Forest Systems
27 (3), eSC03, 5 pages (2018)
eISSN: 2171-9845
https://doi.org/10.5424/fs/2018273-13686
Instituto Nacional de Investigación y Tecnología Agraria y Alimentaria O. A., M. P. (INIA)

SHORT COMMUNICATION

OPEN ACCESS

Stand types discrimination comparing machine-learning algorithms in Monteverde, Canary Islands

Miguel García-Hidalgo*[1], Ángela Blázquez-Casado[1,2], Beatriz Águeda[1,2] and Francisco Rodríguez[1,2]

[1]EiFAB - iuFOR Universidad de Valladolid. Campus Duques de Soria. 42004 Soria. Spain. [2]föra forest technologies. Campus Duques de Soria. 42004 Soria. Spain.

Abstract

Aim of study: The main objective is to determine the best machine-learning algorithm to classify the stand types of Monteverde forests combining LiDAR, orthophotography, and Sentinel-2 data, thus providing an easy and cheap method to classify Monteverde stand types.

Area of study: 1500 ha forest in Monteverde, North Tenerife, Canary Islands.

Material and methods: RF, SVML, SVMR and ANN algorithms are used to classify the three Monteverde stand types. Before training the model, feature selection of LiDAR, orthophotography, and Sentinel-2 data through VSURF was carried out. Comparison of its accuracy was performed.

Main results: Five LiDAR variables were found to be the most efficient for classifying each object, while only one Sentinel-2 index and one Sentinel-2 band was valuable. Additionally, standard deviation and mean of the Red orthophotography colour band, and ratio between Red and Green bands were also found to be suitable. SVML is confirmed as the most accurate algorithm (0.904, 0.041 SD) while ANN showed the lowest value of 0.891 (0.073 SD). SVMR and RF obtain 0.902 (0.060 SD) and 0.904 (0.056 SD) respectively. SVML was found to be the best method given its low standard deviation.

Research highlights: The similar high accuracy values among models confirm the importance of taking into account diverse machine-learning methods for stand types classification purposes and different explanatory variables. Although differences between errors may not seem relevant at a first glance, due to the limited size of the study area with only three plus two categories, such differences could be highly important when working at large scales with more stand types.

Additional keywords: RF algorithm, SVML algorithm, SVMR algorithm, ANN algorithm, LiDAR, orthophotography, Sentinel-2.

Abbreviations used: ANN, artificial neural networks algorithm; Band04, Sentinel-2 band 04 image data; BR, brezal; DTHM, digital tree height model; DTHM-2016, digital tree height model based on 2016 LiDAR data; DTM, digital terrain model; DTM-2016, digital terrain model based on 2016 LiDAR data; FBA, fayal-brezal-acebiñal; FCC, canopy cover; HEIGHT-2009, maximum height based on 2009 LiDAR data; HGR, height growth based on 2009 and 2016 LiDAR data; LA, laurisilva; NDVI705, Sentinel-2 index image data; NMF, non-Monteverde forest; NMG, non-Monteverde ground; P95-2016, height percentile 95 based on 2016 LiDAR data; RATIO R/G, ratio between Red and Green bands orthophotograph data; RED, Red band orthophotograph data; Red-SD, standard deviation of the Red band orthophotograph data; RF, random forest algorithm; SVM, support vector machine algorithm; SVML, linear support vector machine algorithm; SVMR, radial support vector machine algorithm; VSURF, variable selection using random forest.

Authors´ contributions: All authors meet the criteria for authorship described in the instructions for authors. Specifically, the analysis was conceived and designed by: MGH, ABC, BA and FR; the data was analysed by: MGH and ABC; the paper was written by: BA, MGH and ABC. All authors reviewed the paper.

Citation: García-Hidalgo, M., Blázquez-Casado, A., Águeda, B., Rodríguez, F. (2018). Stand types discrimination comparing machine-learning algorithms in Monteverde, Canary Islands. Forest Systems, Volume 27, Issue 3, eSC03. https://doi.org/10.5424/fs/2018273-13686

Supplementary material (Figure S1) accompanies the paper on FS's website.

Received: 13 Jul 2018. **Accepted:** 30 Oct 2018.

