Performance of machine learning algorithms for forest species classification using WorldView-3 data in the Southern Alentejo region, Portugal

Ana Margarida Coelho[1], Adélia M. O. Sousa[2], Ana Cristina Gonçalves[3]

[1]ICT - Instituto de Ciências da Terra, Instituto de Investigação e Formação Avançada, Departamento de Engenharia Rural, Escola de Ciências e Tecnologia, Universidade de Évora, Apartado 94, Évora, Portugal, ana.coelho@uevora.pt;

[2]MED – *Mediterranean Institute for Agriculture, Environment and Development* & CHANGE – *Global Change and Sustainability Institute*, Laboratório de Deteção Remota (EaRSLab), Instituto de Investigação e Formação Avançada, Departamento de Engenharia Rural, Escola de Ciências e Tecnologia, Universidade de Évora, Apartado 94, 7002-544 Évora, Portugal, asousa@uevora.pt

[3]MED – *Mediterranean Institute for Agriculture, Environment and Development* & CHANGE – *Global Change and Sustainability Institute*, Instituto de Investigação e Formação Avançada, Departamento de Engenharia Rural, Escola de Ciências e Tecnologia, Universidade de Évora, Apartado 94, 7002-544 Évora, Portugal, acag@uevora.pt.

Abstract

Recent advances in remote sensing technologies and the increased availability of high spatial resolution satellite data allow the acquisition of detailed spatial information. These data have been used for monitoring the Earth's surface, namely monitoring land use land cover, quantifying biomass and carbon, and evaluating the protection and conservation of forest areas. O WorldView-3 is a high spatial resolution satellite (0.50m) with 8 multispectral bands (visible and infrared) which allows obtaining detailed data from the Earth's surface.

This study aims to map the forest occupation by specie with two WoldView-3 images, and to evaluate the performance of machine learning classifiers (maximum likelihood, support vector machine and random forest) in two regions of Alentejo, south of Portugal. The main forest species are Quercus suber in one region and Quercus rotundifolia in another. The procedures performed were multiresolution image segmentation and object-oriented classification based on 4 bands (blue, green, red and near infrared). As auxiliary data, vegetation indices (NDVI and SAVI) and principal components were calculated.

In the object-oriented classification process, the three classifiers were tested. The support vector machine classifier was the one that presented the best accuracy (kappa and overall accuracy), for both images, allowing to obtain good results in the identification of forest species. In the image dominated by Quercus suber, the values of kappa and overall accuracy were 90% and 95%, and for the image where Quercus rotundifolia predominated, 90% and 96% respectively. The methodology applied to the high spatial resolution satellite data showed very good results in the identification and mapping of main forest species. Higher precision values stand out for the image where the Quercus rotundifolia predominates, where there is less spectral variation, namely fewer land use classes, thus reducing errors between classes that may be spectrally similar.

Keywords: maximum likelihood, support vector machine, random forest, multiresolution image segmentation, object-oriented classification, vegetation indices

Introduction

Recent advances in remote sensing technologies and the increased availability of high-resolution satellite data for Earth's surface enable the acquisition of detailed spatial information. High-resolution satellite images have been used for land surface monitoring, including land use/land cover monitoring, ecological processes analysis, biomass and carbon quantification, as well as assessment of forest protection and conservation efforts.

High resolution satellite images (e.g., WorldView-3) have been used to monitor land use/land cover (Galidaki et al., 2017), ecological processes (Ahmad et al., 2021; Galidaki et al., 2017), biomass (Gonçalves et al., 2019, 2017; Sousa et al., 2015; Galidaki et al., 2017) and carbon (Ahmad et al., 2021; Galidaki et al., 2017) quantification as well as their variability in space and time (Gonçalves et al., 2019, 2017). The classification of the satellite images and their accuracy are of the utmost importance (Meng & Xiao, 2011; Varin et al., 2020) several methods have been used in image classification ((Meng & Xiao, 2011; Varin et al., 2020; Vibhute et al., 2016) among which the multiresolution segmentation and object oriented classification ((Meng & Xiao, 2011). Moreover, several mathematical models have been used in object oriented classification, such as random forest (Varin et al., 2020), support vector machine (Varin et al., 2020) and maximum likelihood (Vibhute et al., 2016). The variability of the landscape in general, and of the forest areas in particular, make image classification a challenge. The diversity of the models reflects the need to accommodate the variability of the areas under analysis.

