lopez *et al*., 2001). We used the Weka workbench (http://www.cs.waikato.ac.nz/ml/weka/) implementation of the above mentioned algorithms except for SVM for which a LIBSVM implementation was used into the ENVI/IDL v4.6 environment. As recommended by Hsu *et al.* (2010), optimal C and γ parameters were found by means of the grid-search method using cross-validation. After several tests and following the literature, the Radial Basis Function (RBF) kernel and the One Against One (OAO) algorithm have been retained (see Hsu *et al*., 2010 for more details on these methods).

Pixel-by-pixel classifications were processed with training sets containing 250 pixels for each of the 4 classes (fixed because number of classes affects accuracy according to Andréfouët *et al.* (2003)) adapted to the level of spatial details available in the images (landscape, community or

species levels). Accuracy assessment was based on 250 pixels per class.

3. RESULTS

No significant differences between overall accuracy (OA) obtained by classifying ATE, NTE and ME were found, what suggest that machine learning algorithms developed during the last decade are able to manage complexity due to the structure of NTE and ME.

Our results show that SVM is a relevant machine learning algorithm for classification of tropical ecosystems (Figure 2) since it outperforms other classifiers in 75% of the situations. *k*NN better performs for the Landsat-7 ETM+ classification, as well as BRT for the AirSAR classification of the NTE and RF for the TerraSAR-X classification of the NTE. Except in this case, we denote that RF performs far worse than the other compared classifiers.

SVM is successfully able to deal with complex classification problems. Indeed smaller is the number of bands (the spectral resolution) and finer is the level of details (the spatial resolution), generally worst is the classification OA (Figure 3). As an example, classification processed on the 30 m-resolution Landsat-7 ETM+ scene performs worse than classification processed on the 0.60 m-resolution Quickbird scene. Likewise, classification processed on the 4 bands Quickbird scene performs worse than the classification processed on the 8 bands WorldView-2 scene. Nevertheless, we point out that the OA achieved by the SVM decreases less than the other classifiers when the complexity of the classification process increases. In other words, the difference between the OA achieved by the SVM and the OA achieved by the other classifiers generally increases when the complexity increases.

