Supervised and Unsupervised Learning for Remote Sensing Land Use/Land Cover

Antonio Di Noia[a], Mariagrazia Cairo[a], Manuela Casole[a], Mariapia Angelino[a], Carmelina Fascione[a]

[a]*Department of Statistics, Sapienza University of Rome, Rome, Italy*

Abstract

In this paper a land cover analysis with statistical machine learning techniques is provided. A dataset of satellite images from Landsat 8 was collected in order to analyse the land cover features. In particular, the study of the province of Rome is the primary objective of the analysis. First of all, an unsupervised clustering procedure is performed in order to detect homogeneous land cover classes. Furthermore, some popular and powerful classification algorithms (e.g. decision trees and random forests) are implemented and successively assessed in terms of accuracy. The obtained results could be useful to map and identify land cover classes, and other issues regarding the environmental health (e.g. the vegetation greenness). This work aims to test different machine learning algorithms to classify land cover using high-resolution imagery. Another interesting purpose is to empirically verify the higher theoretical accuracy of the ensemble methods compared to their non-ensemble version.

Keywords: Remote sensing, Ensemble methods, Supervised Learning, Classification, Land Use-Land Cover, Corine, Landsat

1. Introduction

Understanding the interactions between human being and earth resources is fundamental. In literature this framework is denoted as Land Use/Land Cover (LULC) analysis which is mainly faced by Remote Sensing (RS) data and methodologies. The results obtained in this field of research are

Email addresses: dinoia.1959841@studenti.uniroma1.it (Antonio Di Noia), cairo.1955600@studenti.uniroma1.it (Mariagrazia Cairo), casole.1924024@studenti.uniroma1.it (Manuela Casole), angelino.1968154@studenti.uniroma1.it (Mariapia Angelino), fascione.1742024@studenti.uniroma1.it (Carmelina Fascione)

turning out to be more and more important for several purposes such as urban and regional planning, environmental vulnerability and impact assessment, natural disasters and hazards monitoring and estimation of soil erosion and salinity (Talukdar et al. 2020). More specifically, the analysis is focused on the Land Cover side providing the quantification of the vegetation index for the province of Rome.

The RS dataset was opportunely preprocessed using QGIS 3.18.3 and then analysed involving the most appealing statistical machine learning techniques, both supervised and unsupervised, providing also a comparison of these methodologies in terms of performance and accuracy. The first exploratory step consisted of an unsupervised clustering procedure using k-means algorithm. However, the most interesting modelling step has been essentially a supervised classification problem using decision trees and random forests: the use of an image downloaded from Corine Land Cover has been fundamental at this point. In this phase the sample was randomly split in training and test sets: this passage is crucial since the size and the quality of training data has a great impact on the accuracy of the classification as well as the test set which is fundamental to evaluate the performance of the proposed models.

2. Dataset description and preprocessing

2.1. Study area

The study area of this paper is the province of Rome which is part of the region of Lazio (Italy) located on latitude $41°53'35''$N and longitude $12°28'58''$E. It measures approximately $5000\ km^2$ and covers less than one third of the region.

The province has a temperate climate with the highest average temperature of 30°C in summer months when rainy and cloudy days are less frequent. This aspect is extremely relevant since the clarity of the images used in the study is directly related to atmosphere conditions. In this area there are different land covers like forests, urban areas, cultivated lands and water bodies: for this reason the study of the land cover in this area turned out to be very interesting.

2.2. Dataset

The determination of the Land Use/Land Cover was based on the use of Landsat 8 satellite images. Landsat 8 satellite was chosen among other satellites (Cao et al., 2020, Song et al., 2021)

because it provides multispectral images that can be used actually for land cover classification. The collected dataset derived from Landsat 8 OLI (Operational Land Imager) and TIRS (Thermal Infrared Sensor) Level-1: these data are well-characterized in terms of radiometric quality since they are cross-calibrated among the different Landsat sensors.

