Supervised Image Classification of a railway section in Guidonia Montecelio, Italy

Summary

[Aim and scope 2](#_Toc176081321)

[Methods 2](#_Toc176081322)

[Study area 2](#_Toc176081323)

[Tools and software 3](#_Toc176081324)

[Statistical model: Random Forest 3](#_Toc176081325)

[Data collection and aggregation 4](#_Toc176081326)

[Repository location 4](#_Toc176081327)

[Technical Validation 5](#_Toc176081328)

[Accuracy assessment and Confusion Matrix 5](#_Toc176081329)

[Results 6](#_Toc176081330)

[Land use/land cover Classification 6](#_Toc176081331)

[Corine Land Cover Level III Classification 7](#_Toc176081332)

[Bibliography 8](#_Toc176081333)

# Aim and scope

Supervised classification is a fundamental machine learning technique in remote sensing, essential for tasks like creating Land Use/Land Cover maps and monitoring changes in the environment. Google Earth Engine is particularly well-suited for conducting supervised classification on a large scale. The purpose of this brief report is to expose the methodology used to identify different categories of Land Use in an area of interest adjacent to a railroad section located in the municipality of Guidonia Montecelio, in the province of Rome, and to quantify the areas occupied by each of these categories, to support more detailed planning in the design of rail works and track modifications.

# Methods

## Study area

The study area was carried out in the Region of Lazio, central Italy, in the municipality of Guidonia Montecelio (41°59′48.55″N 12°43′34.1″E), between the localities of Tivoli Terme and Villalba, with a medium a.s.l. of 95 m. The climate of the area is of Mediterranean type: the mean minimum air temperature of the coldest month (January) was 7.4 °C, the mean maximum air temperature of the hottest months (August) 24.5 °C, an average of 30 days of frost per year and 60 days per year with a maximum temperature of 30 °C or more. The average annual rainfall was 813 mm, on average distributed over 82 days, with a minimum in summer and a maximum in autumn. The average annual relative humidity was 71.2% with a minimum of 66% in July and August and a maximum of 78% in November (Data from Guidonia Airport Meteorological Station for the years from 1971 to 2000).

Specifically, the area considered for the analysis was calculated by creating a buffer of 20 meters per side of the railway section between the coordinates points 41°58'14"N 12°43’26"E and 41°57'07"N 12°43’41”E, corresponding respectively to the upper and lower margins of the raster images owned and used for the classification.