Apart from the selection of the models to classify the images, the selection of the explanatory variables plays also a key role in the accuracy of the classification (Varin et al., 2020), bands are frequently used, yet they do not enable to identify and delimit with accuracy some land uses (Varin et al., 2020). Vegetation indices, combining two or more bands enable a better differentiation and delimitation of different land uses (Varin et al., 2020). Moreover, they enable the isolation of tree canopies from other land use types because the enhance their differences (Fonseca & Fernandes,

2004; Marcussi et al., 2010; Che et al., 2019).

Two vegetation indices area frequently used Normalized Difference Vegetation Index (NDVI) and Soil-Adjusted Vegetation Index (SAVI). The NDVI is a widely used vegetation index for vegetation identification and discrimination between species. It is employed because vegetation exhibits high reflectance in the near-infrared region and, conversely, absorbs a significant amount of light in the red region (Rouse et al., 1973). SAVI is a vegetation index also very used for vegetation studies, it helps mitigate the effects of soil interference and accounts for areas with less dense vegetation (Huete, 1988). Both NDVI and SAVI play a crucial role in discerning and characterizing vegetation types, particularly in the context of forest species, enhancing the accuracy of classification outcomes. In image classification, there is redundant information due to the high correlation between bands. Principal Component Analysis (PCA) allows reduction of information eliminating the strong correlation between bands and generate new variables, known as principal components (Thenkabail, 2016).

In object-based image classification, the image segmentation process, where the minimum unit is the object, aims to group pixels considering their spectral, spatial and dimension characteristics (Ma et al., 2017) that represent different objects on the surface. Three adjustment parameters were considered: spectral and spatial detail, and the minimum segment size. These parameters allow for the grouping of adjacent pixels and identification of segments based in each input dataset. Spectral detail allows for object segmentation based on color characteristics. A small spectral detail value results in fewer segments, each covering a larger area. Spatial detail enables object segmentation based on object proximity. Lower values allow for finer segmentation between closely grouped objects. The minimum segment size parameter specifies the minimum size, in pixels, for a group of contiguous pixels. All segments with a pixel count less than the specified minimum are merged with the nearest neighboring segment for a better fit. While there isn't a specific interval for this parameter, smaller values result in less homogeneous segments (Visalli et al., 2021). The adjustment of these parameters is tailored to the specific study area's characteristics and the desired level of segmentation accuracy.

The objective of this study is mapping the tree canopies for the dominant forest species using two WorldView-3 images. Additionally, the study aims to evaluate the performance of machine learning classifiers such as Maximum Likelihood (MLH), Support Vector Machine (SVM), and Random Forest (RF) for cartographic production of forest species in two regions of Alentejo, southern Portugal. In the northern region of Alentejo (39° 5' 15.16"N; 8° 1' 13.35"W), cork oak (Quercus suber) dominates, while in the southern region (38° 3' 5.71"N; 7°38' 43.54"W), holm oak (Quercus rotundifolia) is dominant.

Material and Methods

2.1. Study Area

The study areas correspond to two satellite images of 25 km² each, located in the Alentejo region (Figure 1), southern Portugal. The region is characterized by a Mediterranean climate, where summers are hot and dry, and winters are cold and humid, with greater temperature ranges and intensity in land. The terrain is marked by plains, with an average elevation of approximately 200 m. Image a) is predominantly composed of cork oak (Quercus suber) and image b) of holm oak (Quercus rotundifolia).

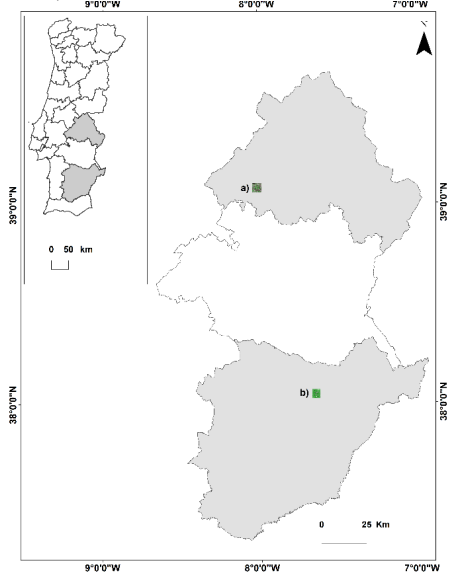


Figure 19. Study area location (a) dominance of cork oak; b) dominance of holm oak.

