Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 159 (2019) 650–657

Procedia Computer Science

www.elsevier.com/locate/procedia

23rd International Conference on Knowledge-Based and Intelligent Information & Engineering Systems

Fully automatic multi-temporal land cover classification using Sentinel-2 image data

Sergio Baamonde[a,b,*], Martiño Cabana[c], Neftalí Sillero[d], Manuel G. Penedo[a,b], Horacio Naveira[c], Jorge Novo[a,b]

[a]Department of Computer Science, University of A Coruña, 15071 A Coruña, Spain
[b]CITIC-Research Center of Information and Communication Technologies, University of A Coruña, 15071 A Coruña, Spain
[c]Department of Biology, Faculty of Sciences, University of A Coruña, A Coruña, Spain
[d]CICGE, Centro de Investigação em Ciências Geo–Espaciais, Faculdade de Ciências da Universidade do Porto, Portugal

Abstract

The analysis of remote sensing images represents a highly important issue to be performed in many relevant fields such as climate change studies or land cover mapping. Traditional proposals usually identify the land cover classes from general related groups such as different tree species or different crop varieties. Additionally, these proposals commonly use information from a precise time span or season, not accounting for the variability of the data over the entire year, specially in regions with several seasons.

In this work, we propose a multi-temporal classification system to identify and represent diverse land cover classes over any period of the entire year by using Sentinel-2 satellite image data. To this end, 526 representative samples were labelled from 5 complex and variable different land cover types over the Special Area of Conservation (SAC) Betanzos–Mandeo in the northwest of the Iberian Peninsula. The method achieves a satisfactory mean accuracy value of 84.0% for the testing set using the best configuration with a radial Support Vector Machine classifier. This system will be used in the study of the population connectivity of two threatened herptiles, but it can be easily extended to other species of interest in the future.

Keywords: Remote sensing; Sentinel–2; Land cover classification; Machine learning

1. Introduction

Remote sensing (RS) satellite images provide highly relevant information for multiple applied domains such as analysis procedures related with climate change, agriculture, environmental monitoring or habitat conservation [23]. RS images cover wide geographic regions, allowing the specialists to obtain multi-spatial and multi-temporal infor-

* Sergio Baamonde
E-mail address: sergio.baamonde@udc.es

10.1016/j.procs.2019.09.220

mation of any region over the earth's surface. Currently, a wide array of RS satellite image products are available both from commercial and non-commercial sources [3].

The Sentinel-2 satellite constellation is an Earth observation mission developed by the European Spatial Agency (ESA) under the Copernicus programme with the aim of providing accurate and easily accessible information about many characteristics of the earth's surface. The higher resolution that is achieved for the visible light spectrum products and the free, full and open data policy of the Copernicus programme place nowadays the Sentinel-2 image products among the most used RS solutions worldwide. Two satellites (A and B, in orbit since 2015 and 2017, respectively) provide high-resolution multi-spectral images across 13 different spectral bands (413–2190 nm) with spatial resolutions of 10m (visible light spectrum and near-infrared bands), 20m (6 red-edge and shortwave-infrared bands) and 60m (3 atmospheric correction bands) with a high revisit frequency of 5 days at the Equator. Given the significant potential of the availability of all this information, the Sentinel-2 products have been exploited in a wide array of applications, such as aboveground biomass estimation [4], fire severity estimation [17], vegetation phenology [22], carbon prediction [2] or land cover mapping [7].

Land cover change is regarded as one of the most important components of global change that also affects the biological diversity, the climate change and the terrestrial ecosystems, among others [21]. Projects such as the CORINE (Coordination of Information on the Environment) Land Cover (CLC) [1] aim at creating a land cover structure of the state of the European surface. However, these projects typically use traditional mapping techniques that require manual or semi-manual work to classify the land surface, implying a long and tedious work that makes infeasible the labelling of significant amounts of temporal image sequences. For that reason, given the lack of sufficient processed images, these projects commonly suffer from poor temporal resolution, without being capable of capturing, analysing and representing the rapid changes of the land cover. Additionally, land cover data are not easily available to the general public and often need a high storage capacity due to the considerable amount of included data in the available products. These factors contribute to the recognition of remote sensing data as "Big Data" [13, 15, 8] that need to be automatically processed and classified in order to obtain an exploitable product within reasonable accuracy values [6].

