https://doi.org/10.25686/7229.2020.6.58827

ON-THE-FLY LAND COVER MAPPING USING MACHINE LEARNING WITH MULTISPECTRAL SATELLITE IMAGERY ON GOOGLE EARTH ENGINE

S. Papaiordanidis, C. Minakou, I.Z. Gitas
Laboratory of Forest Management and Remote Sensing, School of Forestry and Natural Environment, Aristotle University of Thessaloniki

Land cover is one of the most important environmental variables used to describe natural ecosystems. Constant changes on the Earth's surface create the need for new, up-to-date, and accurate land cover maps. During the last decades, remote sensing products have been used in conjunction with field measurements for the production of land cover maps in a cost-efficient manner. The aim of the present study was the development of a method for reliable on-the-fly land cover mapping using the Random Forest classifier and Sentinel-2 multispectral imagery on the Google Earth Engine cloud platform. The Random Forests algorithm was employed, and two classification schemes were adopted, one using the original 44 land cover classes used by the CORINE land cover product, and one using five general land cover classes (Artificial surfaces, Agricultural areas, Forest and semi-natural areas, Wetlands, and Water bodies), resulting in the generation of two land cover maps. The 2018 CORINE land cover product was used to identify training samples and validate the resulting land cover maps. The results showcased that the 44-class land cover map had an overall accuracy of 72.85% and the 5-class land cover map 88.32%. Overall, the results of this research indicate the capability of Random Forest algorithm in the reliable land cover classification.

Keywords: *Land cover mapping, Sentinel-2, Google Earth Engine, Random Forest, Remote sensing, Pixel-based classification*

Introduction

Land cover type is highly correlated with the biophysical properties of an area, and thus it is considered one of the most important environmental variables (Mason et al., 2003). Natural processes on the Earth's surface have a notable impact on climate, biodiversity, and the ecosystems ability to fulfill the needs of the modern human society (Mahmood et al., 2014). Land cover mapping is essential for planning and management of natural resources, as well as modeling environmental variables (Gómez et al., 2016). Traditionally, field measurements of land cover are considered as one of the most reliable land cover mapping techniques, however they are also costly, time-consuming, and present spatial limitations (Friedl et al., 2002). Additionally, constant changes in the Earth's surface, often render the existing land cover maps obsolete, creating the need for new updated maps. These obstacles lead to shortages of up-to-date and reliable land cover maps, which limit management bodies in the decision-making process.

In order to address this challenge, many scientists have investigated the potential of various classification algorithms to produce accurate land cover maps using satellite imagery (Anthony et al., 2007; Mahdianpari et al., 2018; Moser et al., 2012; Rodriguez-Galiano et al., 2012). One of the most commonly used algorithms in land cover mapping is the Random Forests ensemble. Random Forests have been shown to provide high overall accuracy in complex land cover classification problems (Maxwell et al., 2019; Na et al., 2010; Pelletier et al., 2016; Stefanski et al., 2013; Waske and Braun, 2009).

One of the challenges of using Random Forests is the preparation and set-up of the algorithm and the required software, as well as the input dataset preprocessing, and selecting and labeling the training samples. Currently, new cloud-based technologies offer easy access to data and large amounts of processing power to their users. Google Earth Engine (GEE) constitutes one of the most commonly used platforms which provides instant access to data and high processing power. Users can access GEE platform either from its online integrated development environment (IDE), or from the Application Program Interface (API) that is offered. Through GEE, users gain access to a large variety of datasets, including geospatial, satellite imagery, meteorological, census, etc. (Gorelick et al., 2017).

The aim of this study is to evaluate the effectiveness of Random Forest algorithm in land cover mapping, using Sentinel-2 multispectral data through the Google Earth Engine platform. The specific objectives are:

- The implementation of a method for calling and filtering the Sentinel-2 imagery archive and calculating spectral indices used for land cover mapping.
- The implementation of a method for training sample extraction from the Coordination of Information on the Environment (CORINE) land cover dataset.
- Land cover classification using the Random Forest algorithm.
- The quantitative and qualitative evaluation of the results in terms of land cover map overall accuracy and implementation effort.

Study area

The study area is located at the northern part of Greece, around the city of Thessaloniki, extending from 41°54'54.81" to 40°53'62.74" North, and from 22°19'91" to 23°47'71.06" East (fig. 1). The altitude ranges from sea level to 2,029 meters, and the surface area is 10,000 km^2. The climate is typical Mediterranean, with warm and dry summers, and cool and wet winters. This particular area was chosen because of its wide variety of land cover classes. Agricultural, forested, and artificial surfaces can be found in different variations, making the study area a suitable location to test the land cover mapping capabilities of the Random Forests algorithm.