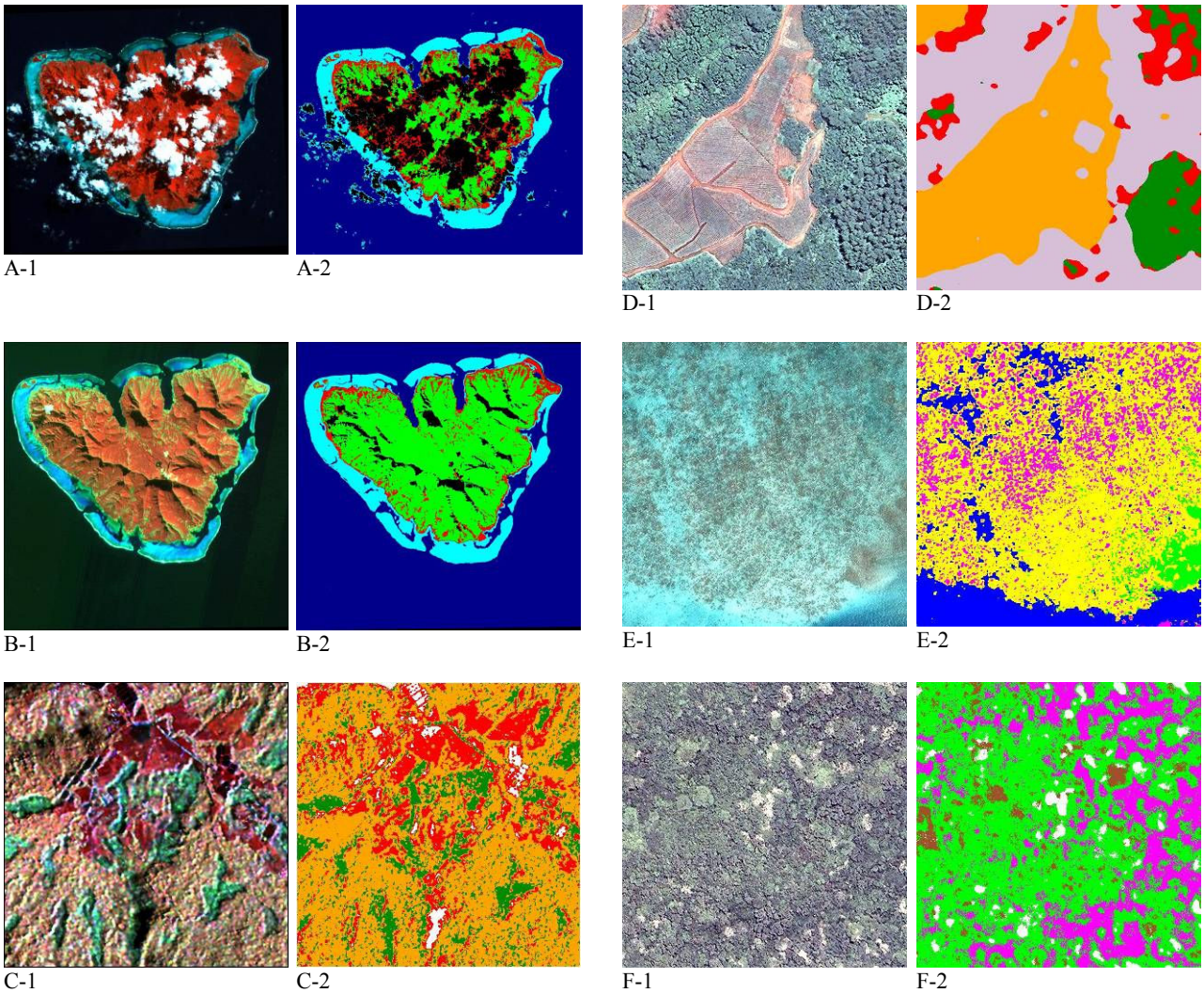

Figure 2. A sample of the scenes used for the analysis and the corresponding SVM classifications. A-1 is the Landsat-7 ETM+ scene and B-1 is the SPOT one. In A-2 and B-2, green represents non-anthropogenic areas, red is anthropogenic areas, cyan is coral reef and blue is deepwater. C-1 is a composite view of an anthropogenic area from AirSAR with Cvv band as red, Lhv as green and Lhh as blue. In C-2, orange is secondary forest, green is tree plantations (*Pinus caribaea* and *Falcataria moluccana*), red is crop lands and white is ponds (water stocks, shrimp farming). D-1, E-1 and F-1 are Quickbird scenes. In D-2, orange is pineapple fields, green is *Tectona grandis* plantation, red is *Casuarina equisetifolia* plantation and thistle is secondary forest (mainly *Falcataria moluccana*). In E-2, blue is detritic sedimentation, green is sand, yellow is living coral and magenta is dead coral. In F-2, green is dominated by *Hibiscus tiliaceus*, magenta by *Neonauclea forsteri*, brown by *Inocarpus fagifer* and white by *Aleurites moluccana*.