Traditional and novel machine learning approximations have been successfully applied in specific land cover classification tasks [20]. Thus, Immitzer *et al.*[11] used a traditional Random Forest (RF) classifier to identify crop and tree species classifications in Central Europe with Sentinel-2 data with corresponding overall cross-validated accuracies of 65% and 76%. Pesaresi *et al.*[16] developed a novel classification method called Symbolic Machine Learning (SML) in a more stable and constant domain, in particular in a detailed urban land cover mapping with a balanced accuracy of 81%. A more novel approach using a deep learning technique was developed by Kussul *et al.*[12] for the classification of crops in Ukraine, achieving an accuracy over 85% using an ensemble of convolutional neural networks (CNN). Similarly, Helber *et al.*[10] used state-of-the-art CNN models (GoogLeNet [19] and ResNet-50 [9]) to validate a proposed novel dataset based on Sentinel-2 images associated with cities across Europe, achieving an overall classification accuracy of 98.57%.

In this work, we propose an automatic method to identify multiple heterogeneous land cover classes using Sentinel-2 multi-temporal satellite image data in a temperate and variable landscape of the northwest of the Iberian Peninsula. In particular, this system is planned to be exploited in the study of the effect that these habitats can produce on the population connectivity of two threatened herpetofauna species: the golden–stripped salamander, *Chioglossa lusitanica*, and the Iberian rock lizard, *Iberolacerta monticola*. Specifically, we retrieve information from the Special Area of Conservation (SAC) Betanzos–Mandeo over the Mandeo and Mendo rivers as well as the Betanzos estuary, being exploited for the identification and classification of a variety of land classes: autochthonous forest, eucalyptus plantations, scrubland, farmland and infrastructures. With the retrieved information, different representative classifiers were adjusted and studied to measure the suitability of the designed system.

This paper is organized in the following sections: Section 2 explains the characteristics of the proposed methodology. Section 3 shows the results of the designed validation experiments. Finally, Section 4 comments on the obtained results and proposes related future work.

2. Methodology

The general scheme of the proposed methodology is illustrated in Figure 1. Firstly, to capture the representative variability of the complex landscape that is analysed in this work, a yearly input dataset is selected and downloaded from the available data sources and prepared for the posterior analysis. Then, the obtained Sentinel-2 images are processed in order to be used by the classification models. Finally, we verify the quality of the optimal classification models and measure their suitability and performance.

Fig. 1: General scheme of the proposed methodology.

2.1. Input dataset generation

This subsection describes the batch process that queries and downloads the required Sentinel-2 multi-temporal datasets over the SAC Betanzos–Mandeo Area of Interest (AOI) for its analysis, as the Figure 2 illustrates. These datasets are retrieved and prepared as the input information to be exploited by the proposed system.

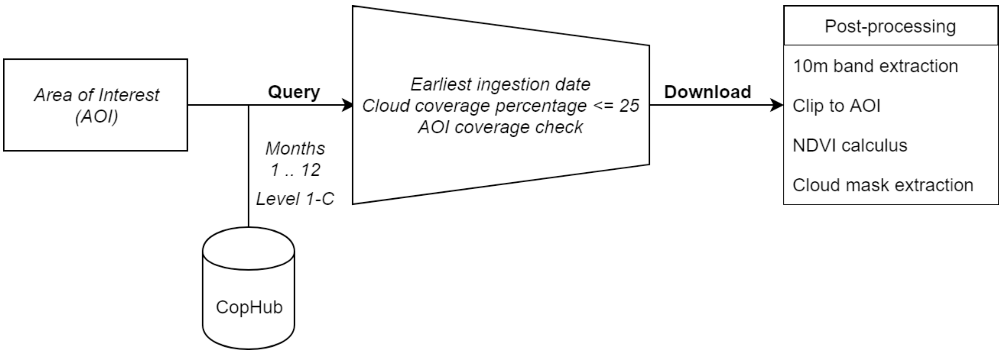


Fig. 2: Dataset generation process.