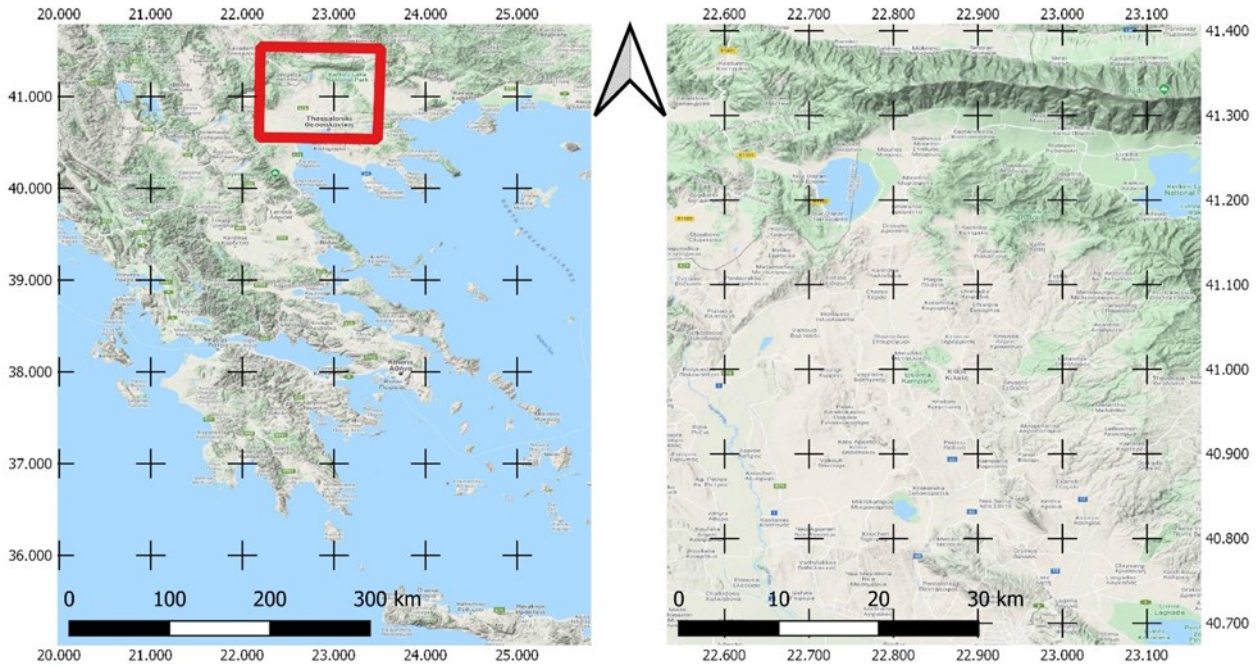

Fig. 1. Study area: Thessaloniki, Greece

Materials and methods

Sentinel-2 data

The satellite data that were used in this study included two images from Sentinel-2 MultiSpectral Instrument (MSI). The first image was acquired on October 26th,2018 and the second on August 31st,2020. Sentinel-2 imagery has a spatial resolution of 10, 20, and 60 meters (depending on the band), and a temporal resolution of 3-5 days depending on the location. The MSI instrument records spectral data in 13 bands ranging from the blue part of the electromagnetic spectrum (443 nm) to the short-wave infrared (2,190 nm).

The acquired images were already preprocessed to level 2A, which means that the images had been geometrically, radiometrically, and atmospherically corrected using the Sen2Cor algorithm by the Sentinel team (Main-Knorn et al., 2017).

CORINE data

The CORINE Land Cover (CLC) inventory provided by the Copernicus Land Monitoring Service was also used in this study (Büttner et al., 2004). The product is a thematic map with 44 land cover classes and a spatial resolution of 100 meters (fig. 2).