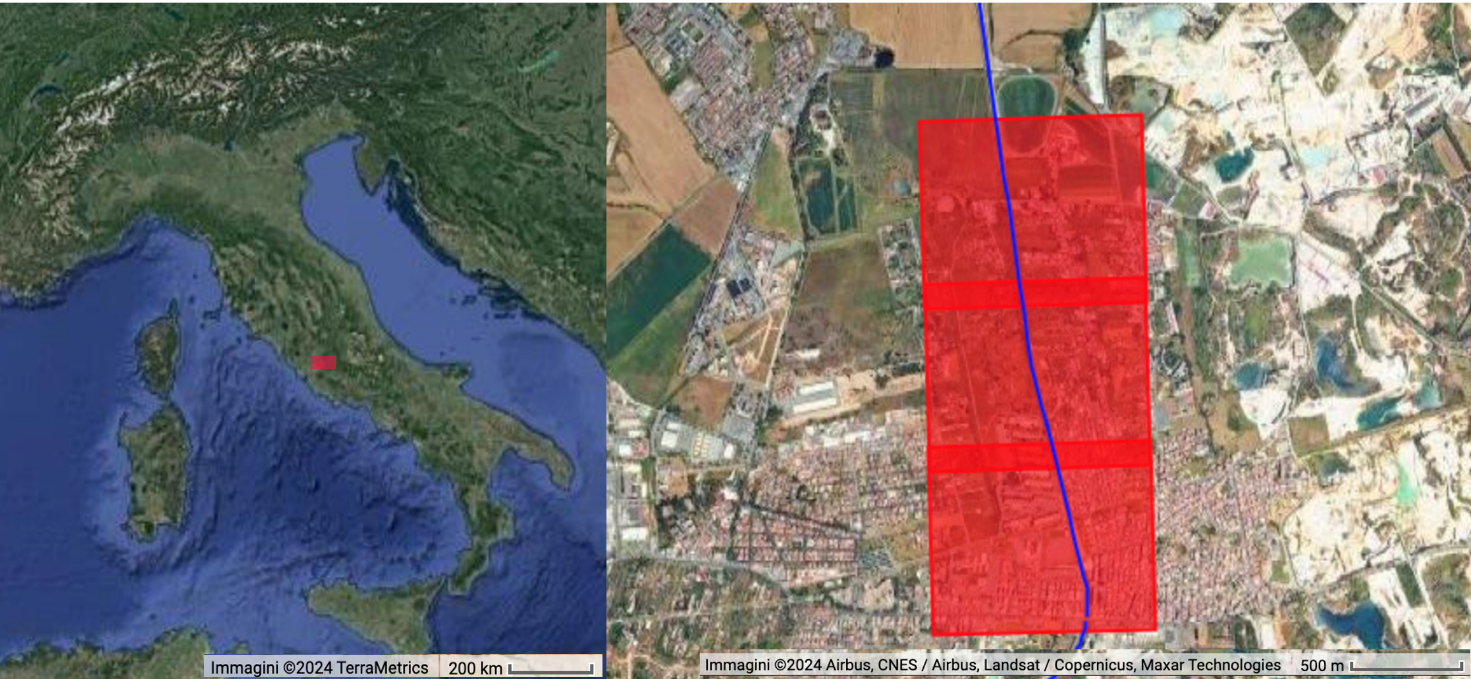


Figure 1: Region of interest in Guidonia Montecelio is defined by the three red boxes that represents the extents of the aerial images used for the classification. The blue line represents the railway infrastructure considered.

## Tools and software

Google Earth Engine (GEE) is a free cloud computing platform platform designed to address key challenges in large-scale land cover mapping (Gorelick, 2013). It hosts a large quantity of remotely-sensed data for the lasts 40 years, such as [Landsat](https://www.sciencedirect.com/topics/earth-and-planetary-sciences/landsat), MODIS, Sentinel 1, 2, 3 and 5-P; it also includes climate-weather and geophysical datasets. GEE platform permits to significantly reducing computational time since it utilizes Google's computational infrastructure to facilitate parallel processing of geospatial data, allowing users to analyze extensive collections of remotely sensed images directly through a web-based Integrated Development Environment code editor, eliminating the need to download data locally (Phan et al., 2020).

The availability of numerous packages with a wide range of algorithms makes script writing and remote sensing tools easier for both experts and novices. It offers several packages, such as machine learning, image processing, image collection, geometry-feature handling, reducers, charting tools, and specialized algorithms *(Tamiminia et al., 2020).*

## Statistical model: Random Forest

Random forest (RF) is a tree-based method and it’s increasingly used in ecology with minimal parameters adjustment (Valavi et al., 2021), its efficiency and management on large databases. It is an ensemble of classification trees based on recursing partitioning methods, which assembly repeatedly data into potential partitions were they have an homogenous response. RF fits many individual trees, usually several hundred to thousands, and combines their predictions in order to generalise better compared to a generalisation of data in individual trees (Valavi et al., 2021), producing an internal unbiased estimate of the generalisation error as the forest building progresses. RF, and in general tree-based methods, can deal with many predictors and elevate order interactions (Elith et al., 2019). This model can estimate missing data, maintaining accuracy even if a large proportion of the data are missing, and it balances errors in class population unbalanced datasets.  
During the process of constructing each tree, the random forest algorithm selects factors from the training samples to determine the optimal splits. While this approach may compromise the strength of individual trees, it enhances the diversity among trees, thereby reducing overall misclassification (Ahmed S. A. et al., 2023). The reliability and robustness of random forest have contributed to its widespread adoption, particularly in handling complex and exceptional cases (Eisavi et al., 2015). Studies have demonstrated the effectiveness of random forest in accurately classifying various types of land use and land cover (Waske and Braun, 2009).