In the first place, we performed a sequential search in the Copernicus Open Access Hub (CopHub) for each month over a studied target year using the available R package getSpatialData [18]. This package provides an easy access to the CopHub data as well as useful tools to specify and restrict the performed queries.

Given the significant variability of the analysed land categories (crops, scrubs, forest, etc.) along the different seasons of the year in the Betanzos–Mandeo SAC, as said, a representative amount of information of all the months of a natural year is retrieved to capture all the possible situations. Additionally, the retrieved information of the Sentinel-2 satellite may be altered by the presence of clouds, introducing erroneous data that may confuse the system. Thus, in order to obtain the cloud mask information and avoid these situations, we selected only the Sentinel-2 products with processing level 1-C which include cloud mask files in the downloaded products. After each query, we filtered the results in order to select the Sentinel-2 product with the earliest ingestion date that had less or equal cloud coverage percentage than 25%. Before making the final selection, an additional AOI coverage check was performed to account for partial matches between the covered product area and the defined AOI, discarding the selected product in that case and performing the check on the next one.

The identified information is post-processed and prepared for its analysis and classification. In particular, the post-processing stages are represented in Figure 3[1]. We downloaded the selected data packages, extracting the 10m resolution bands (B2, B3, B4 and B8) that contain information about the visible light spectrum (Figure 3a) and the Near Infra-Red (NIR) band. The three classical RGB bands (R_{band}, G_{band} and B_{band}) are portrayed by the B4, B3 and

[1] Contains modified Copernicus Sentinel data [2018].

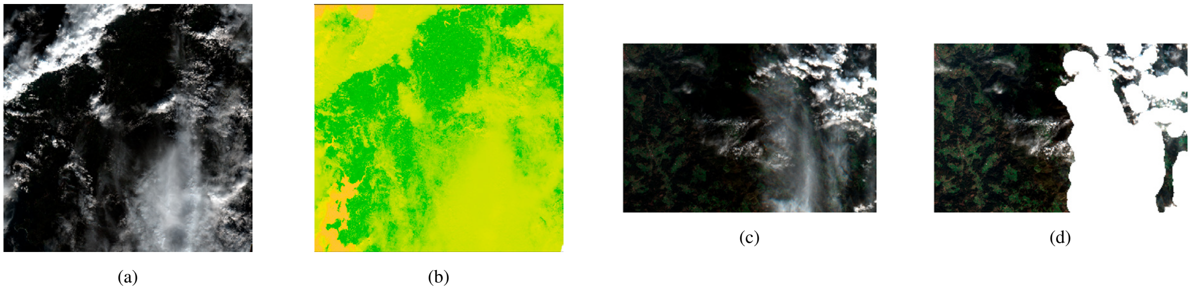


Fig. 3: Post-processing steps of the input dataset preparation. (a) represents the complete granule region. (b) shows a visual reconstruction of the NDVI values for the same region. (c) shows the AOI sub-region of the complete product. (d) represents the cloud mask points as white regions over the AOI.

B2 bands, whereas the NIR band is represented by the B8 band. These bands are used to calculate the Normalized Difference Vegetation Index (NDVI) (Equation 1) for the studied region, represented in Figure 3b. The NDVI index generally represents the difference between the reflectance of the NIR band and the red band from the visible spectrum, being previously used in numerous land cover applications [5, 14] due to its suitability to identify phenological differences among vegetation types. Finally, we clip every band to the AOI extent (Figure 3c) and checked if any point of the cloud mask is contained on the AOI (Figure 3d), extracting the cloud mask file for its future use in that case.

$$NDVI = \frac{NIR - R_{band}}{NIR + R_{band}} \tag{1}$$

2.2. Target land categories

We defined five classes of habitat that we consider to be the most important to determine the potential distribution of the two analysed species as well as other threatened species potentially studied in the future: autochthonous forest, eucalyptus plantations, scrubland, farmland and infrastructures. In this case, the native forest is the habitat most used by both species and is located mainly in the surroundings of the riverbanks of the Mandeo and Mendo rivers. Eucalyptus plantations are among the most relevant threats to the conservation of the species studied in this area due to the drying out of the environment and the destruction of natural habitats. We would like to highlight the heterogeneity and variability of the faced classes, with particular variations and visualizations of each case, like the deciduous autochthonous forest or the farmland as reference, along the natural year, different from the rest.