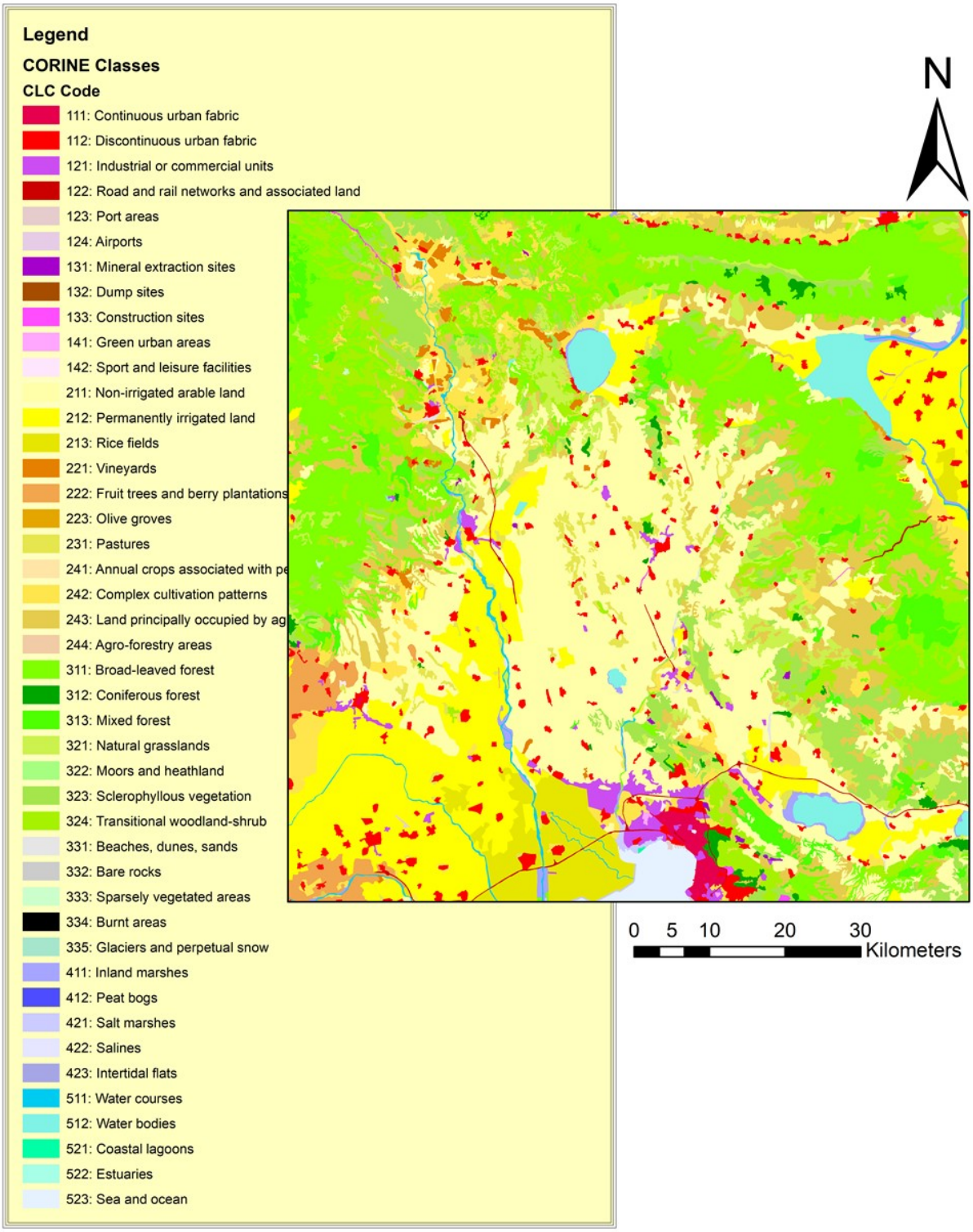

Fig. 2. CORINE land cover map of the study area with 44 classes

Additionally, a second map based on the CORINE product was constructed with the initial land cover classes summarized into five, namely Artificial surfaces, Agricultural area, Forest and semi-natural areas, Wetlands, and Water bodies (fig. 3).

