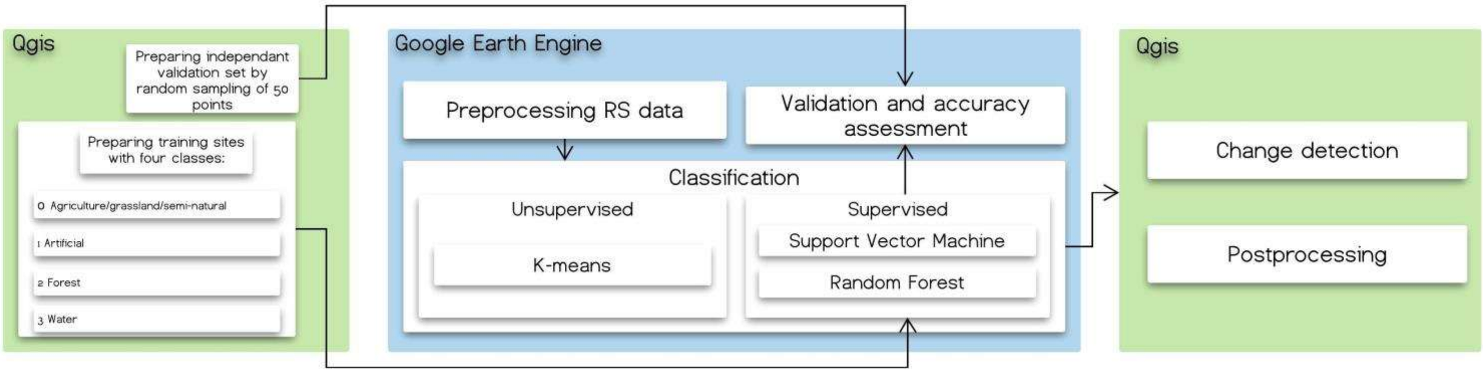
Assignment 1

Tools

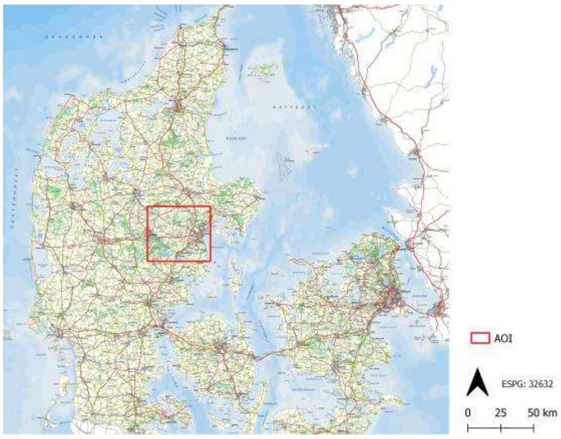
QGIS was used for data preprocessing such as creating training samples, preparing validation dataset with random points and the final visualization of results. For change detection, the semi-automatic classification plugin was used.

The classification task itself was conducted in Google Earth Engine. GEE is a cloud-based solution for accessing and processing satellite images, other remote sensing products and scientific datasets stored in GEE's data archive, it also has an option of uploading other datasets. The workflow for this assignment can be seen in figure below, and the GEE code is provided at the end of this document.



Landsat8 imagery and study area

A bi-temporal set of Landsat8 images was collected, consisting of two summer images, from 2013 and 2020 respectively, of the study area around Aarhus, as seen in figure to the right. The images were cloud-free at the area of interest, so no cloud-masking was performed. This is though rarely the case and normally cloud-free images are obtained from a stack of multiple images.



Classes and training/testing data

In order to classify the two images a number of classes should be selected. For this purpose, Corine Level 1 nomenclature was used with four following classes:

0. Agricultural, grassland and semi-natural areas
1. Artificial surfaces. It includes urban fabric, residential and industrial areas, roads etc.
2. Forest
3. Water

Reference training samples were collected by drawing polygons representative for each class on the background of the Danish Orthophotos in QGIS. 10 polygons per class were drawn, so that they were distributed throughout the entire study area.
The independent validation dataset consisting of 50 points was generated randomly in QGIS within the study area, and these ground truth points were assigned classes based on visual interpretation of orthophotos. The whole process resulted in four datasets: a training and validation dataset for 2013, and two corresponding datasets for 2020.