2.3. Training and classification

Once the complete dataset has been processed and prepared, coherently representing the different months of the year, we labelled representative points of each class, extracting the information of each point across the complete set of the prepared raster files, restricted to the points that are not contained in the cloud regions defined by the cloud mask. Samples located in cloud regions have different spectral values than the rest of the samples and could confuse the system and, therefore, decrease the quality of the classification models.

The resulting dataset was used to train and test representative classifiers as: Random Forest (RF), Radial Support Vector Machine (SVM) and k-Nearest Neighbours (kNN) classifiers. As previously mentioned, these classifiers have been proved suitable in different RS image analysis tasks [20].

3. Results

The methodology has been validated using representative samples that were obtained from the previously mentioned region of the northwest of the Iberian Peninsula, the SAC Betanzos–Mandeo. We analyse this region due to its relevance in the study of the population connectivity of threatened herpetofauna species by the posterior exploitation from the analysis of the aforementioned land cover classes: autochthonous forest, eucalyptus plantations, scrubland, farmland and infrastructures.

The Sentinel-2 1C products provide 100 by 100 kilometre squared ortho-images in UTM/WGS84, with one tile per spectral band. Additionally, the studied spectral bands (B2, B3, B4 and B8) achieve a spatial resolution of 10 metres, resulting in a total of 2,160,000 data instances for the complete region. To account for memory, processor and storage restrictions, the number of selected samples is considerably smaller than the number of available samples. For this proposal, 526 instances were labelled, resulting in 6,038 data points across the entire year. The maximum number of available samples is 6,312 but, as previously mentioned, instances located in cloud points are ignored in order to avoid erroneous data and improve the accuracy of the classifiers. The training set is comprised of 4,229 samples, whereas the remaining 1,809 samples are used to validate the trained models. To avoid the overfitting of the classifiers to the training set, 5 repetitions of a 10-fold cross-validation process were performed on the training set to obtain the optimal hyperparameters of each classifier. Figure 4 presents the confusion matrix using each tested classifier as well as the mean accuracy and Kappa measurements. In particular, the confusion matrix details the results for all the analysed categories. As we can see, the radial SVM classifier achieves a higher mean accuracy value (84.0%) than the RF (82.7%) and KNN (79.6%) classifiers. We consider that these results represent a satisfactory performance given the diversity of analysed classes that present specific variabilities in each case over the seasons of the year, as well as a significant similarity between some pairs of them. Across all the classifiers, the scrubland class has lower accuracy values than the other classes, partially caused by the difficulty of aerial differentiation of scrubs from other arboreal structures. The infrastructure class also achieves similar high accuracy values across all the classifiers, due to the easier differentiation of buildings and other structures from natural landmarks.