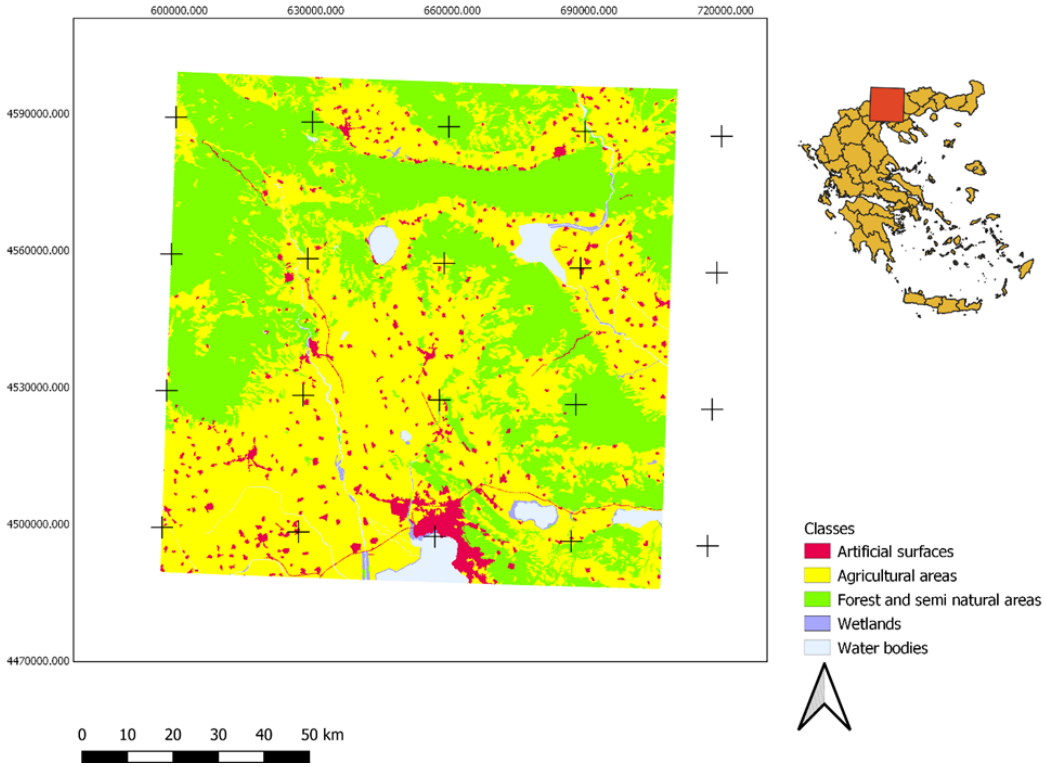

Fig. 3. CORINE land cover map of the study area with five classes

This was performed in order to investigate the potential effect of the employed classification scheme on the classifier's performance.

Methodology

The complete procedure of the methodology took part in GEE's online code editor (https://code.earthengine.google.com/). The steps followed are presented in the flowchart below (fig. 4).

The Sentinel-2 images were initially filtered by location and date. More specifically, the location filtering was carried out by setting a point in the study area and only keeping the Sentinel-2 images that their extent intersected with this point. The date filtering was done by limiting the images to ones that were acquired during 2018 and images that were acquired during 2020. The reason for that was to extract training

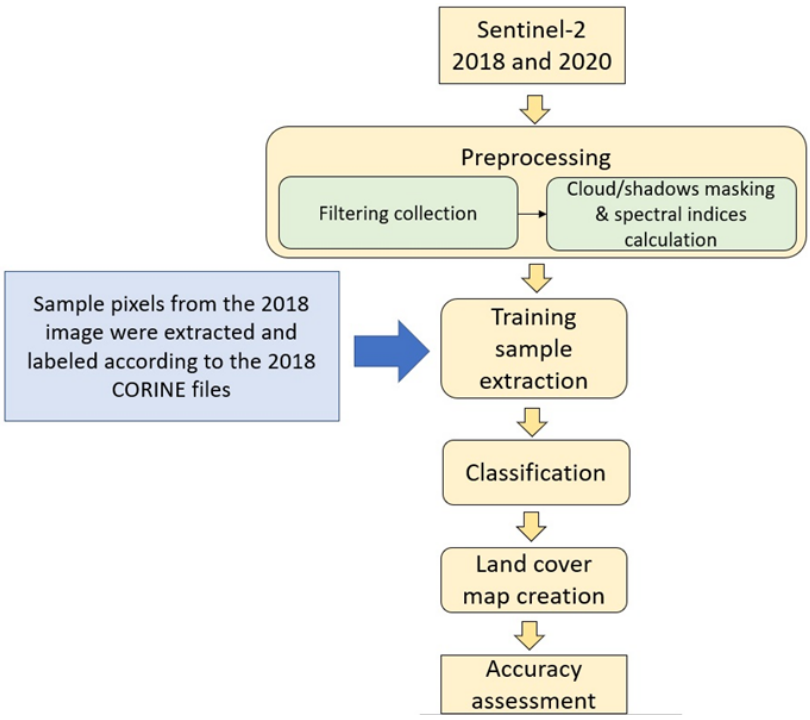

Fig. 4. Flowchart of the methodology

samples from the 2018 image, label them accordingly using the CORINE land cover map using random points, and then validate the trained algorithm using again random points (different from the ones used for training) labeled by the CORINE land cover product. The 2020 image was used to employ the trained algorithm and provide a land cover map for 2020.

The filtered images were then sorted in order of cloud cover using the Sentinel-2 metadata variable "CLOUDY_PIXEL_PERCENTAGE" and the least cloudy images for 2018 and 2020 were eventually selected. Next, a cloud and shadow mask was applied to the selected images by using the cirrus band and the cloud mask provided with Sentinel-2 2A level products.

After the examination of several spectral indices, BSI (Bare Soil Index), NDVI (Normalized Difference Vegetation Index), NDWI (Normalized Difference Water Index), and MCARI (Modified Chlorophyll Absorption in Reflectance Index) were selected to participate in the classification process (table 1).

Table 1

Spectral indices calculated

Index name	Formula	Citation
BSI	$((B6+B4)-(B5+B2))/((B6+B4)+(B5+B2))$	(Li and Chen, 2014)
NDVI	$((B8-B4))/((B8+B4))$	(Rouse Jr et al., 1974)
NDWI	$((B8-B11))/((B8+B11))$	(Gao, 1996)
MCARI	$[(B5-B4)-0.2*(B5-B3)]*(B5/B4)$	(Wu et al., 2008)

After the index calculation, 1000 random pixel samples were extracted from the 2018 Sentinel-2 image, labeled using the CORINE land cover map, and then the Random Forests algorithm was trained using 650 trees. Afterwards, 2000 random pixels (different from the ones used for training) were then selected in order to validate the classification accuracy using the CORINE land cover map. The validation was performed by GEE's native error matrix construction functions and the overall accuracy was calculated.