2.2. Remote sensing data

The two images from the WorldView-3 satellite used in this study were acquired in: a) September 21, 2020, and b) June 14, 2020. Both images have four bands corresponding to wavelengths in the electromagnetic spectrum's blue (B) (491.9 nm), green (G) (541.1 nm), red (R) (660.1 nm), and near-infrared (NIR) (824.0 nm) regions, and a spatial resolution of 0.50 m, which result from the fusion process with the panchromatic band.

Multiresolution segmentation and object oriented classification were used to attain the forest cover maps. The explanatory variables were the four bands, two vegetation indices (NDVI and SAVI) and principal components of the for bands (Figure 2). Multiresolution segmentation was tested with four data sets of explanatory variables: i) near-infrared band, NDVI e second principal component (NIR, NDVI, CP2); ii) near infrared band, first and second principal component (NIR, CP1, CP2); iii) NDVI, SAVI and second principal component (NDVI, SAVI, CP2); iv) red and NIR bands e and NDVI (R, NIR, NDVI). Three parameters were considered in multispectral segmentation spectral detail, spatial detail and minimum segment distance.

2.3. Methodology

In the Figure 2 is presented a flowchart with the main steps of the methodology, from the original bands until the resulted LULC forest maps.

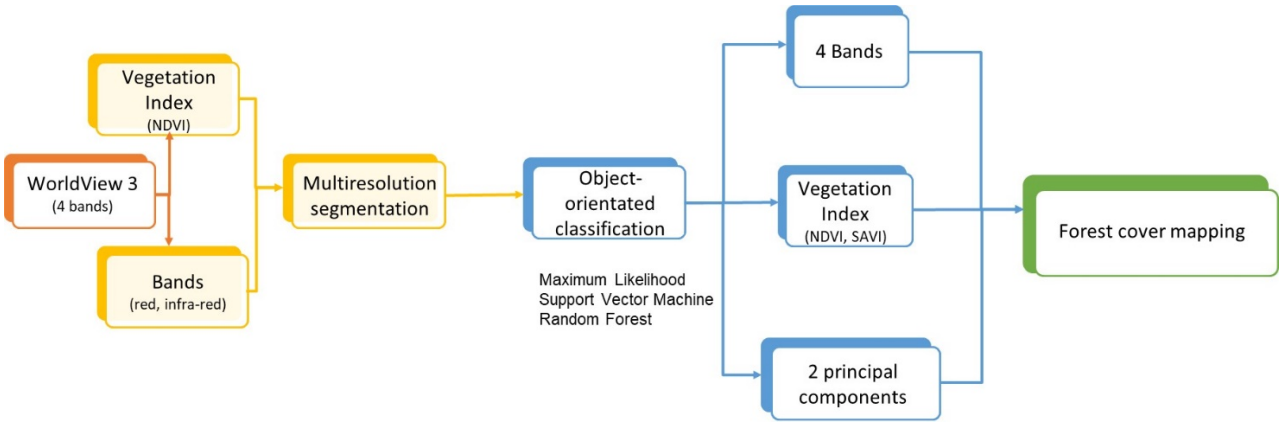


Figure 20. Methodology flowchart

For the spectral and spatial detail parameters (both of which can vary between 1 and 20), a higher value was considered for the former to allow for spectral differentiation of various land use/land cover classes. For the latter, a medium value was chosen to obtain objects of small dimensions, such as isolated canopies, or larger objects when contiguous canopies are present. In the segmentation process, the values that best suited the isolation of tree canopies, considering spectral and spatial detail, and the minimum segment size were determined as follows: for the cork oak-dominated image, the values were 19, 10, and 100 respectively; for the holm oak-dominated image, the values were 18, 15, and 100 respectively. In sparse forest areas, it was possible to obtain objects corresponding to individual tree crowns. Whereas in dense forest areas where the crowns of the individual trees touch each other, isolating individual crowns was not possible, instead segmentation resulted in clusters of crowns.

