Classification of satellite images using Dynland technology

Martins Pukitis*[a], Ints Mednieks[a]
[a]Institute of Electronics and Computer Science, 14 Dzerbenes str., Riga LV1006 Latvia

ABSTRACT

Satellite images are widely used for land cover classification into predefined categories. To perform it in a practicable way, field data are needed to train and test a classifier. The collection of accurate field data requires significant resources, and image data of a category may have sophisticated distribution in the feature space. "Dynland" technology was developed to overcome these problems and offer a robust and practical way to classify such images. It is based on a nonparametric clustering method applicable to data with different distributions with a following assignment of classes to clusters based on their overlapping with available reference data which can be scarce and imprecise.

Dynland clustering algorithm provides meaningful clusters but requires huge computational resources currently limiting the size of processed images. To overcome this problem, we propose to apply a sampling procedure i.e. use each n-th pixel on both axes for clustering, and distribute the remaining pixels to created clusters based on nearest neighbors' search. However, such an approach reduces the classification accuracy; the analysis results of this reduction will be presented.

Another way how to deal with images of large areas is to split them into fragments. To classify a fragment, reference data for all categories of interest should be available for the area it represents. This limitation should be resolved so that the available reference data are used for areas not limited to one fragment. To deal with that, a method is proposed allowing to exploit the classification result of a neighboring fragment for which the reference data were available in full.

The proposed solutions widened the applicability of the Dynland technology e.g. enabled the production of a classified map of mires and peatlands for the whole territory of Latvia.

Keywords: Image classification, land cover classification, image clustering, Dynland technology

1. INTRODUCTION

Dynland is a semi-supervised technology which is aimed at practicable classification of land cover from Sentinel-2 imagery [1]. It is based on a nonparametric clustering algorithm applicable to data with different distributions with a following assignment of classes to clusters based on their overlapping with available reference data which can be scarce and imprecise. Dynland technology enables the use of machine learning methods in applications where it was not possible before due to the lack of sufficient training data. It is generic enough to add value also in other industries, not only in most geospatial image classification tasks. Until now it has been applied in classifying multispectral imagery into general land cover classes, and for analysis of forests, peatlands and mires.

The data clustering algorithm is the core of the Dynland technology. This algorithm unites similar data units and combines them into the group based on their similarity. Algorithm operates with missing data, excludes multicollinearity, and thus can simplify and explain the data structure. Algorithm is non-parametric through the full process – similarity/distance calculation and uniting data in clusters based on their similarity. As a result, an algorithm can capture extremely complex data forms.

Ninth International Conf. on Remote Sensing and Geoinformation of the Environment (RSCy2023), edited by K. Themistocleous, et al., Proc. of SPIE Vol. 12786, 127860E · © The Authors. Published under a Creative Commons Attribution CC-BY License · doi: 10.1117/12.2681896

This algorithm on the contrary of the other popular unsupervised classification methods performs cleaning of the resulting clusters as the clusters can contain false positives (pixels erroneously included in the cluster) and false negatives (pixels erroneously not included in the cluster).

Current implementation of the DynLand clustering algorithm requires huge computational resources. Using the desktop computer with 64GB of RAM it is feasible to perform clustering on images with up to 500 000 pixels. Larger images can be processed by splitting them up into manageable fragments. Alternatively, we propose using down-sampling of the image with some clustering (overleap) step. Clustering step value of n means that the clustering algorithm is only applied on every n-th pixel along each axis. Afterwards, the rest of the pixels are added to initial clusters based on spectral similarity. However, such an approach reduces the classification accuracy; the effects of this reduction are described in Section 3.2.

Splitting the large image into fragments may cause another problem in classification. To classify a fragment, reference data for all categories of interest should be available for the area it represents. If that is not ensured, the algorithm for assignment of classes to clusters is unable to reveal clusters related to the classes not represented in the reference data. A method is proposed dealing with this problem and allowing to exploit the classification result of a neighboring fragment for which the reference data were available in full. That is described in Section 3.3.

2. DATA

We have applied the Dynland technology mostly to classification of freely available Sentinel-2 data [2]. To illustrate the influence of the clustering step value, three Sentinel-2 images from different areas of Latvia were chosen with sizes close to the upper limit of the clustering algorithm. Ten bands of the Sentinel-2 multispectral image were used with 10m and 20m resolution. Level 1C data were used. The images are illustrated in Figures 1, 2 and 3.

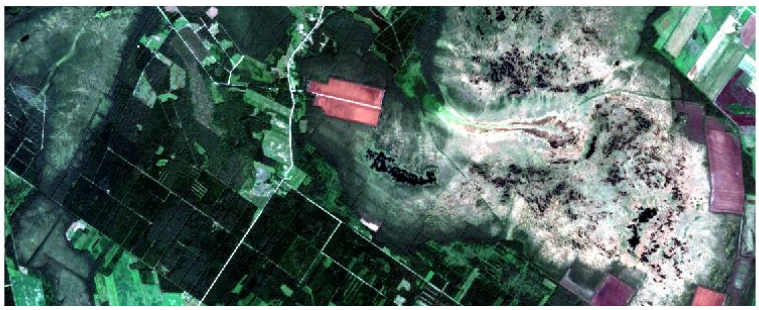

Figure 1. **Cenu** bog area true color Sentinel-2 image taken on 1st August, 2018. Center coordinates 56.86°N, 23.81°E.