Funding: Research by Ángela Blázquez-Casado was funded by DI-14-06953 fellowship (Ministerio de Economía, Industria y Competitividad, Spanish Government). Research by Beatriz Águeda was funded by PTQ-16-08411 fellowship (Ministerio de Economía, Industria y Competitividad, Spanish Government).

Competing interests: The authors have declared that no competing interests exist.

Correspondence should be addressed to Miguel García-Hidalgo: garhidmi@gmail.com

Introduction

During the second half of the 20th century traditional forestry practices used in the Monteverde forest in Canary Islands changed dramatically. Moreover, the beginning of this century is marked by a shift in the preferences of society as regards the ecosystem services provided by the Monteverde forest from traditional forest

resources to conservation orientated services, especially as part of one of the current areas of Macaronesia and the importance of its relict flora (Arozena & Panareda, 2013). Nowadays, Monteverde conservation is explained by its uniqueness and the need for monitoring Laurisilva dynamics which lead to spot new and innovative classification tools. Recent remote sensing technologies can help to improve its management, providing easy and cheap classification of its stand types reducing cost and time consumption from traditional forest management procedures based on expensive field studies.

Data derived from active and passive remote sensors are of great interest in forestry. In particular, the combination of LiDAR information with Sentinel-2 multispectral images provides a powerful tool for classifying forests with high densities and stocking rates, thus reducing the cost of the estimation process (Zhu *et al.*, 2017). In addition, the volume of data we are dealing with is constantly growing, including the aim at retrieving a wide variety of geographic and ecological characteristics. Consequently, the analyses can only be tackled using computational methods.

From the wide range of algorithms used to find the rules for object classification in forest sciences, the random forest algorithm (RF) has shown high rates of accuracy. Nevertheless, the Linear and Radial Support Vector Machine (SVML, SVMR) and Artificial Neural Networks (ANN) algorithms are increasingly being taken into consideration in this area (Nitze *et al.*, 2012; Valbuena *et al.*, 2016; Vega Isuhuaylas *et al.*, 2018; Xu *et al.*, 2018). Although its accuracy is not always taken into account, its effectiveness could be easily increased only keeping in mind the most suitable one.

The main objective of this work is to determine the best machine-learning algorithm to classify the three Monteverde stand types, Canary Islands, combining LiDAR, orthophotography and Sentinel-2 data.

Material and Methods

Study area

The study area is located in a 1500-ha evergreen forest in the North of the island of Tenerife in the Canary Islands, between 200-1300 m.a.s.l Fig. S1 [suppl.]). The ecosystem is highly valuable in economic terms but also as regards ecosystem services, characterised by a wide variety of species.

Field data

A total number of 259 objects were measured during May of 2017. The objects were irregular in shape, collecting homogenous remote information. Each of ones included visual information about stand type (main species and GPS coordinates). Objects were selected with the aim of picking up the highest spectral variability of stand types in Monteverde, three forest categories were defined according to its relevance in management (Arozena & Panareda, 2013): (i) Brezal (BR), composed by *Erica arborea* shrubby stands with variable cover and occasional presence of scrub; (ii) Fayal-Brezal-Acebiñal (FBA), composed by shrub or tree stands with a high density and average diameter of saplings 5-10 cm; the proportion of *E. arborea* in the specific composition of the stand varies and stands may be dominated by other species such as *Morella faya*, *Laurus novocanariensis*, or *Viburnum rigidum* and other companion species at different stand types of development; (iii) Laurisilva (LA), mixed stands with a significant presence of *L. novocanariensis* and *V. rigidum* with average diameter greater than 10 cm and the rare presence of *E. arborea*. In addition, two more stand types were included in the database: (iv) Non-Monteverde ground (NMG), defined as bare ground or scrub less than 2 m high; and (v) Non-Monteverde forest (NMF) defined as stand cover composed by other species. The number of samples set out in each category were 32 for BR, 143 for FB, 27 for LA, 6 for NMG, and 51 for NMF, respectively.