With the segmented image, the next step was the object-based image classification process. Three machine learning algorithms were tested: Maximum Likelihood (MLH), Support Vector Machine (SVM), and Random Forest (RF). Initially, classes are assigned to the identified objects, acting as training data for the classifiers tested. In the cork oak-dominated image, nine classes were considered (cork oak, holm oak, stone pine, eucalyptus, water, soil, urban area, agriculture, and shadow). In the holm oak-dominated image, seven classes were considered (holm oak, eucalyptus, water, soil, urban area, agriculture, and shadow). The data set for the classification included as explanatory variables the four original bands, two vegetation indices (NDVI and SAVI), the first and second principal components (PC1 and PC2) and the objects that resulted from the multiresolution segmentation. For RF were used 500 decision trees and 20 knots, which according to several authors (Breiman, 2001; Ienco et al., 2019; Karasiak et al., 2017; Pageot et al., 2020) are suitable to reach good accuracies. For SVM, the maximum number of samples was used to define each class was 0, all samples are used in the training of the classifier.

To assess the accuracy of the four land use/land cover maps resulting from the object-based image classification, a random sampling of 50 points was done for each class (8 LULC classes), for a total of 450 points for image a) and 350 for image b). Each point was assigned a LULC class based on visual analysis with the help of true-color and false-color composite imagery and base map provided in ARCGIS and Google Earth. This information was compared with the

classification results from an accuracy assessment using a confusion matrix, Kappa coefficient (Eq. 1), and overall accuracy (Eq. 2). The confusion matrix (Congalton et al., 1983, p. 1673; Stehman, 1997, p.1221) displays the number of pixels correctly classified against the number of pixels predicted for each class during classification, thus assessing the degree of agreement with reality. The information in the rows represents user accuracy (omission errors), while the columns represent producer accuracy (commission errors). A commission error occurs when pixels are included in an incorrect class, while an omission error occurs when pixels are excluded from the class to which they belong (Fonseca & Fernandes, 2004).

The confusion matrix enables the calculation of the Kappa coefficient (Kappa) and overall accuracy (OA).

$$Kappa = \frac{N\sum_{i=1}^{r} x_{ii} - \sum_{i=1}^{r}(x_{i+}x_{+i})}{N^2 - \sum_{i=1}^{r}(x_{i+}.x_{+i})} \times 100 \quad \text{Eq. 1}$$

$$OA = \sum_{i=1}^{k} \frac{N_{ii}}{N} \quad \text{Eq. 2}$$

For the processes calculation of the vegetation indices, principal component analysis, multiresolution segmentation, and object-based image classification, Geographic Information System (GIS) tools were employed using the ARCGIS software version 10.8.

Results and Discussion

In the multiresolution segmentation, the input data set that yielded the best result was the one of red and near-infrared bands, and NDVI, with a Kappa coefficient of 67% and an overall accuracy of 79%. According to Landis & Koch (1977), Kappa coefficients ranging between 41-60% correspond to good classifications and from 61-80% very good. In this study variable combinations including NIR, CP1, CP2, and NIR, NDVI, CP2 are good, while those including NDVI, SAVI, CP2, and R, NIR, NDVI are very good (Table 1). The accuracy improved with the inclusion of vegetation indices (NDVI and SAVI). However, the most significant contribution to accuracy came from the original red and near-infrared bands in combination with NDVI. It was also observed that the principal components did not contribute significantly to increase accuracy

Table 9. Multiresolution segmentation accuracy.

N	Variables	Segmentation			Classification	
		Spectral detail	Spatial detail	Segment size	*Kappa*	Overall accuracy
1	NIR, CP1, CP2	19	10	100	54%	69%
2	NIR, NDVI, CP2				58%	73%
3	NDVI, SAVI, CP2				64%	77%
4	R, NIR, NDVI				67%	79%