Classification

Band selection

For classification, bands 2-7 were selected. Furthermore, several indices, such as NDVI, NDBI and MNDWI, were added to this band collection to enhance different features – vegetation, build-up and water respectively. These are some of the common indices, but the land cover classification can be further expanded to include other indices.

NDVI, or Normalized Difference Vegetation Index, is used to describe the density of greenery and vegetation health. Vegetation strongly reflects in NIR and absorbs Red spectrum, so following formula is used for calculation: *NDVI = (NIR-Red)/(NIR+Red)*. The resulting values range from -1 to 1, where the higher values represent dense vegetation, while negative values indicate the absence of greenery.

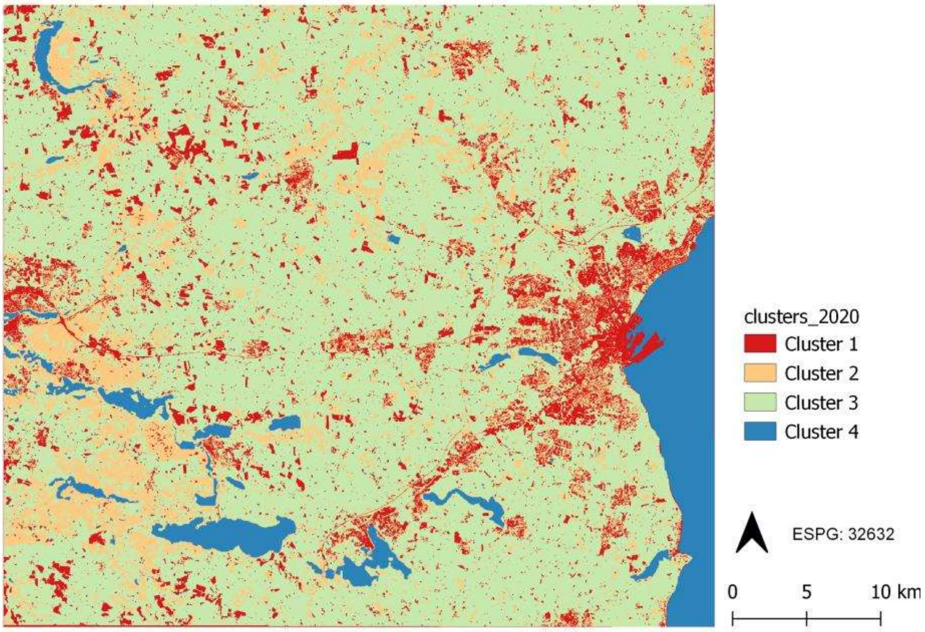
NDBI, or Normalized Difference Build-Up Index is used to enhance build-up areas. It is calculated using formula *NDBI = (SWIR-NIR)/(SWIR+NIR).* The resulting values also range from -1 to 1, where the higher values represent build-up areas, while negative values represent water.

The last index that was used MNDWI, or Modified Normalized Difference Water Index, enhances open water features, and distinguishes it from vegetation and build-up areas. It is calculated by *MNDWI = (Green - SWIR)/(Green + SWIR).*

As the values of the calculated indices range from -1 to 1, and all the bands were scaled to have values between 0 and 1, so that all the bands would have a common scale.

Unsupervised classification

It is recommended to run an unsupervised classification first to get an idea of spectral variation and of natural spectral groups, or clusters, in the image. The training data is not used in this classification type. The unsupervised algorithm was set to define four classes and the result classification is shown in the figure below. The clusters have no labels, but one can see that it identified artificial surfaces, water bodies, vegetation and presumably forest.