In the GEE platform, the function *ee.Classifier.smileRandomForest* was used to create the algorithm; it was observed that the optimal number of trees was 70, as the model achieved a maximum accuracy by entering this value, while increasing the number of trees did not show any improvement in accuracy.

## Data collection and aggregation

Railway data at Italian territory level were downloaded from Geofabrik, a consulting and software development firm based in Karlsruhe, Germany, specialized in OpenStreetMap services: data are accurated and up-to-date.

Preliminarily, railway data were analyzed and filtered using the R program (Team, R. C., 2020). R: A language and environment for statistical computing.), in order to have a nationwide shapefile that can be used for any study area: sections of bridges and tunnels have been removed, and only railroad types equivalent to light rail, narrow gauge, rail sections have been maintained.

Regarding the raster files used, it was not possible to use high-resolution open access satellite images, such as Harmonized Sentinel-2 MSI (Claverie et al., 2018), since a 10 meters resolution was not defined enough to distinguish railway features. Therefore, very high resolution (less than 1 meter) private aerial photographs containing red, green and blue bands were used. A mosaic of the aerial images of the railway section in Guidonia Montecelio was classified; the images were uploaded in the GEE platform and their resolution decreased to 1 meter. The region of interest (ROI) was defined by creating a buffer of 20 meters on each side with respect to the geometry of the railroad infrastructure present in the area covered by the mosaic of the three images used. Subsequently, the classification was performed considering the areal of 40 meters total width and about 3 kilometers in length.

## Repository location

The GitHub repository contains all relevant scripts and data required for reproducing the analyses described in this study. Specifically, it includes JavaScript scripts for performing the supervised image classification, as well as an R script detailing the modifications made to the initial dataset on railway infrastructure. The README file provides detailed instructions on how to adapt the scripts to suit specific research interests. While some of the necessary files for these analyses are included in the repository, certain files could not be uploaded due to their large size. These files are available upon request. The GitHub repository can be accessed at <https://github.com/SofiaPrandelli/SupervisedImageClassification>, and the analyses on Google Earth Engine are available at <https://code.earthengine.google.co.in/58891e0600359fa8ed6612988d296ed4>.

# Technical Validation

## Accuracy assessment and Confusion Matrix

Accuracy assessment is a crucial component of any classification project, as it quantitatively evaluates the reliability of the classification results. In this study, the confusion matrix was generated using the ee.Classifier.confusionMatrix() method, which constructs a simple error matrix from the validation data. This approach is effective for assessing the classifier's performance, highlighting which classes are classified accurately and identifying those that require improvement (Ahmed S. A. et al., 2023). For instance, a poorly modeled class may benefit from the collection of additional training points.

The training samples were divided into two random fractions: 70% of the Ground Control Points (GCPs) were utilized to train the classifier for raster image classification, and the remaining 30% were randomly selected to validate the model by calculating the confusion matrix. The overall accuracy, which measures the proportion of correctly classified values along the diagonal from the upper-left to the lower-right of the confusion matrix relative to the total number of validation values, was found to be 91%.