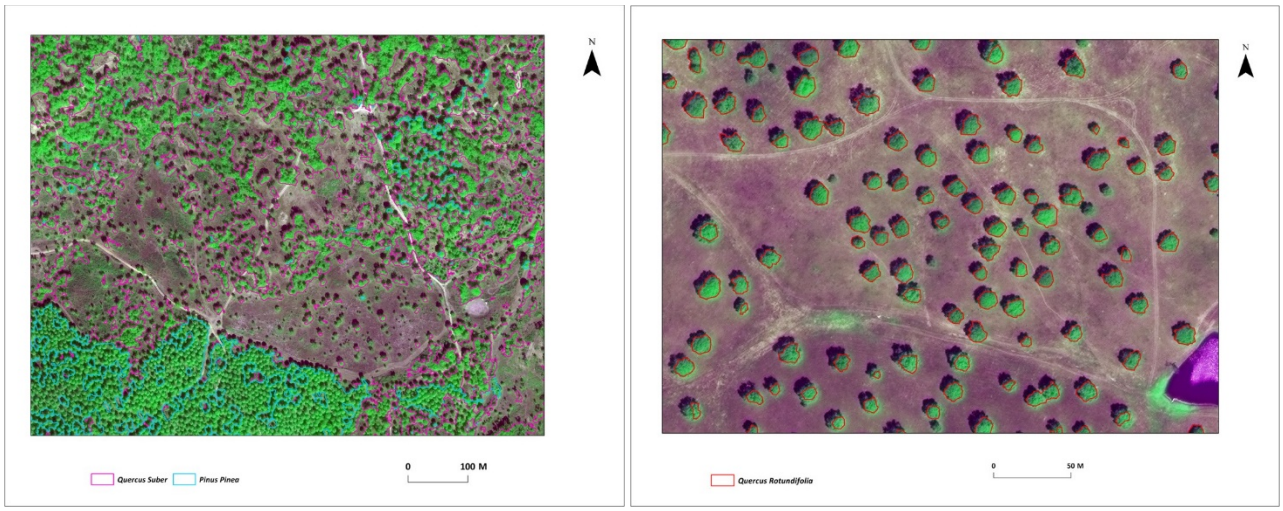
The multiresolution segmentation was used to object-based image classification, with SVM attaining the best accuracy when compared with RF and MLH in both images. The accuracy was higher for image b) composed mainly of holm oak then for image a) with predominance of cork oak, due to the its smaller variability in land cover.

A similar study by Sousa et al. (2010) using high-resolution imagery from the Quickbird satellite, with three LULC classes had high accuracy (91%). This discrepancy can be attributed to the greater complexity of our study area, characterized by high spectral variation due to the presence of multiple LULC classes and spatial variation due to object size. As in this study, errors along the tree canopy edges were observed, possible due to the irregular characteristics of each forest species' canopies, stemming from their shapes, textures, and lighting conditions.

These results are in agreement with those found by Volke & Abarca-Del-Rio (2020), which demonstrated that the Support Vector Machine algorithm is efficient for this classification method, even when using LANDSAT and ASTER satellite imagery, achieving an accuracy of 97% for LULC mapping. The segmented images, with defined objects corresponding to isolated or aggregated tree canopies, can be observed in the figures 3, illustrating the two forest species areas in question in this study.

Table 10. Kappa coefficient and overall accuracy for both images and for the three machine learning algorithms.

Image	Variables	Segmentation			Classifiers	Classification	
		Spectral detail	Spatial detail	Segment size		*kappa*	Overall accuracy
a)	R, NIR, NDVI	19	10	100	MLH	84%	92%
					SVM	90%	95%
					RF	89%	94%
b)		18	15	100	MLH	66%	85%
					SVM	90%	96%
					RF	90%	95%



a) b)

Figure 3. Forest classification, a) *quercus suber* dominance and b) *quercus rotundifolia* dominance.

Conclusions

In this study, it was identified and delineated the tree crowns of forest species present in the image through segmentation and object-based image classification tools in the ARCGIS v10.8 software. The segmentation process, of high-resolution images, allows for detailed adjustments to surface objects.

Object-based image classification using WorldView-3 satellite imagery produced good results in identifying forest species crowns. The Support Vector Machine classifier, with the original bands and vegetation indices, in particular NDVI, led to land use/land cover maps with an overall accuracy of 95% and 96% and Kappa coefficients of 90%. The reflectance values for each class obtained by selecting objects to train the classifier discriminated very well between most of the classes identified. The image's acquisition date, corresponding to the dry season, enhanced a greater contrast between forest species and other land use classes, reducing the confusion with shrub vegetation and soil. However, some errors still occurred due to spectral similarity between certain forest species (e.g., cork oak vs. holm oak), crown characteristics, and shadow presence near canopies.

Funding

This work is funded by National Funds through FCT - Foundation for Science and Technology under the Project ***UIDB/05183/2020***

Acknowledgements

The work was supported by Programa Operativo de Cooperação Transfronteiriço Espanha-Portugal (POCTEP); CILIFO project with grant number 0753_CILIFO_5_E. The authors are grateful to Teresa Pinto Correia for the valuable suggestions and comments. The authors are grateful to two anonymous reviewers for their comments.