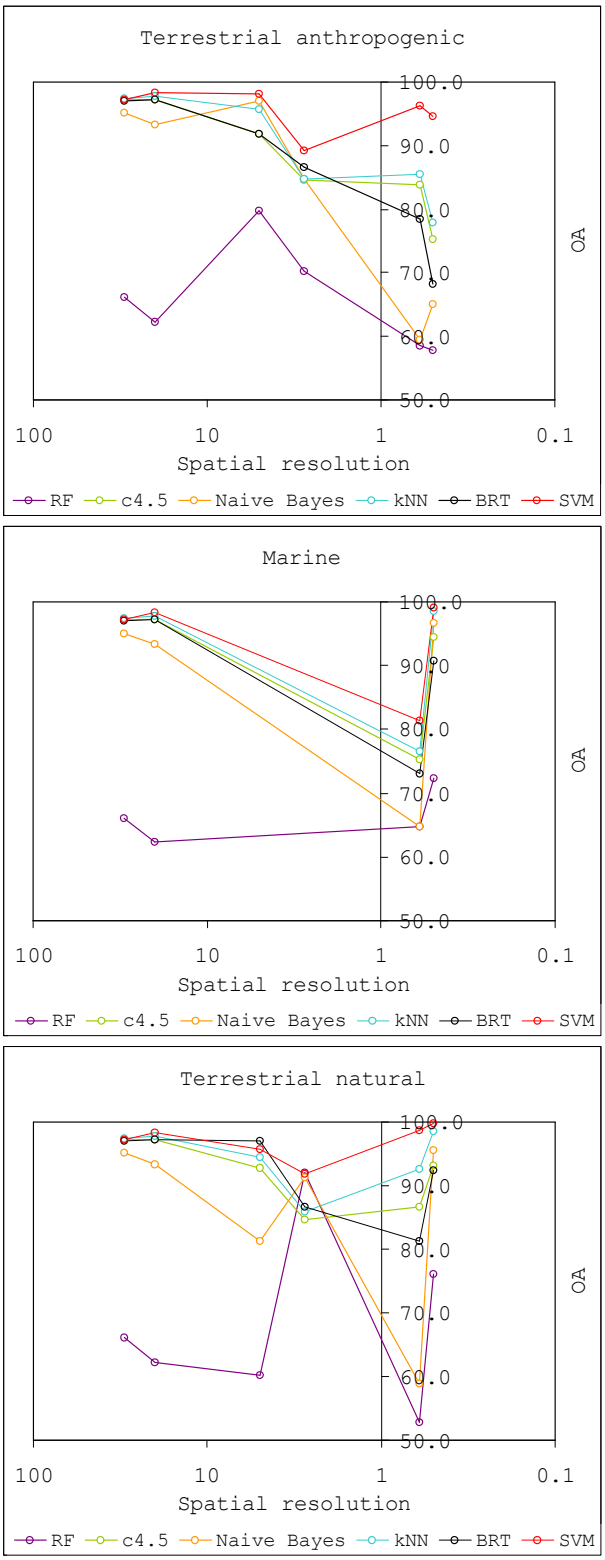

Figure 3. Overall accuracy (OA) achieved by a range of machine learning algorithms applied on multiple sensors with multiple spatial resolutions. For algorithms abbreviation, please see the first paragraph of part 2.3.

4. DISCUSSION

SVM is successfully able to deal with complex classification problems with high level of details, speckle noise and few bands. We argued that this ability is intrinsic to the paradigm of SVM classification. Indeed, the classification process occurs in a high dimensional feature space (Vapnik, 1998) which can arguably explain good performances achieved by SVM even when bandwidth is narrow. Moreover, high spatial resolution and presence of a lot of speckle noise produce a large amount of unmeaningful pixels. SVM seems less sensitive to them than other classifiers since, as pointed out by Foody and Mathur (2006), its classification process is based on few meaningful pixels, *i.e.* support vectors.

From a practical point of view, regarding these results, we demonstrate that SVM does not provide any major benefit face to other compared classifiers (except RF) when used in "simple" classification problems (with a coarse spatial resolution or a satisfactory number of bands). On the contrary, when a user is dealing with a complex problem (with a high spatial resolution and a low spectral resolution), SVM is significantly the best candidate among those tested in this experiment.

ACKNOWLEDGEMENTS

The authors are grateful to the Government of French Polynesia and its *Service de l'Urbanisme* (Urbanism Department) for providing remotely sensed data.

REFERENCES

S. Andréfouët, L. Roux, "Characterisation of ecotones using membership degrees computed with a fuzzy classifier," *International Journal of Remote Sensing*, vol. 19, pp. 3205-3211, 1998.

S. Andréfouët, P. Kramer, D. Torres-Pulliza, K. E. Joyce, E. J. Hochberg, R. Garza-Pérrez, P. J. Mumby, B. Riegl, H. Yamano, W. H. White, M. Zubia, J. C. Brock, S. R. Phinn, A. Naseer, B. G. Hatcher, F. E. Muller-Karger, "Multi-sites evaluation of IKONOS data for classification of tropical coral reef environments," *Remote Sensing of Environment*, vol. 88, pp. 128-143, 2003.

L. Breiman, "Random Forest," *Machine Learning*, vol. 45, pp. 5-32, 2001.

G. M. Foody, A. Mathur A., "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*, vol. 103, pp. 179-189, 2006.

H. Franco-Lopez, A. R. Ek, M. E. Bauer, "Estimation and mapping of forest stand density, volume, and cover type using the *k*-nearest neighbors method," *Remote Sensing of Environment*, vol. 77, pp. 251-274.

R. Lawrence, A. Bunn, S. Powell, M. Zambon, "Classification of remotely sensed imagery using stochastic gradient boosting as a refinement of classification tree analysis," *Remote Sensing of Environment,* vol. 90, pp. 331-336, 2006.

R. Pouteau, B. Stoll, S. Chabrier, "Ground truth method assessment for SVM-based landscape classification," *Proc. IGARSS*, pp. 2715-2718, 2010.

J. R. Quinlan, "Improved use of continuous attributes in C4.5," *Journal of Artificial Intelligence Research*, vol. 4, pp. 77-90, 1996.

I. Rish, "An empirical study of the naïve Bayes classifier," *International Joint Conferences on Artificial Intelligence*, 2001.

V. Vapnik, *Statistical learning theory*, New York, Wiley, 1998.