Finally, a land cover map was produced using as input the 2020 Sentinel-2 image to the trained algorithm. This methodology process was repeated two times, using the 44-class and the 5-class CORINE land cover map, respectively.

Results & discussion

The produced land cover maps are presented below (fig. 5, fig. 6):

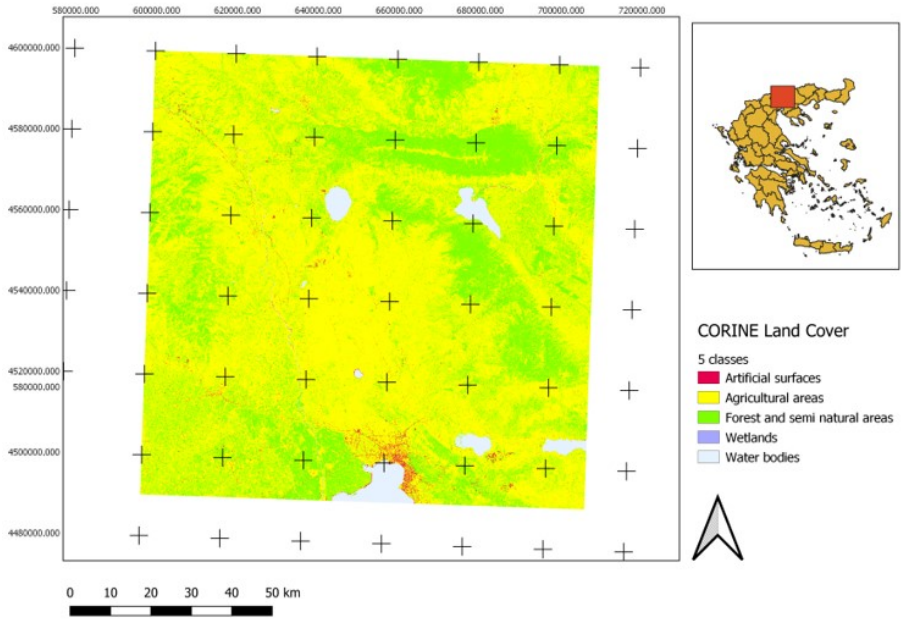

Fig. 5. Resulting 5-class land cover map produced by employing the Random Forests algorithm on the Sentinel-2 2020 image