More specifically an image of the province of Rome was downloaded from the United States Geological Survey (USGS) website after the definition of the time frame: from 1^{st} July 2018 to 15^{th} July 2018, since in this time range no rainfall occurred. Among different images the most clear and cloud-free was selected and became the object of the study. Six spectral bands of this image were selected for their radiometric and spectral characteristics: Blue (Band 2), Green (Band 3), Red (Band 4), Near infra-red, (Band 5), Short Wave infra-red 1 (Band 6), Short wave infra-red 2 (Band 7).

In order to implement a supervised classification, an image from Corine Land Cover (CLC) was picked. It was downloaded from CLC 2018 which is based on 2017 and 2018 satellite images: it contained spatial information of different physical coverage classes of Italy and other south Europe countries.

The first level (L1) of thematic detail of Corine Land Cover nomenclature was chosen to compare the pixel characteristics of CLC image with satellite ones from Landsat 8. L1 includes five classes of land cover: artificial surfaces, agricultural areas, forests and semi-natural areas, wetlands and water bodies. In this study wetlands are left out since they are not a significant land cover class for the study area.

2.3. Image preprocessing

In order to perform some learning algorithms, the preprocessing of images is crucial (Fichera et al., 2012, Ghosh and Hijmans, 2019, Qian et al., 2015, Talukdar et al., 2020, Rogan et al., 2008, Aguilera, 2020, Goldstein, 2021). The collected images, indeed, had a different scale: Landsat 8 images represented the province of Rome while CLC image was related to Italy and other south Europe countries. Moreover, Landsat 8 images showed a black frame which was removed in order not to affect the results. Since the perfect matching of the images data points is necessary for the analysis, they were imported in QGIS (Varga et al. (2021)), a free and open-source application which supports viewing, editing, analysis of geospatial data. The CLC image was cropped out using Landsat 8 image. At the end of this process the cropped images were exported in `.tif` format:

the initial format of the downloaded images.

Once obtained the images of the same study area, an important step was the definition and derivation of the Normalized Difference Vegetation Index (NDVI) as combination of Red and Nir bands. Let Band4 be the red band and Band5 be the Nir one it follows that

$$NDVI = \frac{Band5 - Band4}{Band5 + Band4}.$$

NDVI is a robust index of vegetation greenness, very used to distinguish between different land cover types and fundamental to pick out different levels of vegetation and soil: vegetation indeed has a unique spectral signature which distinguishes it among other types of land cover metrics. The role played by this index and the chosen six bands of Landset 8 as input data was crucial to define, analyse and predict different land cover classes using statistical machine learning techniques.

3. Statistical analysis

3.1. Unsupervised learning

The involved statistical methods are both unsupervised and supervised. First of all, to have an intuition of the most significant spectral bands, a canonical Principal Component Analysis (PCA) is performed. In particular the aim is to find low dimensional approximations to original data projecting on a linear subspace. Namely, let $X \in \mathbb{R}^d$ and $\mathcal{L}_k$ be all the k-dimensional linear subspaces. The k-th principal subspace is

$$\ell_k = \underset{\ell \in \mathcal{L}_k}{\operatorname{argmin}} \, \mathrm{E}\left[\min_{y \in \ell} \|X - \mathrm{E}[X] - y\|^2\right],$$

hence the dimension reduced X is

$$T(X) = \mathrm{E[X]} + \pi_{\ell_k} X,$$

where $\pi_{\ell_k} X$ is the projection of X onto ℓ_k, i.e. which turns out to be a linear combination of original features. Recall that each principal component is correlated with spectral bands in particular the most significant spectral bands are those being most correlated to the first principal components.

The modelling step was preceded by an unsupervised clustering step where the aim is to find

k significant clusters (eventually L1 of Corine) to predict land cover classes by means of some classification algorithm. The applied clustering algorithm was the k-means which basically consists of fixing the number of centroids and in every step minimize the euclidean distance between the updated centroid and each data point i.e.

$$\underset{c_i \in C}{\operatorname{argmin}} \|c_i - x\|^2$$

where c_i is the i-th centroid, C is the set of centroids and x is a generic data point.