LiDAR data

The island of Tenerife was scanned using a LiDAR sensor in 2009 and 2016, with an average nominal point density of 0.5 pulses m^2 (PNOA project, Spanish Government). Data from the study area were provided in digital files of 2x2 km extension. Point clouds were automatically classified and coloured, taking RGB orthophotos as a reference. LiDAR data was processed using FUSION software (McGaughey, 2007) and several raster variables (5 m resolution) were generated: digital tree height model (DTHM), digital terrain model (DTM), canopy cover (FCC), height percentile 95 (P95), height percentile 25 (P25), height growth between 2016 and 2009 (HGR), along with standard deviations for all these variables.

Multispectral imagery data sources

European Space Agency Sentinel-2 satellite images (10 m and 20 m resolution) of the study area captured in December 2015 and January 2017 were employed in order to avoid clouds and deciduous tree reflectance. These orthorectified and atmospherically corrected images were downloaded from Copernicus Open Access Hub (https://scihub.copernicus.eu/). Several vegetation

indexes were calculated based on imagery data: NDVI, RNDVI, GNDVI, SAVI, LAI-SAVI, SR, and EVI (10 m resolution), and NDVI705, NDWI, RNDWI, NDII, NDI45, NBR, MSI (20 m resolution) (Henrich *et al.*, 2012). In addition, aerial orthophotograph images were provided by CNIG-PNOA (Spanish Government), with a resolution of 25 cm in the official reference geodetic system, REGCAN95 - UTM zone 28N projection. Medium value and standard deviation of each of its bands, and the ratio between Red and Green bands, were calculated. All these spectral indices were included as potential predictors in the classification model.

Segmentation process

To define object segmentation in the study area, we executed an Object Based Image Analysis which created an image-object through the aggregation of pixels by image segmentation from the Orfeo Tool Box (OTB Development Team, 2017). The two variables we worked with are: (i) the spatial resolution and (ii) range domains, which is the allowable spectral range within each segment for each band at a minimum scale of 40 m^2.

Data features from each source of information were assigned to the generated objects from segmentation using QGIS Zonal Statistics Plugin (QGIS Development Team, 2017).

Data analysis

Prior to model training, feature selection using the Variable Selection Using Random Forest (VSURF) was performed (Genuer *et al.*, 2015). In order to conduct a useful comparison between RF, SVML, SVMR, and ANN, caret package in R Software was run using the *rf, svmLinear, svmRadial* and *nnet* methods with default parameters (Kuhn *et al.*, 2018). Furthermore, the 'overfitting' problem was reduced by Cross Validation using 10 folds with three repetitions.

Results and discussion

The application of VSURF procedure selected 10 features (Fig. 1). Five LiDAR variables were found as the most efficient for classifying each object, while only one Sentinel-2 index and one Sentinel-2 band was valuable. Additionally, standard deviation and mean of the Red orthophotography colour band, and ratio between Red and Green bands were also found to be suitable.

When the data dispersion was analysed, selected LiDAR features showed differences according to the classification factor variable. The feature DTHM-2016 reveals importance at its clear boundaries between the different forest typologies and it shows the difference between forest typologies stage, together with P95-2016 and HGR variables (Fig. 1). The rest of the LiDAR data support distinction and split soil from the other types. The selection of Sentinel-2 NDVI705 reflects red edge radiation and its usefulness with very high spectral resolution reflectance data (Sinergise, 2018). In contrast, the average and the standard deviation of Red orthophotography colour band showed a crucial disjunction among NMF and Monteverde.

From the Cross Validation results, most of the models reached 0.90 mean accuracy (Table 1). SVML was confirmed as the most accurate method while ANN presented the lowest accuracy value. SVMR and RF obtained the intermediated accuracy values of 0.902 and 0.904 respectively. Simultaneously, Cohen Kappa values did not vary from the achieved accuracy values. Given its low standard deviation, SVML was found as the best method thanks to the variable influence shown in Table 2 for Monteverde stand classification types. Usefulness of LiDAR variables in the classification of this case study is demonstrated by the fact that four (DTMH-2016, P95-2016, HEIGHT-2009, and DTM-2016) out of ten selected variables have the highest accuracy, being the one LiDAR variable remaining (HGR) the sixth one (Table 2).