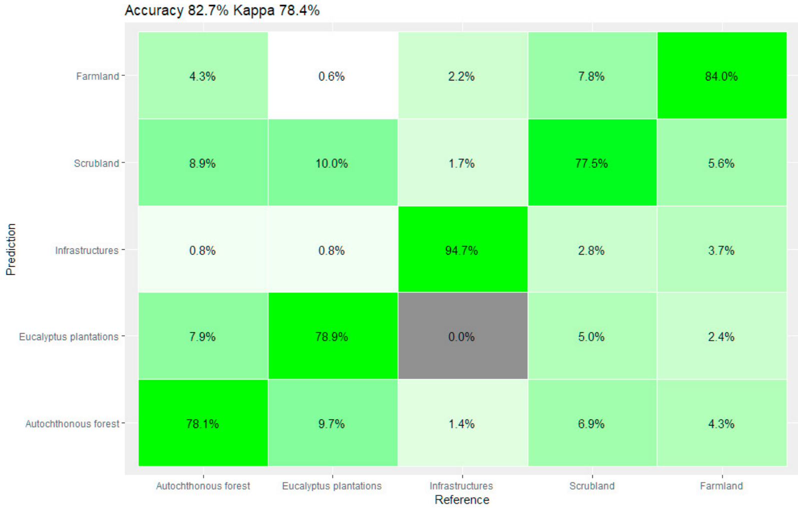
As illustration, representative classification results are presented in Figure 5[2]. As we can see from the examples, the method is capable of identifying representative landmarks and structures, such as farms, forests or highways. In particular, the optimal classifier predicts the terrain class of each point of a Sentinel-2 satellite image (Figure 5, 1st row) using the 10m resolution bands. The resulting image (Figure 5, 2nd row) shows the assigned class to each pixel in a visual and intuitive way for a direct inspection of the specialist. Combining the prediction with the original image (Figure 5, 3rd row) further illustrates the similarity between the classification and the terrain that is depicted in the original image.

4. Discussion and conclusions

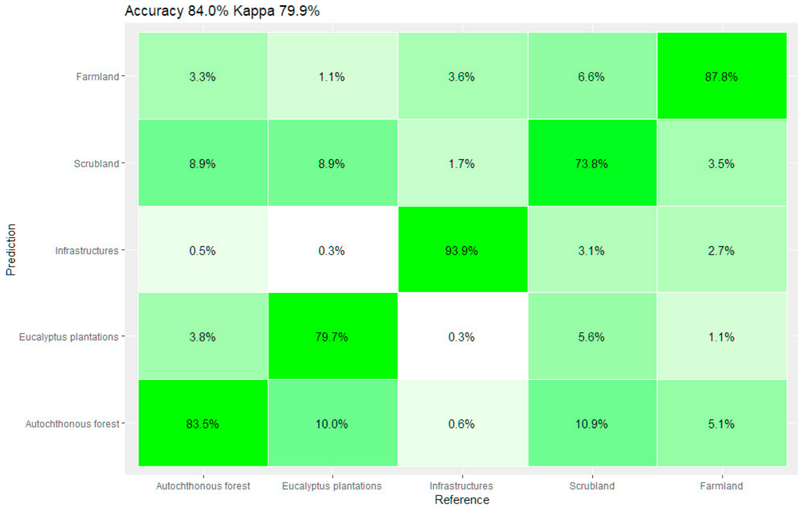
Land cover identification and classification is a highly relevant issue in multiple prominent applications, such as their use in the analysis of climate change, risk prediction or landscape connectivity and habitat suitability studies. Usually, the identification techniques rely on precise training sets that are temporally situated during a small period or specific season, limiting the classification tasks to products that are obtained during those restricted periods.

In this work, we present an automatic classification procedure to predict land cover uses from multi-temporal Sentinel-2 satellite image data. This proposal addresses the challenging topic of classifying multi-temporal images from any season of a natural year due to the visual variability of the different land cover classes across the seasons. Additionally, this method also identifies and segments multiple diverse land cover classes, not being limited to, for example, different species of trees (from the total tree surface) or different crop types (from the total farmland surface). The proposed complete and heterogeneous set of land cover classes (autochthonous forest, eucalyptus plantations, scrubland, farmland and infrastructures) takes into account a highly different land evolution with different characteristics that can be exploited to identify those land cover classes. In regard to the classification task, the proposed method achieves a mean test accuracy of 84.0% when using a radial SVM classifier, with a kappa value of 79.9%. This high capacity of identification allows the specialists to adequately determine the distribution of the different types of habitats in the studied region. Posteriorly, the system is planned to be exploited in the analysis of the impact of those habitats in two particular threatened herpetofauna species, with the potential of the future extension to the analysis of others of interest.