References

Ahmad, A., Gilani, H., & Ahmad, S. R. (2021). Forest aboveground biomass estimation and mapping through high-resolution optical satellite imagery—A literature review. Forests, 12(7), 914.

Alrassi, F., Salim, E., Nina, A., Alwi, L., Danoedoro, P., & Kamal, M. (2016, November). GEOBIA For Land Use Mapping Using Worldview2 Image In Bengkak Village Coastal, Banyuwangi Regency, East Java. In IOP Conference Series: Earth and Environmental Science (Vol. 47, No. 1, p. 012009). IOP Publishing.

Baatz, M., Schäpe, M., 2000. Multiresolution segmentation — na optimization approach for high quality multi-scale image segmentation. In: Strobl, J., Blaschke, T.,Griesebner, G. (Eds.), Angewandte Geographische Informations-Verarbeitung XII. Wichmann Verlag, Karlsruhe, pp. 12–23.

Belward, A. S., & Skøien, J. O. (2015). Who launched what, when and why; trends in global land-cover observation capacity from civilian earth observation satellites. ISPRS Journal of Photogrammetry and Remote Sensing, 103, 115-128.

Blaschke, T. (2010). Object based image analysis for remote sensing. ISPRS journal of photogrammetry remote sensing, 65(1), 2-16.

Blaschke, T., Hay, G. J., Kelly, M., Lang, S., Hofmann, P., Addink, E., ... & Tiede, D. (2014). Geographic object-based image analysis–towards a new paradigm. ISPRS journal of photogrammetry and remote sensing, 87, 180-191.

Chen, L., Wang, Y., Ren, C., Zhang, B., & Wang, Z. (2019). Assessment of Multi-Wavelength SAR and Multispectral Instrument Data for Forest Aboveground Biomass Mapping Using Random Forest Kriging. Forest Ecology and Management, 447, 12–25. https://doi.org/10.1016/j.foreco.2019.05.057.

Congalton, R. G., Oderwald, R. G., & Mead, R. A. (1983). Assessing Landsat classification accuracy using discrete multivariate analysis statistical techniques. Photogrammetric engineering and remote sensing, 49(12), 1671-1678.

Dorren, L. K. A., Maier, B. and Seijmonsbergen, A. C. 2003. Improved Landsat-based forest mapping in steep mountainous terrain using object-based classification. Forest Ecology and Management 183: 31– 4.

Fonseca, A. D., & Fernandes, J. C. (2004). Deteção Remota (LIDEL).

Galidaki, G., Zianis, D., Gitas, I., Radoglou, K., Karathanassi, V., Tsakiri–Strati, M., ... & Mallinis, G. (2017). Vegetation biomass estimation with remote sensing: focus on forest and other wooded land over the Mediterranean ecosystem. International journal of remote sensing, 38(7), 1940-1966.

Gibril, M. B. A., Shafri, H. Z., & Hamedianfar, A. (2017). New semi-automated mapping of asbestos cement roofs using rule-based object-based image analysis and Taguchi optimization technique from WorldView-2 images. International Journal of Remote Sensing, 38(2), 467-491.

Gonçalves, A.C., Sousa, A.M.O., Mesquita, P., 2019. Functions for aboveground biomass estimation derived from satellite images data in Mediterranean agroforestry systems. Agroforestry Systems 93, 1485–1500. https://doi.org/10.1007/s10457-018-0252-4

Gonçalves, A.C., Sousa, A.M.O., Mesquita, P.G., 2017. Estimation and dynamics of above ground biomass with very high resolution satellite images in Pinus pinaster stands. Biomass and Bioenergy 106, 146–154. https://doi.org/10.1016/j.biombioe.2017.08.026

Gudmann, A., Csikós, N., Szilassi, P., & Mucsi, L. (2020). Improvement in Satellite Image-Based Land Cover Classification with Landscape Metrics. Remote Sensing, 12(3580), 1–19. https://doi.org/10.3390/rs12213580.

Hao, S., Cui, Y., & Wang, J. (2021). Segmentation Scale Effect Analysis in the Object-Oriented Method of High-Spatial-Resolution Image Classification. Sensors, 21(23), 7935.

Huete, A. R. (1988). A soil-adjusted vegetation index (SAVI). Remote sensing of environment, 25(3), 295-309.

ICNF.(2015). Inventário Nacional 6.