3.2. Supervised learning

Priorly to the classification step a stratified sampling was performed in order to balance the sample with respect to the k-means clusters. The supervised step was performed by a classification tree since it has a simple interpretation of variable importance and it is one of the most suggested in literature for our purposes. The basic idea of tree growing is to choose a split among all the possible splits at each node so that the resulting child nodes are as homogeneous as possible within. The best split point is the one which is optimal with respect to the splitting criterion involved. Briefly, denote by R_m a region of the sample space induced by the node m and N_m the number of observations included in R_m, let

$$\hat{p}_{mk} = \sum_{x_i \in R_m} I_{\{y_i = k\}},$$

be the proportion of class k observation in node m. We classify the observations in node m to class

$$k(m) = \operatorname{argmax} \hat{p}_{mk}$$

which is the majority class in node m. Since we are dealing with an outcome taking values $1, 2, \ldots, K$ the error or node impurity could be defined in the following way

$$Q_m(T) = N_m^{-1} \sum_{i \in R_m} I_{\{y_i \neq k(m)\}} = 1 - \hat{p}_{mk},$$

where T is the considered subtree. Note that a simple and common way to stop the splitting process is to define a minimum tree size (here it is set to 5) and minimize the following

$$C_\alpha(T) = \sum_{m=1}^{|T|} N_m Q_m(T) + \alpha|T|,$$

where $|T|$ is the number of terminal nodes of the tree and α is the tuning parameter which governs the tradeoff between tree size and goodness of fit.

Another supervised learning algorithm often suggested in LULC literature is the Random Forest, the ensemble version of classification tree. It produces a collection of non-correlated trees and then averages them out. The principle on which the Random Forest relies on, is the so called bagging or bootstrap aggregation. Briefly, a bootstrap sample is drawn, then a tree T_b is grown to the bootstrap sample in such a way that at each split a random subset of m variables is extracted from the total p variables (typically $m = \sqrt{p}$) and used to grow T_b. Proceeding iteratively an ensemble of trees $\{T_b\}_1^B$ is obtained. For classification purposes the predicted class is given by the majority class or majority vote (majvote) among the individual predictions, namely

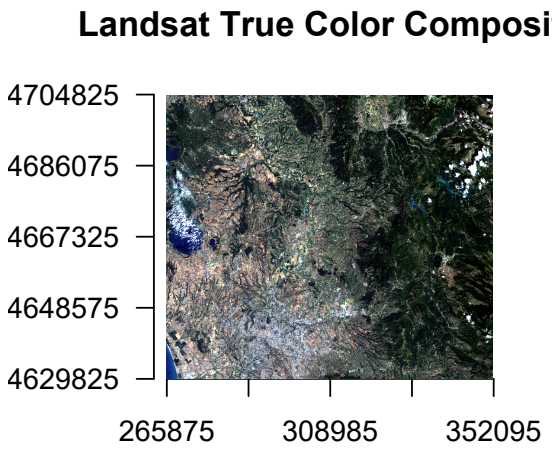
$$\hat{C}_{RF}^B(x) = \text{majvote}\{\hat{C}_b(x)\}_1^B.$$

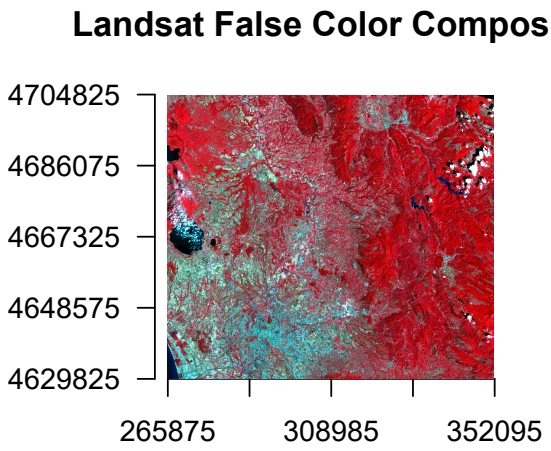
The model validation step is done by assessing the model accuracy defined in some way. First of all a k-fold cross-validation is implemented. In this technique the data used to fit the model is split into k groups (typically 5 groups). In turn, one of the groups will be used for model testing, while the rest of the data is used for model training (fitting). The procedure leads to the computation of the confusion matrix which permits to compute several measures of accuracy. Among all, the common overall accuracy, the Cohen's kappa which takes into account the possibility of the agreement occurring by chance, the user and the producer accuracy in order to detect class specific accuracy.