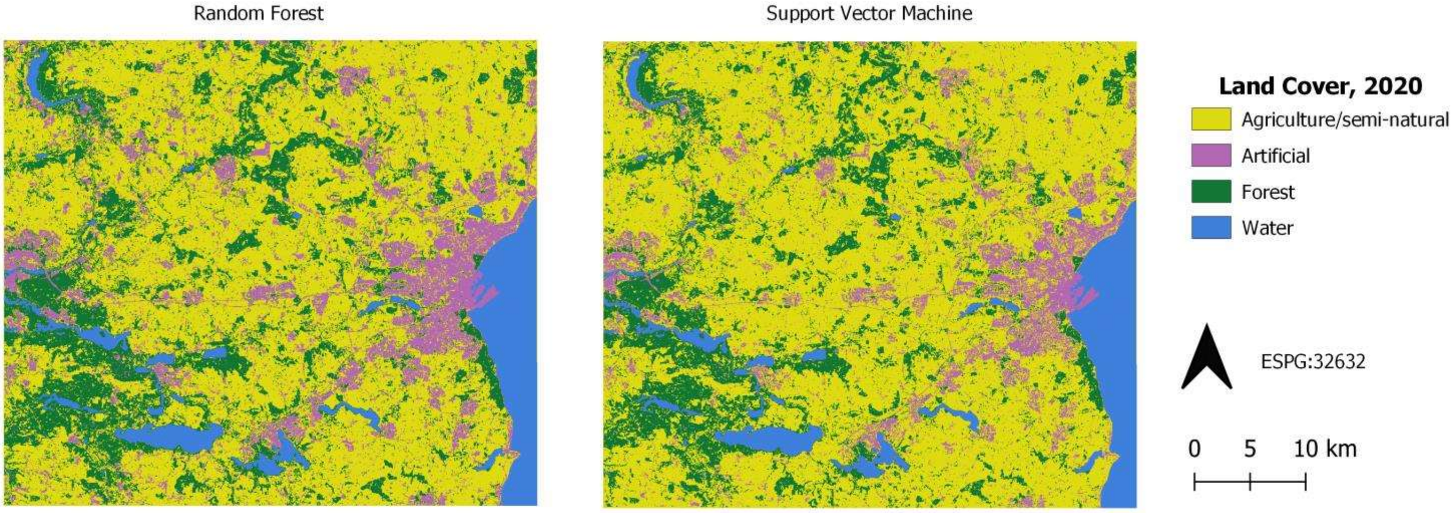


Supervised classification

The goal of supervised classification is to categorize all the pixels into the predefined land cover classes. For supervised classification two classifiers were used – Random forest and Support Vector Machine. A Random Forest classifier (RF) is a supervised learning classifier that uses decision trees to form an “ensemble” to make a 'decision forest’. Each of the individual decision trees votes about the outcome, and the result of the random forest is the class with the most votes.

Support vector machine (SVM) finds a decision boundary in a multidimensional space that distinctly separates the data into classes. The idea is to find such a boundary, also known as hyperplane, where the distance between data points of the given classes is the biggest.

Both classifiers were trained and validated in Google Earth Engine, and the visualization of the resulting land cover classification for 2020 can be seen below. It is noticeable, that RF classifies more pixels as artificial surfaces.



Validation and accuracy assessment

In order to assess the classification and look at how well it performs, an error matrix was produced in GEE along with the overall accuracy, user’s and producer’s accuracy and Kohen’s Kappa. Overall accuracy is the total number of the elements that were correctly classified divided by the total number of the ground truth samples. Both RF and SVM achieved the overall accuracy of 84% in land cover classification for 2020, while Cohen’s Kappa, which measures the agreement between the predicted and the actual class, is around 0.7, which indicates a substantial agreement. It can be said, that our classification is around 70% better than a classification resulting from random assignment of pixel classes. Overall accuracy and Kappa differ in values, because the overall accuracy includes only diagonal elements without the errors of omission/commission, while Kappa includes non-diagonal elements from the error matrix.

Producer’s accuracy is the number of sites classified correctly divided by the total number of validation sites for that class. It indicates how often the actual features will be correctly depicted on the land cover map, while user’s accuracy indicates how often the land cover class on the produced map will actually be found in the reality.

From the error matrix one can see, that artificial surfaces are confused with agriculture/semi-natural vegetation areas and forest is confused with agriculture. An explanation can be, that the pixels used for training were mixed, as vegetation is often represented in urban fabrics.

The producer’s accuracy for artificial surfaces, produced by SMV, is only 50%, which means that only in 50% of the cases artificial surfaces will be correctly shown on the land cover map, while the RF classification of the artificial class has user's accuracy of 57%, meaning that it might exaggerate actual occurrences of artificial surfaces. This might explain why the SVM map, seen in the figure below, shows fewer urban/artificial pixels than the RF map.