Obtained results differ with other previous ones in the literature. For instance, Valbuena *et al.* (2016) found that the best machine learning algorithm to determine Mediterranean forest development stages are RF and ANN, and Vega Isuhuaylas *et al.* (2018) met with SVM and RF to classify Andes mountain forests and shrubland land cover classes.

The high accuracy values confirm the importance of taking into account diverse machine-learning methods for stand classification purposes and different explanatory aspects. Notwithstanding the small deviation between accuracy values, our work proves that SVML is the best algorithm for the Monteverde forest classification due to the minimal results' scattering. Over 90% of cases Monteverde stand type (BR, FBA, LA, NMG, or NMF) determined from remote sensing data are correct, though small size of the study area and the only three, plus two, stand types considered here should not be forgotten. Our results confirm machine learning classification is a suitable tool to optimize classification in Monteverde forest and, thus, its management.

Boost of machine learning algorithms applied to classify forest is broadly enough demonstrated. Although differences between errors in accuracy and scattering may not seem significant at a first glance,

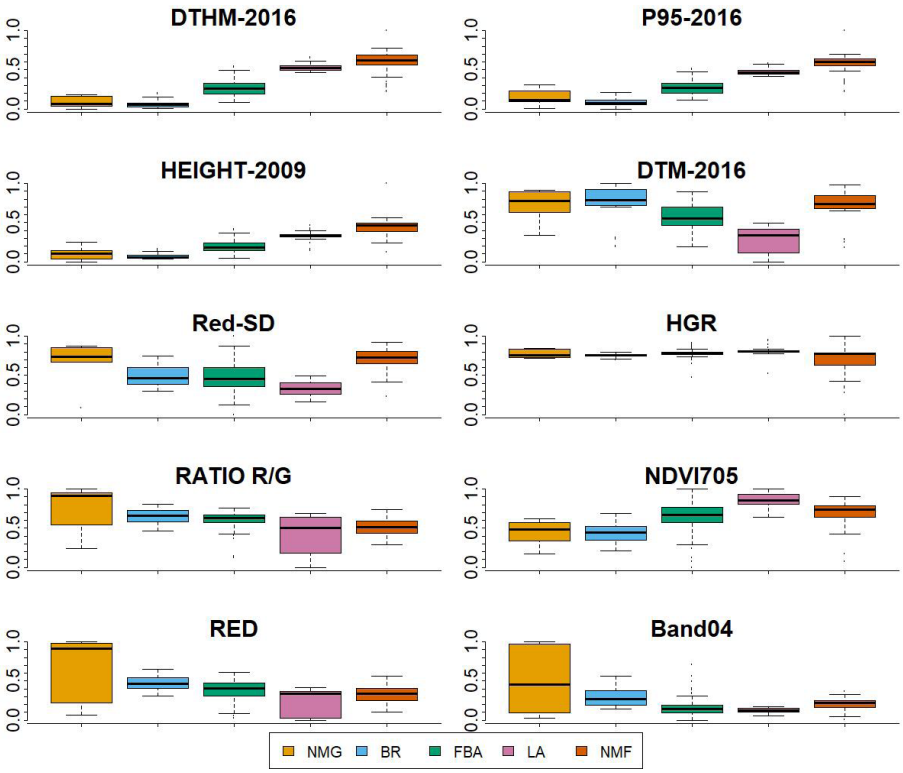

Figure 1. Distribution of selected features as a result of variable selection applying random forest (VSURF) according to Monteverde typologies. Values under normalization from minimum '0' to maximum value '1'. Abbreviations should be checked at the 'abbreviations used' section.

such differences are highly important when working at large scales. Errors may be higher when classifying larger areas with more stand types. So, comparisons between algorithms should be considered when stand classification analyses are performed owing to different behaviour of algorithms relying on stand types, features and size of the study area.

Acknowledgements

The Authors wish to thank the Cabildo de Tenerife staff for their invaluable help on the fieldwork. LiDAR data was provided by CNIG-PNOA (Spanish Government), who also provided orthophotographic images. Sentinel images were provided by ESA (European Space Agency). Tecnosylva provided support to the authors in the image segmentation

Table 1. Accuracy, Accuracy Standard Deviation, Kappa Value and Kappa Value Standard Deviation resulted from each model.