4. Results

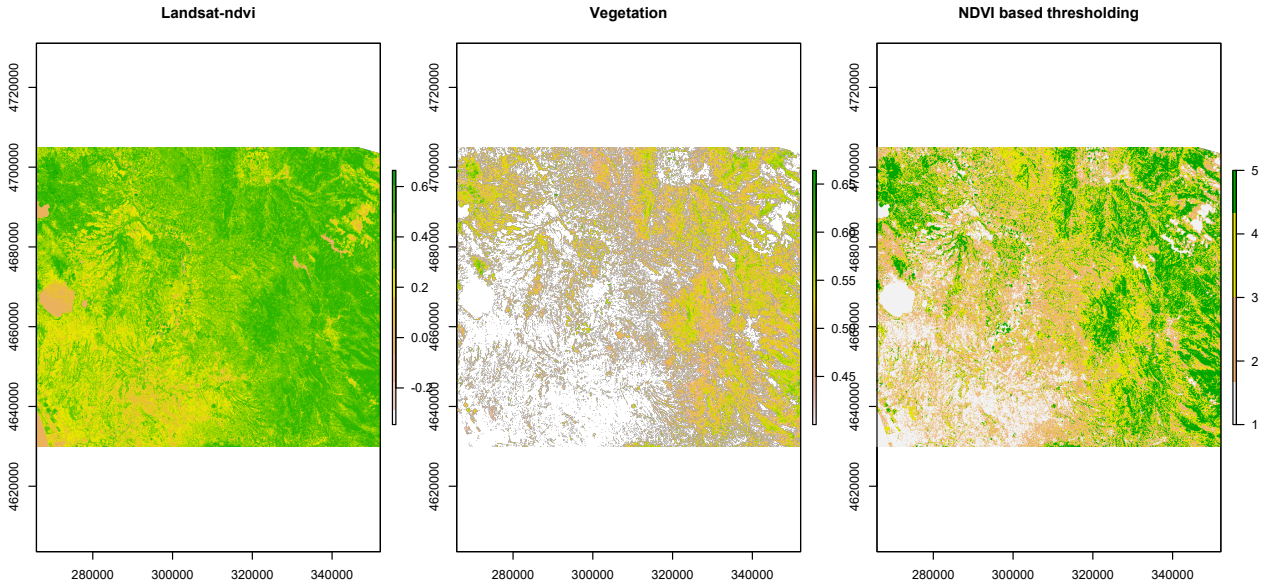
4.1. Image visualization

Before proceeding to PCA and k-means some basic exploratory results of image visualization are reported, in particular true and false colour images are represented in the following figure.





The computation of the NDVI leads to the following images obtained by hiding and thresholding NVDI values. The first is the basic NDVI, the second is obtained by hiding values lower then 0.4 and the third is obtained by thresholding for different values of the NDVI. Those permit to visualize areas with high vegetation land cover.