Table 1

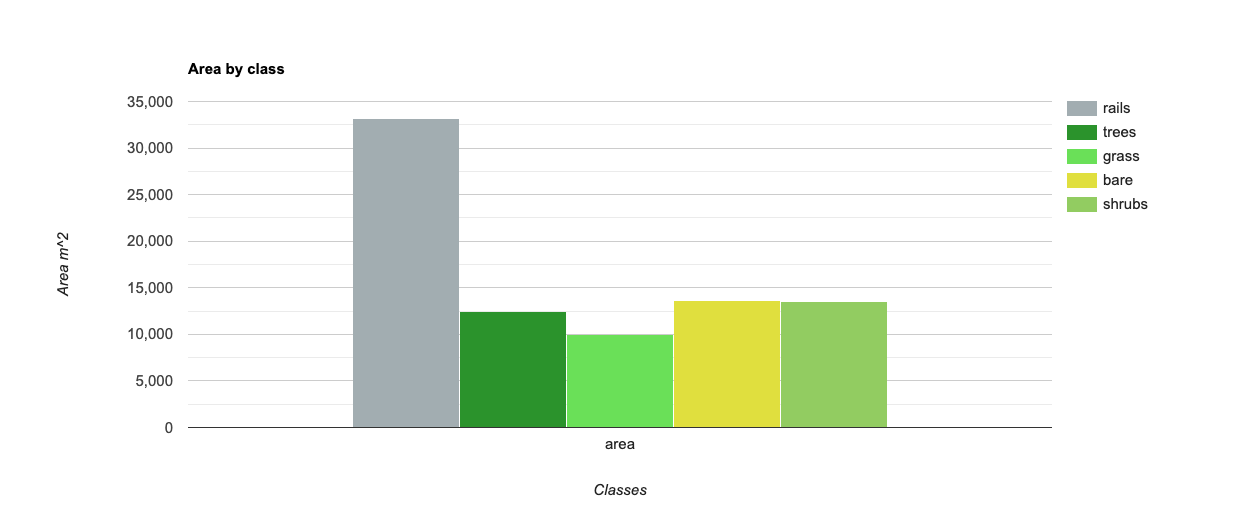
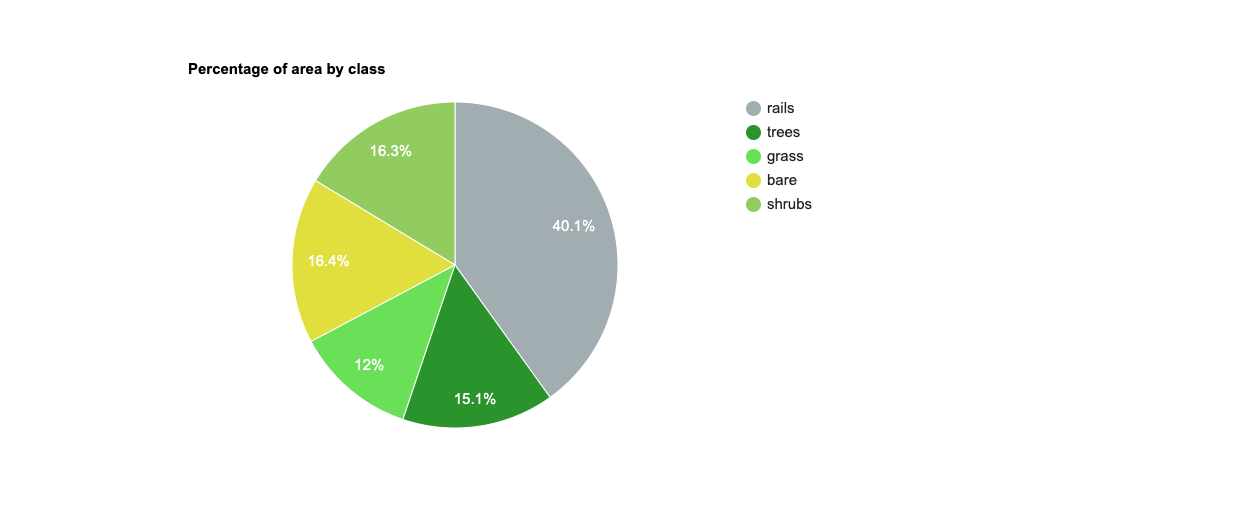
|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Predicted | Truth | | | | | |
|  | Rails | Trees | Grass | Bare | Shrubs |
| Rails | 13 | 0 | 0 | 0 | 0 |
| Trees | 0 | 12 | 0 | 0 | 0 |
| Grass | 0 | 1 | 9 | 0 | 3 |
| Bare | 0 | 1 | 0 | 7 | 0 |
| Shrubs | 0 | 0 | 0 | 0 | 12 |

As shown in Table 1, out of 58 evaluated points, 5 were incorrectly classified between the classes of bare vegetation and grass cover. Additionally, when observing the classification raster layer on the platform, the pixels demonstrate a strong correspondence to the actual data, further supporting the accuracy of the classification.