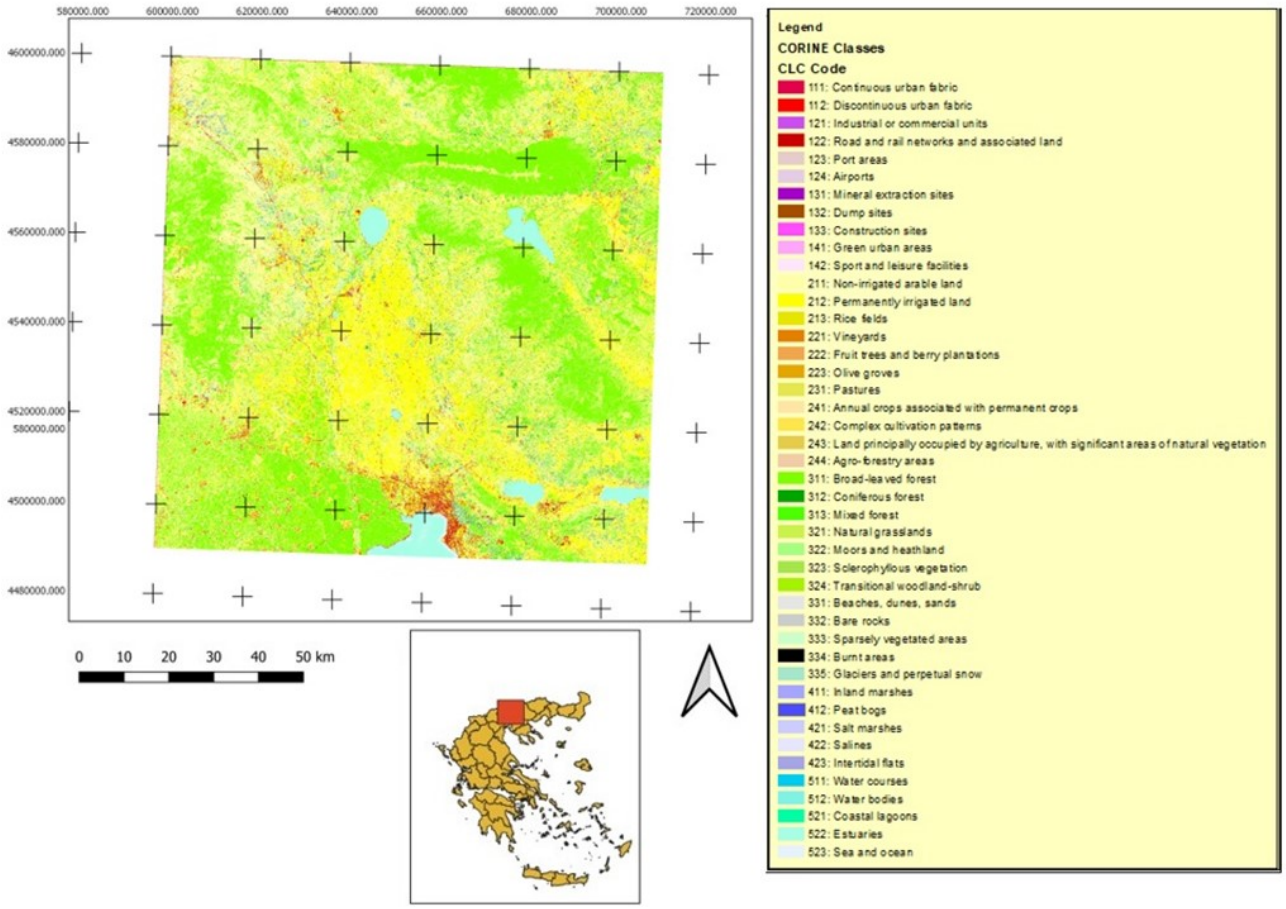

Fig. 6. Resulting 44-class land cover map produced by employing the Random Forests algorithm on the Sentinel-2 2020 image

The overall accuracies for the 5-class and the 44-class land cover maps were 88.32% and 72.85% respectively. Below, a comparison between each map and the corresponding original CORINE land cover map is presented (fig. 7, fig. 8).

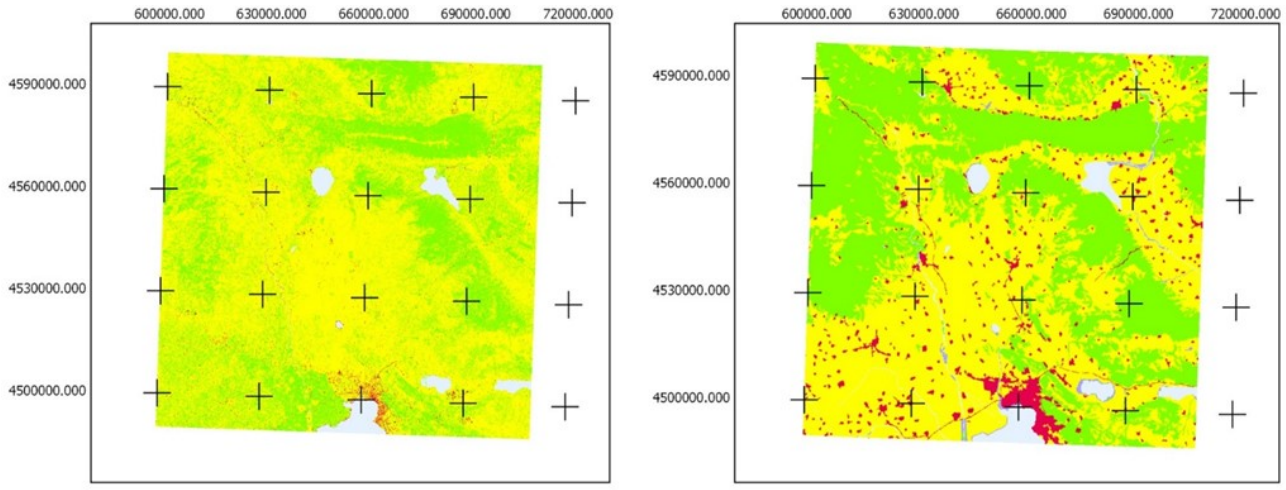

Fig. 7. The resulting 5-class land cover map (left) and the original 5-class CORINE land cover map (right)