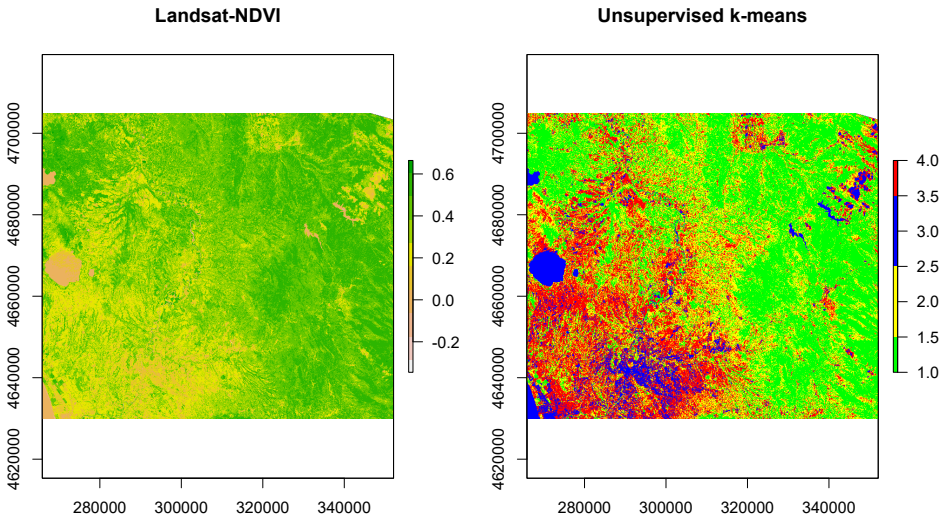


4.2. Unsupervised results

The PCA analysis suggests that all the spectral bands are significant, in particular the first 2 principal components, which explain the 90% of the variance, are correlated with all of spectral bands as it is evident from the following correlation coefficients table.

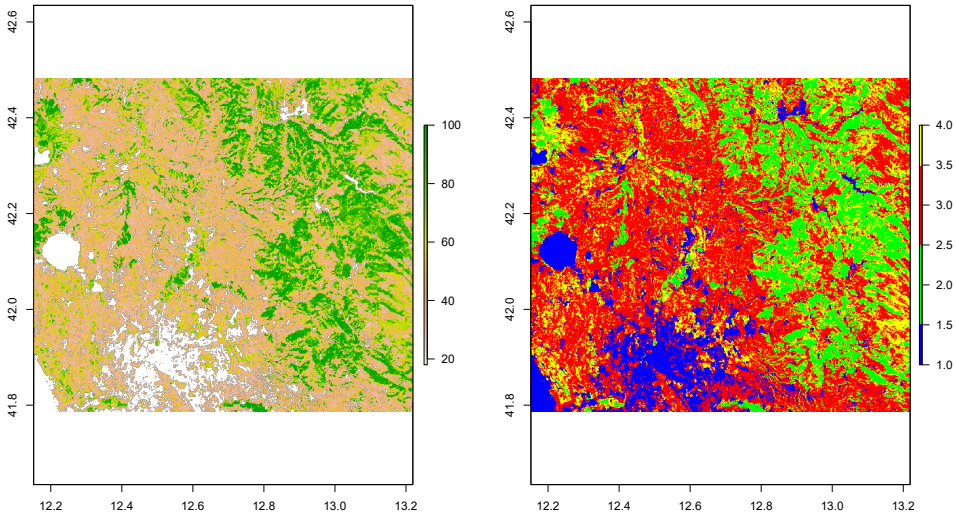
Band \ PC	PC1	PC2	PC3	PC4	PC5	PC6
B2	0.436	0.099	-0.551	0.599	-0.110	0.352
B3	0.459	0.022	-0.364	-0.189	0.170	-0.769
B4	0.468	0.108	-0.050	-0.670	0.219	0.519
B5	0.002	-0.957	-0.222	-0.116	-0.127	0.079
B6	0.417	-0.251	0.582	0.376	0.532	-0.008
B7	0.453	0.005	0.416	-0.039	-0.782	-0.090

The k-means algorithm over the NDVI raster with 4 centres shows robustness in reconstructing the previous image LULC structure. The following image shows the original and the clustered image where green denotes forests and high vegetation areas, red stands for urban and artificial areas, blue indicates water bodies, and yellow agricultural or cultivated areas.