Random Forest 2020

Overall accuracy 84%		Predicted				
Kappa Index 0,71		**Agriculture**	**Artificial**	**Forest**	**Water**	**Producer's accuracy %**
Actual	**Agriculture**	28	3	1	0	88
	Artificial	2	4	0	0	67
	Forest	1	0	6	0	86
	Water	0	0	1	4	80
	User's accuracy %	90	57	75	100	

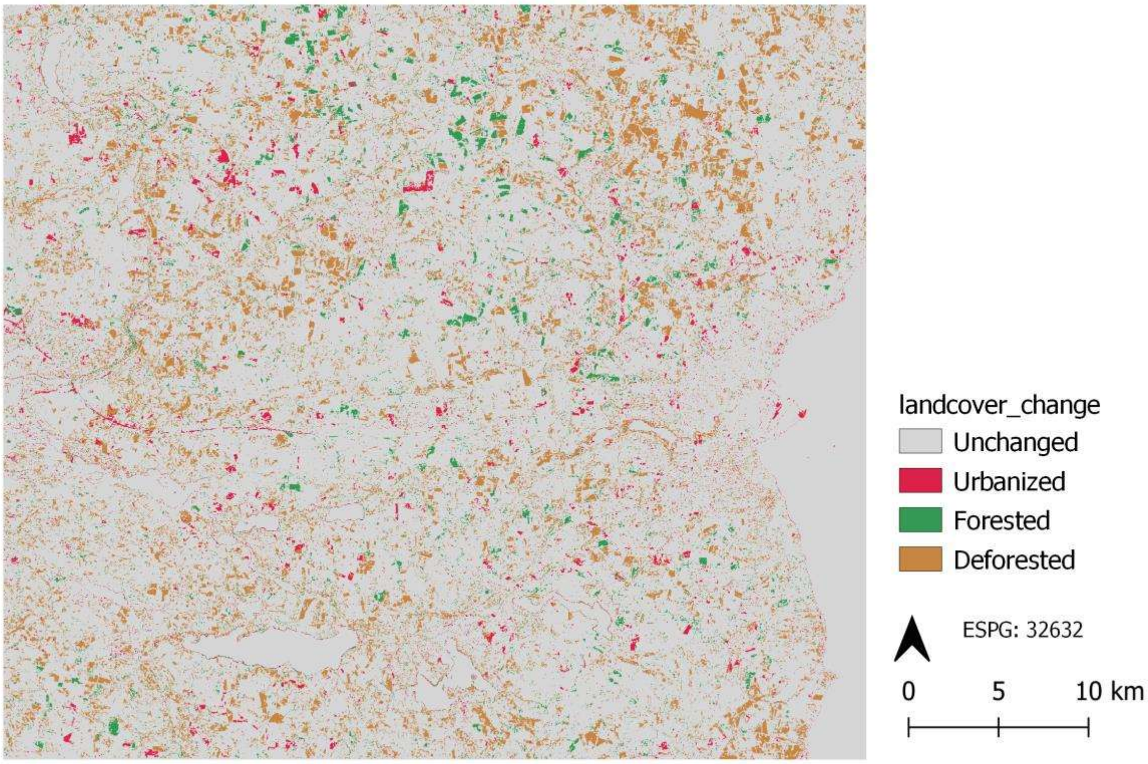
Support vector Machine 2020

Overall accuracy 84%		Predicted				
Kappa Index 0,69		**Agriculture**	**Artificial**	**Forest**	**Water**	**Producer's accuracy %**
Actual	**Agriculture**	30	1	1	0	94
	Artificial	3	3	0	0	50
	Forest	2	0	5	0	71
	Water	0	0	1	4	80
	User's accuracy %	86	75	71	100	

One can also notice some unbalance in the number of validations points per class, where there are 32 points representing the agricultural class, and only 5 points representing the water class. The simple random sampling method, that was used, cannot guarantee that all classes are represented, so the suggestion is to use stratified random sampling approach, where reference points are placed randomly within each class. We can also normalize the absolute sample counts by the map area by calculating a new error matrix with estimated area proportions of each class.

Land cover change

Finally, the land cover change map, as seen in the figure below, was created using Semi-automated classification plugin in QGIS. The map has captured broad land cover changes, for example the expansion of the harbor area in Aarhus and the construction of new roads and urban areas. But one should keep in mind that the error contribution comes from the two classifications – from 2013 and 2020, and the visual comparison of the changes and the orthophotos from 2013 and 2020 has shown, that many of the changes are the results of misclassifications. An alternative to our post-classification approach, where two images were classified independently, can be direct change classification, where a single classification is run on the combined dataset for the two images.



The difficulty level of the task is 5 out of 10.