Isbaex, C., & Coelho, A. (2019). The Potential of Sentinel-2 Satellite Images for Land-Cover/Land-Use and Forest Biomass Estimation: A Review. Em Forest Biomass - From Trees to Energy. https://doi.org/10.5772/intechopen.93363.

Isnaen, Z., Utari, D., Ramadhan, A. F., Putri, R. C., Kusuma, D. W., & Kamal, M. (2019, July). Comparison of Mangrove and Other Objects Spectral Reflectance from Small Format Aerial Photography Image, WorldView-2 Image, and Field Measurement. In 2019 5th International Conference on Science and Technology (ICST) (Vol. 1, pp. 1-5). IEEE.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. biometrics, 159-174.

Ma, L., Li, M., Ma, X., Cheng, L., Du, P., & Liu, Y. (2017). A review of supervised object-based land-cover image classification. ISPRS Journal of Photogrammetry and Remote Sensing, 130, 277-293.

Mallinis, G., Koutsias, N., Tsakiri-Strati, M., & Karteris, M. (2008). Object-based classification using Quickbird imagery for delineating forest vegetation polygons in a Mediterranean test site. ISPRS Journal of Photogrammetry and Remote Sensing, 63(2), 237-250.

Marcussi, A. B., Bueno, C. R. P., Miqueloni, D. P., & Arraes, C. L. (2010). Utilização de Índices de Vegetação para os Sistemas de Informação Geográfica. Caminhos de Geografia, 11(35), 41–53.

Meng, Z., & Xiao, B. (2011, July). High-resolution satellite image classification and segmentation using Laplacian graph energy. In 2011 IEEE International Geoscience and Remote Sensing Symposium (pp. 605-608). IEEE.

Meyer, L. H., Heurich, M., Beudert, B., Premier, J., & Pflugmacher, D. (2019). Comparison of Landsat-8 and Sentinel-2 data for Estimation of Leaf Area Index in Temperate Forests. Remote Sensing, 11(10), 1–6. https://doi.org/10.3390/rs11101160.

Rouse Jr, J. W., Haas, R. H., Schell, J. A., & Deering, D. W. (1973). Monitoring the vernal advancement and retrogradation (green wave effect) of natural vegetation (No. NASA-CR-132982).

Sousa, A., Mesquita, P., Gonçalves, A. C., & Marques da Silva, J. R. (2010). Segmentação e classificação de tipologias florestais a partir de imagens Quickbird. IX Seminário de Atualização em Sensoriamento Remoto e Sistemas de Informações Geográficas Aplicados à Engenharia Florestal, Curitiba, Paraná (Brasil), 19-21.

Sousa, A.M.O., Gonçalves, A.C., Mesquita, P., Marques da Silva, J.R., 2015. Biomass estimation with high resolution satellite images: A case study of Quercus rotundifolia. ISPRS Journal of Photogrammetry and Remote Sensing 101, 69–79. https://doi.org/10.1016/j.isprsjprs.2014.12.004

Stehman, S. (1996). Estimating the kappa coefficient and its variance under stratified random sampling. Photogrammetric Engineering and Remote Sensing, 62(4), 401-407.

Thenkabail, P. S. (2016). Remotely Sensed Data Characterization, Classification, and Accuracies. Em L. Taylor & Francis Group (Ed.), Remote Sensing Handbook (Vol. 1). https://doi.org/10.1201/b19294.

Varin, M., Chalghaf, B., & Joanisse, G. (2020). Object-based approach using very high spatial resolution 16-band WorldView-3 and LiDAR data for tree species classification in a broadleaf Forest in Quebec, Canada. Remote Sensing, 12(18), 3092.

Vibhute, A. D., Dhumal, R. K., Nagne, A. D., Rajendra, Y. D., Kale, K. V., & Mehrotra, S. C. (2016). Analysis, classification, and estimation of pattern for land of Aurangabad region using high-resolution satellite image. In Proceedings of the Second International Conference on Computer and Communication Technologies: IC3T 2015, Volume 2 (pp. 413-427). Springer India.

Visalli, R., Ortolano, G., Godard, G., & Cirrincione, R. (2021). Micro-Fabric Analyzer (MFA): A new semiautomated ArcGIS-based edge detector for quantitative microstructural analysis of rock thin-sections. ISPRS International Journal of Geo-Information, 10(2), 51.