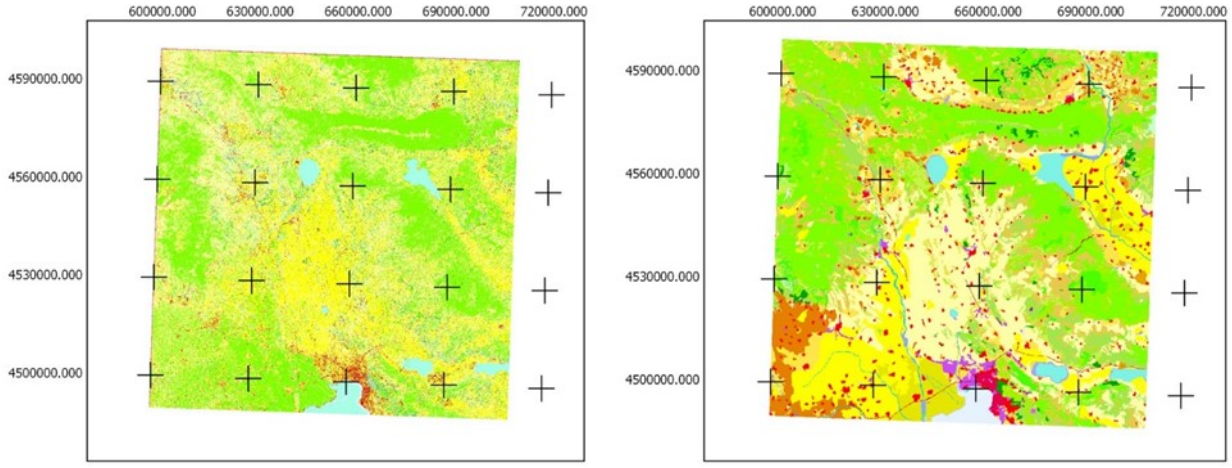

Fig. 8. The resulting 44-class land cover map (left) and the original 44-class CORINE land cover map (right)

It is obvious, that even though the overall accuracies of the resulting maps are quite high, and the general spatial distribution of the classes is correct, the fine details are lost. This classification error can be attributed to the fact that the spatial resolution of the CORINE land cover product is 100 meters, while Sentinel-2 spatial resolution is 10 meters. This could cause a problem in areas where a CORINE 100m pixel contains two different land cover classes. In these cases, only the value of the dominant class is given, while the spectral features recorded by Sentinel-2 are heterogeneous inside the confines of the CORINE pixel. Another issue with this classification is the fact that the distribution of classes was not taken into consideration. This means that since the training samples were selected randomly, minority classes (i.e. Wetlands) were not guaranteed to be part of the training samples.

One of the benefits of this methodology is that it can be used from any workstation with access to the internet and with no special preparation or special hardware requirements. Also, there is no need for downloading the required datasets since the whole procedure takes place in Google's remote servers. Another benefit of the methodology is its' time-efficiency since a Sentinel-2 tile can be classified in under a minute. This gives the ability to land managers to have a general idea of the spatial distribution of land covers in an area where no up-to-date land cover maps are available.

Conclusions

In this study, a method for calling the Sentinel-2 imagery archive and calculating spectral indices for land cover mapping was implemented. A method for training sample extraction from the CORINE land cover dataset was applied and a land cover classification using the Random Forest algorithm was performed. A method for confusion matrix construction and overall accuracy estimation was also employed for the validation of the classification results. Finally, a quantitative and qualitative evaluation of results in terms of land cover map accuracy and implementation effort was conducted.

The results showcased that produced maps had relatively high overall accuracy, but closer inspection of the maps revealed that there were some issues with the classification. For example, while the general class distribution was correct, the fine details of the land cover distribution were lost. This method provides the ability to land managers to trade between availability and accuracy. If an accurate and reliable land cover map is available there is no need to use this method, but when there is no such product, this method can rapidly generate a land cover map that represents the general spatial distribution of land cover in an area. More specifically:

- Using the Random Forest algorithm on Sentinel-2 imagery and CORINE land cover maps on Google Earth Engine cloud platform, greatly decreases the time needed for land cover mapping.
- Using the Random Forest algorithm on Sentinel-2 imagery and CORINE land cover maps on Google Earth Engine cloud platform, still produces comparable results to other land cover mapping products.

Future research could include the use of other machine learning classification techniques (i.e. Deep Learning) and Sentinel-2 imagery on Google Earth Engine. A different training set other than CORINE land cover map, or even different optical satellite datasets with higher resolution (i.e. WorldView). Finally, other types of remote sensing data like LiDAR or SAR could be investigated in combination with the above-mentioned methods.

This work was carried out in the Laboratory of Forest Management and Remote Sensing, in the School of Forestry and Natural Environment, in Aristotle University of Thessaloniki. The lead author (Stefanos Papaiordanidis) would like to thank Professor Ioannis Gitas for the helpful discussion and their insightful comments.

The European Commission support for the production of this publication does not constitute an endorsement of the contents which reflects the views only of the authors, and the Commission cannot be held responsible for any use which may be made of the information contained therein.