# Results

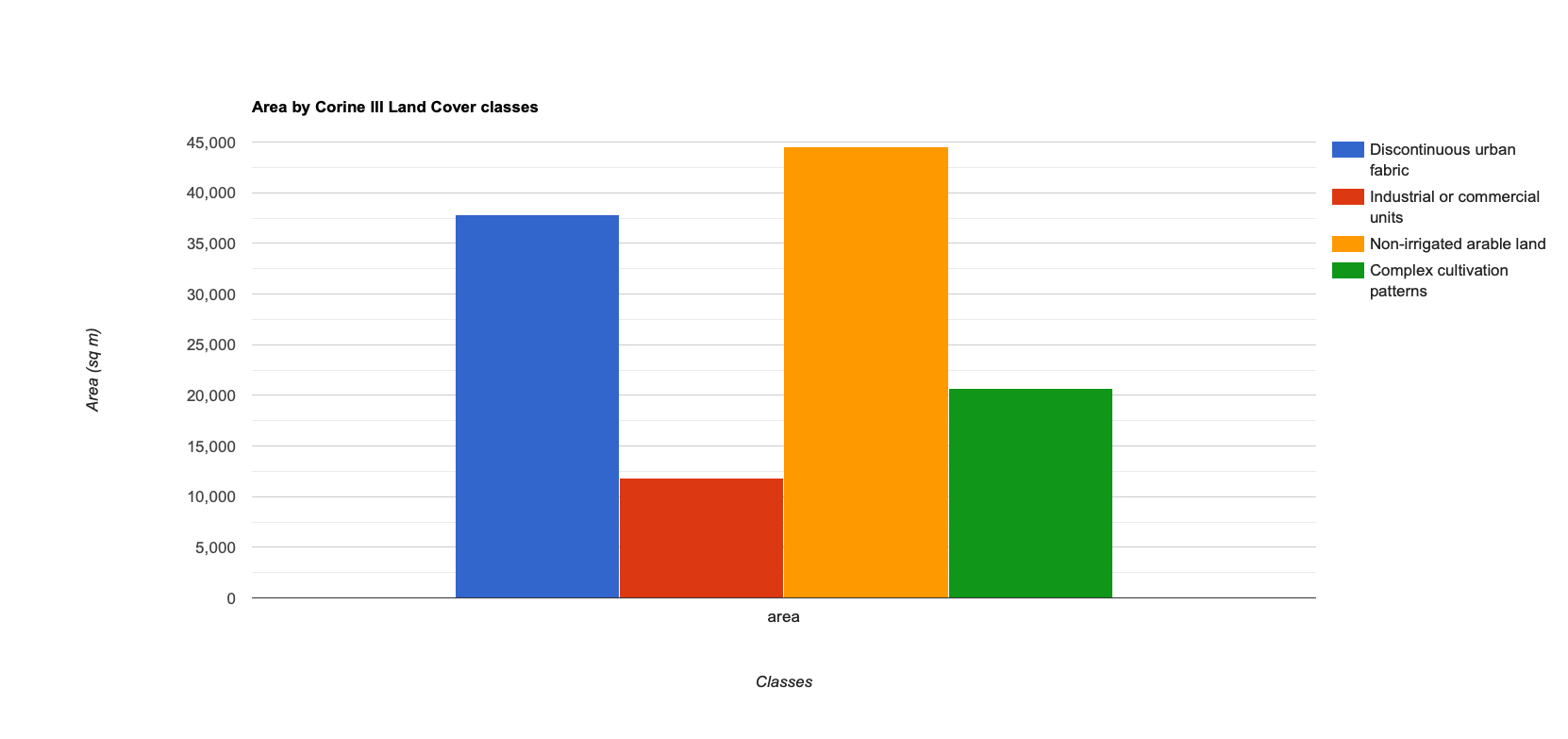