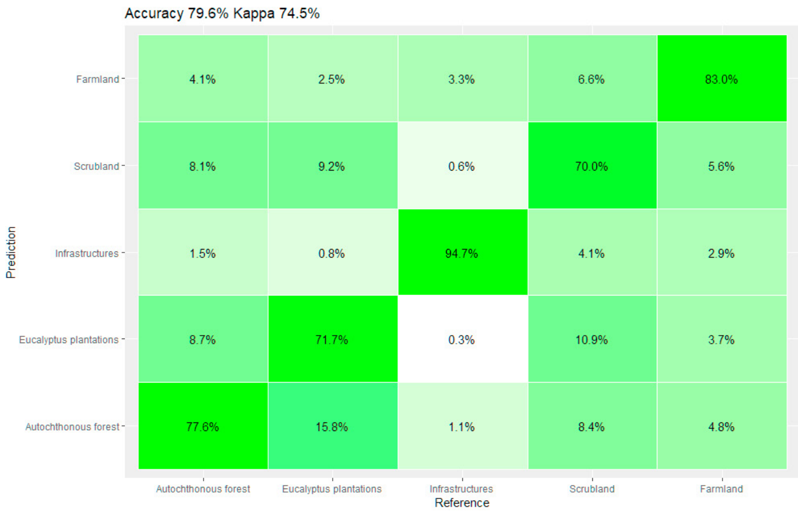
[2] Contains modified Copernicus Sentinel data [2018].



(a)



(b)



(c)

Fig. 4: Confusion matrix results for the studied representative classifiers. (a) Confusion matrix of the RF classifier. (b) Confusion matrix of the Radial SVM classifier. (c) Confusion matrix of the KNN classifier.