	Accuracy	Accuracy SD	Kappa	Kappa SD
ANN	0.891	0.073	0.817	0.129
SVML	0.904	0.041	0.842	0.070
SVMR	0.902	0.060	0.836	0.106
RF	0.904	0.056	0.841	0.092

Table 2. Importance of variables in the linear support vector machine algorithm (SVML) final model for the Monteverde typology classification. Values state the loss of accuracy and/or standard deviation according to the lack of each feature from calculated SVML. Abbreviations should be checked at the 'abbreviations used' section.

	Source	Mean	Standard Deviation
SVML		0.904	0.041
DTHM-2016	LiDAR	0.892	0.055
P95-2016	LiDAR	0.903	0.052
HEIGHT-2009	LiDAR	0.909	0.045
DTM-2016	LiDAR	0.880	0.049
Red-SD	Orthophotography	0.908	0.043
HGR	LiDAR	0.886	0.057
RATIO R/G	Orthophotography	0.907	0.046
NDVI705	Sentinel	0.900	0.042
RED	Orthophotography	0.915	0.044
Band04	Sentinel	0.915	0.049

process. We also thank Adam Collins for the English language advice.

References

Arozena ME, Panareda, JM, 2013. Forest transition and biogeographic meaning of the current laurel forest landscape in Canary Islands, Spain. Physical Geography 34: 211-235. https://doi.org/10.1080/02723646.2013.817181

Genuer R, Poggi JM, Tuleau-Malot C, 2015. VSURF: An R Package for Variable Selection Using Random Forests. https://journal.r-project.org/archive/2015/RJ-2015-018/RJ-2015-018.pdf.

Henrich V, Krauss G, Götze C, Sandow C, 2012. IDB - https://www.indexdatabase.de/, Entwicklung einer Datenbank für Fernerkundungsindizes. AK Fernerkundung, Bochum 45: 10.

Kuhn M, Jed Wing A, Weston S, Williams A, Keefer C, Engelhardt A, Cooper T, *et al.*, 2018. caret: Classification and Regression TrainingR package version 6.0-78. https://cran.r-project.org/web/packages/caret/index.html.

McGaughey R, 2007. Fusion/LDV: Software for Lidar Data Analysis and Visualization; USDA Forest Service, Pacific Northwest Research Station: Portland, OR, USA.

Nitze I, Schulthess U, Asche H, 2012. Comparison of machine learning algorithms Random Forest, Artificial Neural Network and Support Vector Machine to maximum likelihood for supervised crop type classification. Proc GEOBIA, Rio de Janeiro (Brazil), May 7-9, pp: 35.

OTB Development Team, 2017. The ORFEO Tool Box Software Guide. Updated for OTB-5.10.0

QGIS Development Team, 2017. QGIS Geographic Information System. Open Source Geospatial Foundation Project. https://www.qgis.org/en/site/

Sinergise, 2018. Sentinel 2 EO products. https://www.sentinel-hub.com/develop/documentation/eo_products/Sentinel2EOproducts.

Valbuena R, Maltamo M, Packalen P, 2016. Classification of forest development stages from national low-density LiDAR datasets: a comparison of machine learning methods. Revista de Teledetección 45: 15-25. https://doi.org/10.4995/raet.2016.4029

Vega Isuhuaylas LA, Hirata Y, Ventura Santos LC, Serrudo Torobeo N, 2018. Natural forest mapping in the Andes (Peru): a comparison of the performance of machine-learning algorithms. Remote Sens 10: 782. https://doi.org/10.3390/rs10050782

Xu C, Manley B, Morgenroth J, 2018. Evaluation of modelling approaches in predicting forest volume and stand age for small-scale plantation forests in New Zealand with RapidEye and LiDAR. Int J Appl Earth Obs Geoinformation 73: 386-396. https://doi.org/10.1016/j.jag.2018.06.021

Zhu XX, Tuia D, Mou L, Xia GS, Zhang L, Xu F, Fraundorfer F, 2017. Deep Learning in Remote Sensing: A Comprehensive Review and List of Resources. IEEE Geosci Remote Sens Mag 5: 8-36. https://doi.org/10.1109/MGRS.2017.2762307