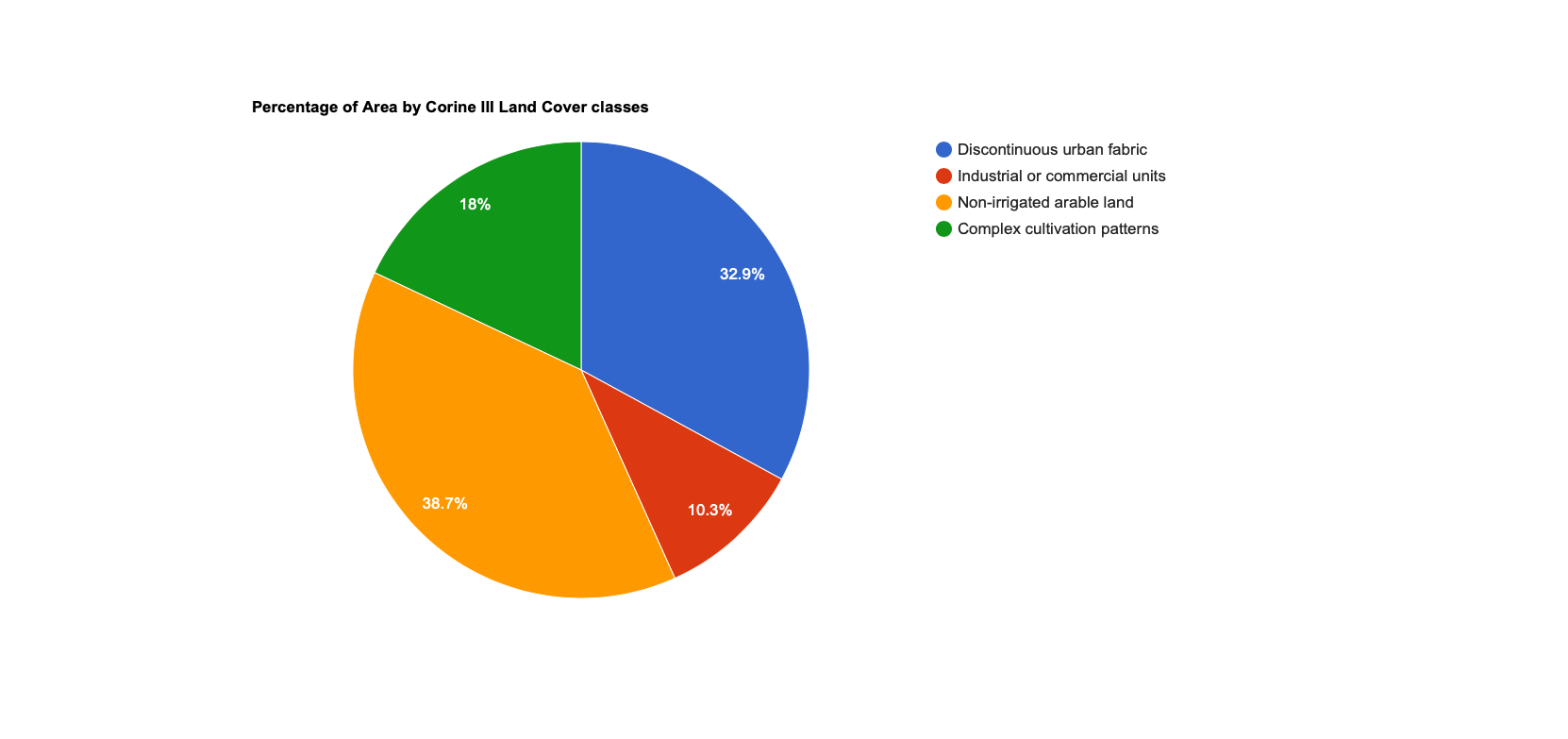
## Land use/land cover Classification