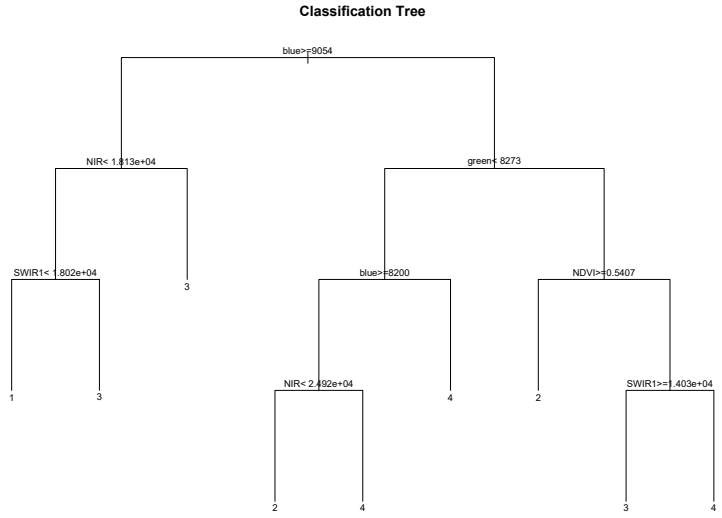
4.3. Supervised results

The first step of the supervised classification has been a stratified sampling significant in order to balance the sample with respect to the four centres of the k-means algorithm. The figure below is useful to recognize a parallelism between the original CLC2018 image and the corresponding image obtained applying k-means algorithm.



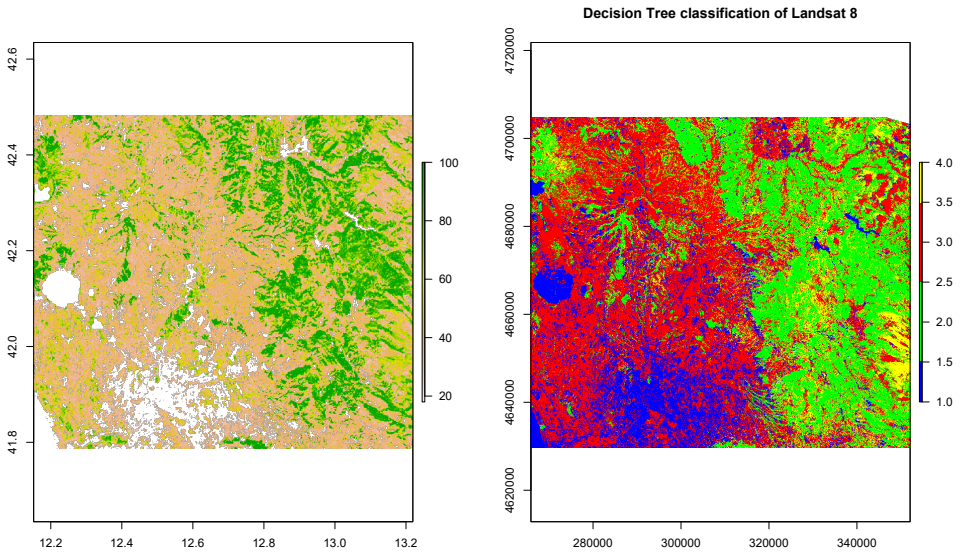
Decision trees and random forests algorithms utilize the pixel values of the training sites from Landsat 8 data to define the four land cover classes. The following figure shows the plot of the

trained classification tree obtained considering five bands: blue, green, Nir, SWIR1, SWIR2 and the vegetation index NDVI. The red band is not used since NDVI is a combination of the Nir and red bands: using this band does not increase the accuracy of the model. In the classification tree the assigned classes are printed at the leaf nodes.