References

1. Anthony, G. Image classification using SVMs: one-against-one vs one-against-all, G. Anthony, H. Gregg, M. Tshilidzi, *arXiv preprint arXiv:0711.2914*, **2007**.
2. Büttner, G. The CORINE land cover 2000 project, Büttner, G., Feranec, J., Jaffrain, G., Mari, L., Maucha, G. and Soukup, T., *EARSeL eProceedings*, **2004**, Vol. 3, No. 3, pp 331-346.
3. Friedl, M. Global land cover mapping from MODIS: algorithms and early results M.A. Friedl, D.K. McIver, J.C.Hodges, X.Y. Zhang, D. Muchoney, A.H. Strahler, et al., *Remote sensing of Environment*, **2002**, Vol. 83, No. 1-2, pp. 287-302.
4. Gao, B.-C. NDWI—A normalized difference water index for remote sensing of vegetation liquid water from space, Gao, B.-C., *Remote sensing of Environment*, **1996**, Vol. 58, No. 3, pp 257-266.
5. Gómez, C. Optical remotely sensed time series data for land cover classification: A review, C. Gómez, J.C. White, and M.A.Wulder, *ISPRS Journal of Photogrammetry and Remote Sensing*, **2016**, Vol. 116, No., pp. 55-72.
6. Li, S. A new bare-soil index for rapid mapping developing areas using landsat 8 data, Li, S. and Chen, X., *The International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences*, **2014**, Vol. 40, No. 4, pp 139.
7. Mahdianpari, M. Very deep convolutional neural networks for complex land cover mapping using multispectral remote sensing imagery, Mahdianpari, M., Salehi, B., Rezaee, M., Mohammadimanesh, F. and Zhang, Y., *Remote Sensing*, **2018**, Vol. 10, No. 7, pp. 1119.
8. Mahmood, R. Land cover changes and their biogeophysical effects on climate, Mahmood, R., Pielke Sr, R. A., Hubbard, K. G., Niyogi, D., Dirmeyer, P. A., McAlpine, C. et al., *International journal of climatology*, **2014**, Vol. 34, No. 4, pp. 929-953.
9. Main-Knorn, M. Sen2Cor for sentinel-2, Main-Knorn, M., Pflug, B., Louis, J., Debaecker, V., Müller-Wilm, U. and Gascon, F., Proc. Conf. *Image and Signal Processing for Remote Sensing XXIII*, **2017**, International Society for Optics and Photonics.
10. Mason, P. The second report on the adequacy of the global observing systems for climate in support of the UNFCCC, P. Mason, M. Manton, D. Harrison, A. Belward, A. Thomas, D. Dawson, et. al., *GCOS Rep*, **2003**, p. 82.
11. Maxwell, A. E. Large-Area, High Spatial Resolution Land Cover Mapping Using Random Forests, GEOBIA, and NAIP Orthophotography: Findings and Recommendations, Maxwell, A. E., Strager, M. P., Warner, T. A., Ramezan, C. A., Morgan, A. N. and Pauley, C. E., *Remote Sensing*, **2019**, Vol. 11, No. 12, pp 1409.
12. Moser, G. Land-cover mapping by Markov modeling of spatial–contextual information in very-high-resolution remote sensing images, Moser, G., Serpico, S. B. and Benediktsson, J. A., *Proceedings of the IEEE*, **2012**, Vol. 101, No. 3, pp 631-651.
13. Na, X. Improved land cover mapping using random forests combined with landsat thematic mapper imagery and ancillary geographic data, Na, X., Zhang, S., Li, X., Yu, H. and Liu, C., *Photogrammetric Engineering & Remote Sensing*, **2010**, Vol. 76, No. 7, pp 833-840.
14. Pelletier, C. Assessing the robustness of Random Forests to map land cover with high resolution satellite image time series over large areas, Pelletier, C., Valero, S., Inglada, J., Champion, N. and Dedieu, G., *Remote sensing of Environment*, **2016**, Vol. 187, No., pp 156-168.
15. Rodriguez-Galiano, V. F. An assessment of the effectiveness of a random forest classifier for land-cover classification, Rodriguez-Galiano, V. F., Ghimire, B., Rogan, J., Chica-Olmo, M. and Rigol-Sanchez, J. P., *ISPRS Journal of Photogrammetry and Remote Sensing*, **2012**, Vol. 67, No., pp 93-104.

16. Rouse Jr, J. Paper A 20, Rouse Jr, J., Haas, R., Schell, J. and Deering, D., Proc. Conf. *Third Earth Resources Technology Satellite-1 Symposium: The Proceedings of a Symposium Held by Goddard Space Flight Center at Washington, DC on December 10-14, 1973: Prepared at Goddard Space Flight Center*, **1974**, Scientific and Technical Information Office, National Aeronautics and Space ….
17. Stefanski, J. Optimization of object-based image analysis with random forests for land cover mapping, Stefanski, J., Mack, B. and Waske, B., *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, **2013**, Vol. 6, No. 6, pp 2492-2504.
18. Waske, B. Classifier ensembles for land cover mapping using multitemporal SAR imagery, Waske, B. and Braun, M., *ISPRS Journal of Photogrammetry and Remote Sensing*, **2009**, Vol. 64, No. 5, pp 450-457.
19. Wu, C. Estimating chlorophyll content from hyperspectral vegetation indices: Modeling and validation, Wu, C., Niu, Z., Tang, Q. and Huang, W., *Agricultural and forest meteorology*, **2008**, Vol. 148, No. 8-9, pp 1230-1241.