Image observations were used to generate the training and validation datasets. For the classification, as said in the previous section, a total of 108 training Samples were utilised. The five categories of LULC were Railway infrastructures, Tree Cover, Grass Cover, Bare soil and Shrub Cover. To identify these classes, therefore, the random forest strategy was effective both in terms of results and accuracy. Following the classification, the total area for each land cover class was calculated to facilitate a comprehensive analysis of the studied territory. It was observed that the primary class, Railway Infrastructures, was somewhat overestimated, reflecting a slightly inflated area compared to the actual coverage. In contrast, the other classes—Tree Cover, Grass Cover, Bare Soil, and Shrub Cover—exhibited relatively uniform values, indicating a consistent distribution of these land cover types across the study area.

The script employed for this classification is adaptable and can be modified according to specific study areas. Users can alter or add new classes to refine the classification process or to study different environments. This flexibility allows for tailored analyses and enhanced precision in land ****cover assessments.

## Corine Land Cover Level III Classification

An additional overview of the test area is given by considering the Copernicus Corine Land Cover Level III datasets, that contains 44 categorical values of Land Use (Copernicus, E.U., 2018). The standard CORINE Land Cover (CLC) project provides consistent information on land cover and land cover changes across Europe. The established CLC nomenclature includes 44 land cover classes, grouped in five main classes at level 1. Level 2 (15 classes) corresponds to physical and physiognomic entities at scales of 1:500 000 and 1:1 000 000 (‘urban zones’, ’forests’, ‘lakes’ etc.). Finally, level 3 is composed of 44 classes for use at a scale of 1:100 000 and higher (Heymann et al., 1994).

The total areas for each of the classes in the study area are calculated and displayed by the graphs below. Since the resolution of the data used is 100 meters, these outputs allow only a less detailed overview of the area under consideration, but they report information that could be useful for the formulation of sustainable land management and planning strategies, to provide the information elements to support decision-making processes.

# Bibliography

* Ahmed, S. A., & N, H. (2023). Land use and land cover classification using machine learning algorithms in google earth engine. *Earth Science Informatics*, *16*(4), 3057-3073.
* Claverie, M., Ju, J., Masek, J. G., Dungan, J. L., Vermote, E. F., Roger, J. C., ... & Justice, C. (2018). The Harmonized Landsat and Sentinel-2 surface reflectance data set. *Remote sensing of environment*, *219*, 145-161.
* Claverie, M., Ju, J., Masek, J. G., Dungan, J. L., Vermote, E. F., Roger, J.-C.,   Skakun, S. V., & Justice, C. (2018). The Harmonized Landsat and Sentinel-2   surface reflectance data set. Remote Sensing of Environment, 219, 145-161.
* Copernicus, E. U. (2018). Copernicus land monitoring service. Corine Land Cover data.
* Eisavi, V., Homayouni, S., Yazdi, A. M., & Alimohammadi, A. (2015). Land cover mapping based on random forest classification of multitemporal spectral and thermal images. *Environmental monitoring and assessment*, *187*, 1-14.
* Elith, J. (2019). 15-machine learning, random forests, and boosted regression trees. Quantitative analyses in wildlife science, page 281.Ahmed S. A. et al., 2023
* Geofabrik, D. S. (2017). Open Street Map
* Gorelick, N. (2013, April). Google earth engine. In EGU general assembly conference abstracts (Vol. 15, p. 11997). Vienna, Austria: American Geophysical Union.
* Guidonia Airport Meteorological Station
* Heymann, Y. (1994). *CORINE land cover: Technical guide*. Office for Official Publ. of the Europ. Communities.
* Phan, T. N., Kuch, V., & Lehnert, L. W. (2020). Land cover classification using Google Earth Engine and random forest classifier—The role of image composition. *Remote Sensing*, *12*(15), 2411.
* species presence-only data with random forests. Ecography, 44(12):1731–1742.
* *Tamiminia, H., Salehi, B., Mahdianpari, M., Quackenbush, L., Adeli, S., & Brisco, B. (2020). Google Earth Engine for geo-big data applications: A meta-analysis and systematic review. ISPRS journal of photogrammetry and remote sensing, 164, 152-170*
* Team, R. C. (2020). RA language and environment for statistical computing, R Foundation for Statistical. *Computing*.
* Valavi, R., Elith, J., Lahoz-Monfort, J. J., and Guillera-Arroita, G. (2021). Modelling
* Waske, B., & Braun, M. (2009). Classifier ensembles for land cover mapping using multitemporal SAR imagery. *ISPRS journal of photogrammetry and remote sensing*, *64*(5), 450-457.