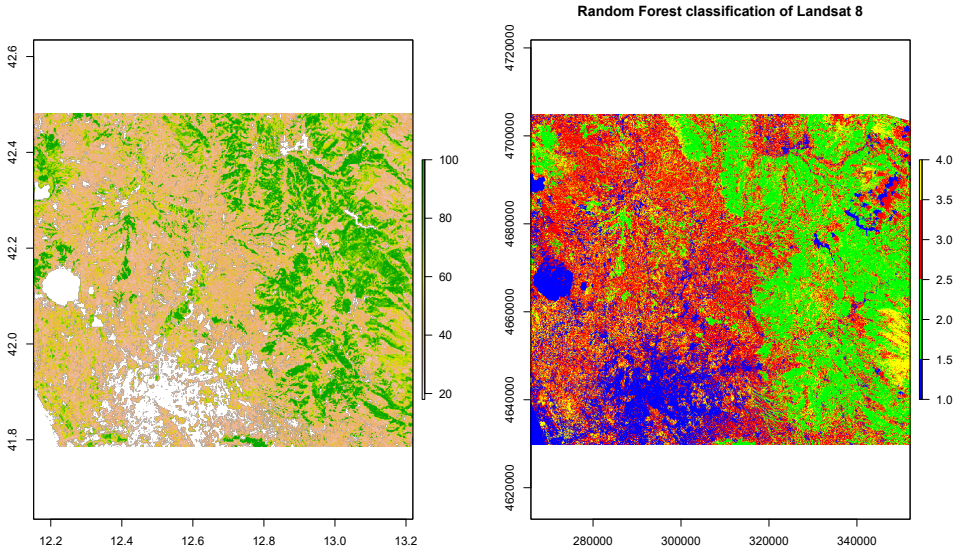


From the figure above it emerges that the blue band is really meaningful to discriminate between different land cover classes: this evidence is in accordance with the preliminary principal component analysis. Also the NIR band and the NDVI are extremely significant: they are fundamental to discriminate the forest from the other classes. As expected from theoretical results high values of NDVI classify pixels into the second class that corresponds to the forest one.

The figure below shows a comparison between the original CLC2018 image and the image obtained applying the classification tree: all the four classes seem to be well predicted and the image structure is reconstructed by the model.



In order to have a comparison between different models, the random forest algorithm is applied to the same CLC2018 image. From a theoretical point of view the expected accuracy of this model should be bigger since random forest uses a large number of decision trees in order to overcome the weaknesses of a single decision tree. The following picture shows a comparison between the original CLC2018 image and the plot turned up applying random forest. Comparing this plot with the plot obtained after performing decision tree it seems that this picture is more detailed.



In the table below the confusion matrix obtained after the implementation of decision tree and random forest is displayed. It summarizes the performance of the two classification algorithms.

The simple classification accuracy indeed can hide the details needed to diagnose the performance of the model over a single class once the dataset is composed of more than two classes. The confusion matrix instead shows the ways in which the classification model is confused when it makes predictions for each class.

Confusion Matrices							
Decision Tree				Random Forest			
Waterbodies	Forest	Artificial	Agricultural	Waterbodies	Forest	Artificial	Agricultural
67	2	27	4	71	0	15	7
6	75	10	9	4	67	6	21
21	8	50	21	14	9	53	20
13	28	30	29	8	28	31	27

To evaluate the obtained results in terms of accuracy for each model a cross validation has been done. The table below is representative to have a look at the accuracy of the two methods.

Accuracy \ Model	Decision tree	Random Forest
Overall accuracy	55%	57%
Kappa	40%	43%

From the analysis of this table the random forest seems to be the best model in terms of accuracy and Cohen's kappa: the random forest indeed seems to be more robust than the decision tree as expected from theoretical results. Although the random forests are more accurate than the decision trees, the computational cost in terms of time and computational effort is significantly different. Decision trees computational cost is indeed extremely lower. The overall accuracy results are satisfactory but the class specific accuracy is even better, around 70%, e.g. for forest areas. These results are reported as the producer and the user accuracy which are easily computed as the ratio between the principal diagonal of the confusion matrix and colsums and rowsums respectively.

	Producer accuracy		User accuracy	
Class \ Model	Decision Tree	Random Forest	Decision Tree	Random Forest
Waterbodies	63%	72%	67%	75%
Forest	66%	65%	75%	69%
Artificial	43%	48%	50%	53%
Agricultural	46%	37%	29%	29%

5. Conclusions and future developments

The focus of this study has been the building of a model to predict and detect LULC classes and the comparison between different machine learning algorithms both unsupervised and supervised for pixel classification of images from Landsat 8 satellite. The performance is satisfactory in particular when dealing with green areas detection. The results confirm the well-known theory on ensemble methods, in fact Random Forest is slightly more accurate despite decision trees are preferable from computational effort point of view.