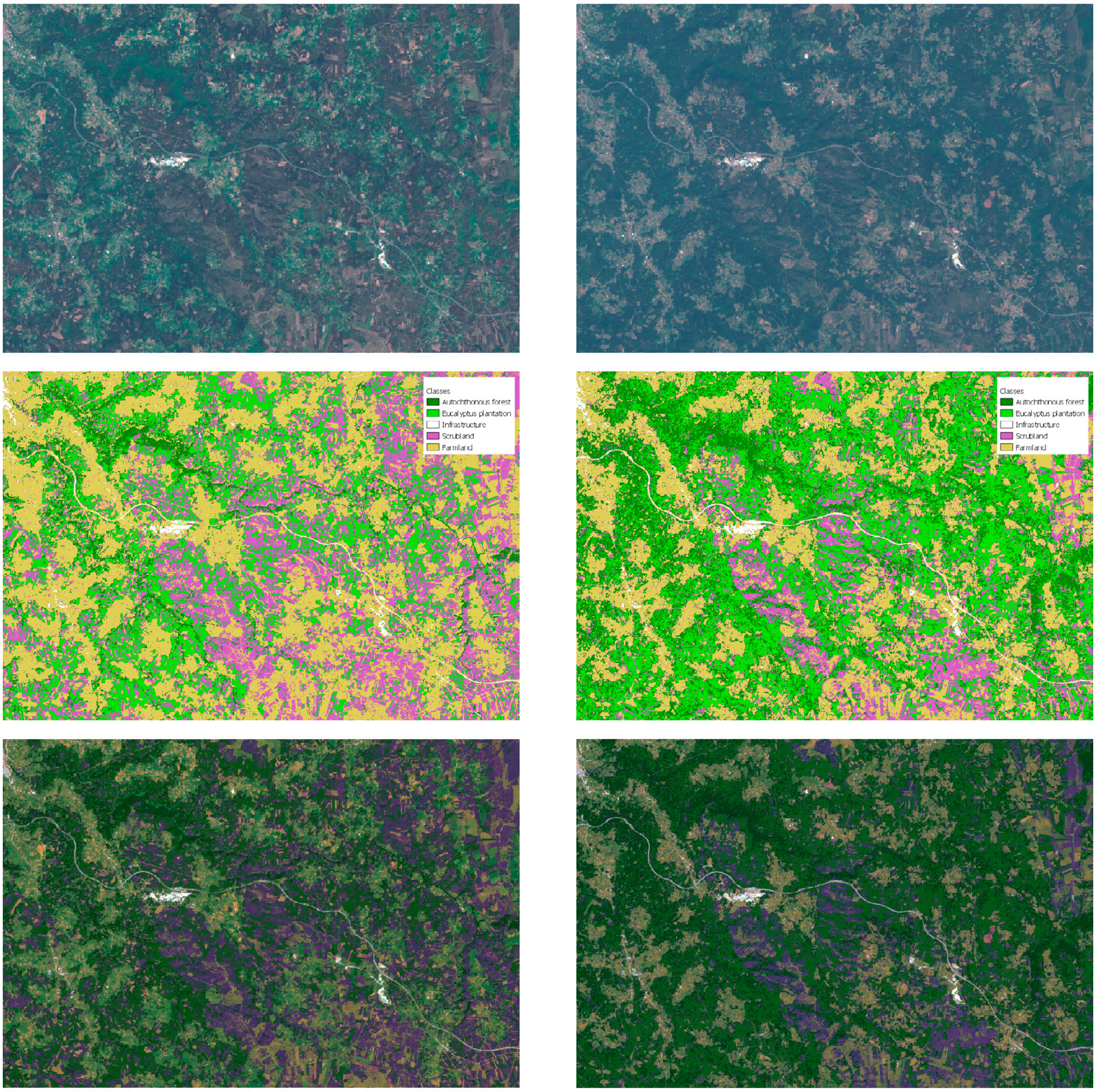


Fig. 5: Examples of classification results of the proposed methodology over the studied Betanzos–Mandeo SAC region. 1st row, original Sentinel-2 image products. 2nd row, prediction results of the proposed method. 3rd row, prediction results superposed with the original Sentinel-2 data.

Future works plan the implementation of deep learning techniques to improve the performance of the methodology and the quality of the results. Similarly, the introduction of more data samples to the training and testing sets will also improve the overall accuracy of the methodology. The dataset could also be improved by refining the studied classes to account for more diverse terrain identification.

Acknowledgements

This work is supported by the Biodiversity Foundation of the Ministry for the Ecological Transition (Spain), through a grant awarded to the UDC in response to call 2017 of research projects in the domain of Adaptations to Climatic Change. Also, this work has received financial support from the European Union (European Regional Development Fund - ERDF) and the Xunta de Galicia, Centro singular de investigación de Galicia accreditation 2016-2019, Ref. ED431G/01; and Grupos de Referencia Competitiva, Ref. ED431C 2016-047.