Although the use of Landsat 8 images represent a good starting point for classification and leads to a satisfying land cover of the province of Rome, the study area lacks of one of the Corine land cover classes: wetlands. In the province of Rome indeed the percentage of wetlands is not significant and for this reason this land cover class is left out. Future developments could lead to deepen the level of the analysis basing the study on L2 or L3 Corine land cover classes instead of L1. Moreover the machine learning methods explored in this report could be easily applied to other satellite images: this study could be further explored analysing the LULC of different areas of study.

References

Aguilera, M. A. Z. (2020). Classification of land-cover through machine learning algorithms for fusion of sentinel-2a and planetscope imagery. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XLII-3/W12-2020.

Cao, Z., Ma, R., Duan, H., Pahlevan, N., Melack, J., Shen, M., and Xue, K. (2020). A machine learning approach to estimate chlorophyll-a from landsat-8 measurements in inland lakes. *Remote Sensing of Environment*, 248(111974).

Fichera, C., Modica, G., and Pollino, M. (2012). Land cover classification and change-detection analysis using multi-temporal remote sensed imagery and landscape metrics. *European Journal of Remote Sensing*, 45(1).

Ghosh, A. and Hijmans, R. (2019). Remote sensing image analysis with r.

Goldstein, S. (2021). Classifying satellite imagery in r.

Gromping, U. (2019). Variable importance assessment in regression: Linear regression versus random forest. *The American Statistician*, 63(4).

Gyamfi-Ampadu, E., Gebreslasie, M., and Mendoza-Ponce, A. (2020). Mapping natural forest cover using satellite imagery of nkandla forest reserve, kwazulu-natal, south africa. *Remote Sensing Applications: Society and Environment*, 18.

Keshtkar, H., Voigt, W., and Alizadeh, E. (2017). Land-cover classification and analysis of change using machine-learning classifiers and multi-temporal remote sensing imagery. *Arabian Journal of Geosciences*, 10(154).

Qian, Y., Zhou, W., Yan, J., Li, W., and Han, L. (2015). Comparing machine learning classifiers for object-based land cover classification using very high resolution imagery. *Remote Sensing*, 7.

Ribeiro, M. T., Singh, S., and Guestrin, C. (2016). "why should i trust you?" explaining the predictions of any classifier. *KDD '16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

Richetti, J., de Albuquerque Silva, L. C., Becker, W. R., Paludo, A., Comineti, H. J., and Johann, J. A. (2019). Machine learning alghorithms to land cover mapping with landsat-8. *XIX Simpósio Brasileiro de Sensoriamento*.

Rogan, J., Franklin, J., Stow, D., Miller, J., Woodcock, C., and Roberts, D. (2008). Mapping land-cover modifications over large areas: A comparison of machine learning alghorithms. *Remote Sensing of Environment*, 112.

Singh, V., Bhattacharjee, V., Jeganathan, C., and Bhattacherjee, N. (2020). Land cover classification using machine learning techniques - a survey. *International Journal of Engineering Research & Technology (IJERT)*, 9(06).

Song, X., Huang, W., Hansen, M., and Potapov, P. (2021). An evaluation of landsat, sentinel-2, sentinel-1 and modis data for crop type mapping. *Science of Remote Sensing*, 3(100018).

Talukdar, S., Singha, P., Mahato, S., Shahfahad, Pal, S., Liou, Y., and Rahman, A. (2020). Land-use land-cover classification by machine learning classifiers for satellite observationa-a review. *Remote Sensing*, 12.

Varga, O. G., Kovács, Z., Bekõ, L., Burai, P., Szabó, Z. C., Holb, I., Ninsawat, S., and Szabó, S. (2021). Validation of visually interpreted corine land cover classes with spectral values of satellite images and machine learning. *Remote Sensing*, 13(5).