References

[1] Bossard, M., Feranec, J., Otahel, J., 2000. CORINE land cover technical guide: Addendum 2000 .
[2] Castaldi, F., Hueni, A., Chabrillat, S., Ward, K., Buttafuoco, G., Bomans, B., Vreys, K., Brell, M., van Wesemael, B., 2019. Evaluating the capability of the Sentinel 2 data for soil organic carbon prediction in croplands. ISPRS Journal of Photogrammetry and Remote Sensing 147, 267–282. doi:10.1016/j.isprsjprs.2018.11.026.
[3] Chauhan, S., Darvishzadeh, R., Boschetti, M., Pepe, M., Nelson, A., 2019. Remote sensing-based crop lodging assessment: Current status and perspectives. ISPRS Journal of Photogrammetry and Remote Sensing 151, 124–140. doi:10.1016/j.isprsjprs.2019.03.005.
[4] Chen, Y., Li, L., Lu, D., Li, D., 2019. Exploring Bamboo Forest Aboveground Biomass Estimation Using Sentinel-2 Data. Remote Sensing 11, 7. doi:10.3390/rs11010007.
[5] DEFRIES, R.S., TOWNSHEND, J.R.G., 1994. NDVI-derived land cover classifications at a global scale. International Journal of Remote Sensing 15, 3567–3586. doi:10.1080/01431169408954345.
[6] Foody, G.M., 2002. Status of land cover classification accuracy assessment. Remote Sensing of Environment 80, 185–201. doi:10.1016/S0034-4257(01)00295-4.
[7] Griffiths, P., Nendel, C., Hostert, P., 2019. Intra-annual reflectance composites from Sentinel-2 and Landsat for national-scale crop and land cover mapping. Remote Sensing of Environment 220, 135–151. doi:10.1016/j.rse.2018.10.031.
[8] Guo, H., Wang, L., Chen, F., Liang, D., 2014. Scientific big data and Digital Earth. Chinese Science Bulletin 59, 5066–5073. doi:10.1007/s11434-014-0645-3.
[9] He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep Residual Learning for Image Recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778.
[10] Helber, P., Bischke, B., Dengel, A., Borth, D., 2018. Introducing Eurosat: A Novel Dataset and Deep Learning Benchmark for Land Use and Land Cover Classification, in: IGARSS 2018 - 2018 IEEE International Geoscience and Remote Sensing Symposium, pp. 204–207. doi:10.1109/IGARSS.2018.8519248.
[11] Immitzer, M., Vuolo, F., Atzberger, C., 2016. First Experience with Sentinel-2 Data for Crop and Tree Species Classifications in Central Europe. Remote Sensing 8, 166. doi:10.3390/rs8030166.
[12] Kussul, N., Lavreniuk, M., Skakun, S., Shelestov, A., 2017. Deep Learning Classification of Land Cover and Crop Types Using Remote Sensing Data. IEEE Geoscience and Remote Sensing Letters 14, 778–782. doi:10.1109/LGRS.2017.2681128.
[13] Liu, P., Di, L., Du, Q., Wang, L., 2018. Remote Sensing Big Data: Theory, Methods and Applications. Remote Sensing 10, 711. doi:10.3390/rs10050711.
[14] Lunetta, R.S., Knight, J.F., Ediriwickrema, J., Lyon, J.G., Worthy, L.D., 2006. Land-cover change detection using multi-temporal MODIS NDVI data. Remote Sensing of Environment 105, 142–154. doi:10.1016/j.rse.2006.06.018.
[15] Ma, Y., Wu, H., Wang, L., Huang, B., Ranjan, R., Zomaya, A., Jie, W., 2015. Remote sensing big data computing: Challenges and opportunities. Future Generation Computer Systems 51, 47–60. doi:10.1016/j.future.2014.10.029.
[16] Pesaresi, M., Corbane, C., Julea, A., Florczyk, A.J., Syrris, V., Soille, P., 2016. Assessment of the Added-Value of Sentinel-2 for Detecting Built-up Areas. Remote Sensing 8, 299. doi:10.3390/rs8040299.
[17] Quintano, C., Fernández-Manso, A., Fernández-Manso, O., 2018. Combination of Landsat and Sentinel-2 MSI data for initial assessing of burn severity. International Journal of Applied Earth Observation and Geoinformation 64, 221–225. doi:10.1016/j.jag.2017.09.014.
[18] Schwalb-Willmann, J., 2019. getSpatialData. https://github.com/16EAGLE/getSpatialData.
[19] Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A., 2015. Going Deeper With Convolutions, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–9.
[20] Thanh Noi, P., Kappas, M., 2018. Comparison of Random Forest, k-Nearest Neighbor, and Support Vector Machine Classifiers for Land Cover Classification Using Sentinel-2 Imagery. Sensors 18, 18. doi:10.3390/s18010018.
[21] Vitousek, P.M., 1994. Beyond Global Warming: Ecology and Global Change. Ecology 75, 1861–1876. doi:10.2307/1941591.
[22] Vrieling, A., Meroni, M., Darvishzadeh, R., Skidmore, A.K., Wang, T., Zurita-Milla, R., Oosterbeek, K., O'Connor, B., Paganini, M., 2018. Vegetation phenology from Sentinel-2 and field cameras for a Dutch barrier island. Remote Sensing of Environment 215, 517–529. doi:10.1016/j.rse.2018.03.014.
[23] Zellweger, F., De Frenne, P., Lenoir, J., Rocchini, D., Coomes, D., 2019. Advances in Microclimate Ecology Arising from Remote Sensing. Trends in Ecology & Evolution 34, 327–341. doi:10.1016/j.tree.2018.12.012.