Figure 2. **Dobele** area true color Sentinel-2 image taken on 1st August, 2018. Center coordinates 56.44°N, 22.82°E.

Figure 3. **Zakumuiza** area true color Sentinel-2 image taken on 10th September, 2019. Center coordinates 56.97°N, 24.43°E.

Reference data were combined from several sources including 2018 Corine Land Cover [3] (CLC) data on man-made features and agriculture areas, 2018 Copernicus Forest Type HRL [4] data as well as information from the Latvian Peat Association on abandoned and licensed peat extraction sites in 2018, and data from the Nature Conservation Agency database "Ozols" on bog habitats in 2021. An image with reference data combined from these sources for the Cenu bog area is presented in Figure 4.

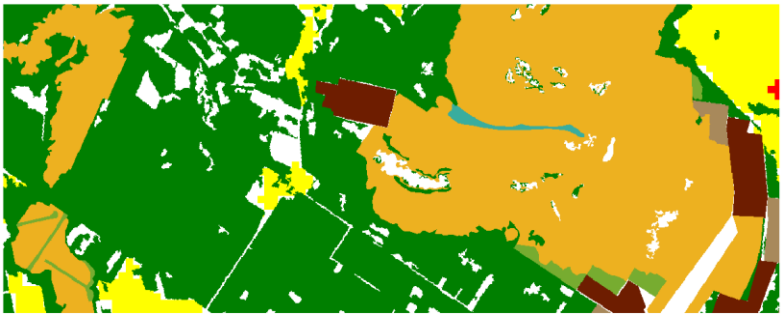

Figure 4. Reference image for the Cenu bog area (yellow: agriculture, green: forests; orange: active raised bogs; light green: degraded raised bogs; light blue: transition mires and quaking bogs; brown: licensed peat extraction sites; tan: abandoned peat extraction sites).

For the analysis of the missing reference problem, two neighboring images of the Kaigu bog were used, illustrated in Figure 5.

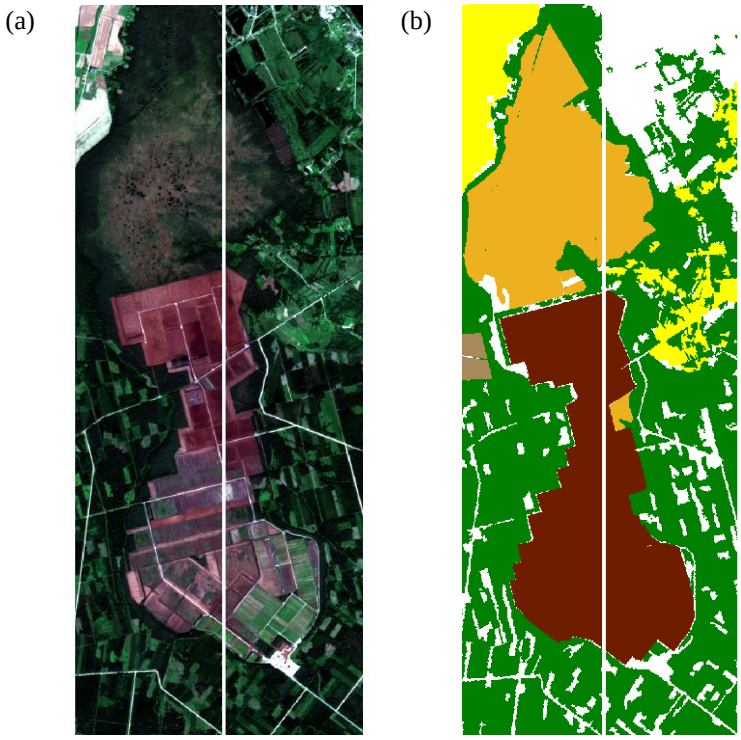

Figure 5. (a) **Kaigu** bog area true color Sentinel-2 image. Center coordinates 56.86°N, 23.81°E. (b) Reference image of the Kaigu bog area (yellow: agriculture, green: forests; orange: active raised bogs; brown: licensed peat extraction sites; tan: abandoned peat extraction sites).

Both images to be processed in a combined way were obtained in the same conditions; actually, they are parts of the same Sentinel-2 tile image. That is crucial to ensure the proper operation of the clustering algorithm.

3. ANALYSIS

3.1. Classification algorithm

Dynland clustering algorithm includes the following main steps:

1) Forming clusters around each pixel including 10 nearest neighbors in the multispectral feature space
2) Uniting clusters with significant overlap (>80%)
3) Eliminating clones to leave only unique clusters
4) Repeating 2) and 3) until there are no changes
5) Performing growth of clusters by adding nearest neighbors
6) Performing cleaning of the set of clusters leaving only one instance of multiple similar clusters
7) Repeating 5) and 6) until there are no changes in the set of clusters

The algorithm is nonparametric; it is described in detail in the paper [1]. As a result, all pixels of the image are distributed into clusters of different size and shape in the 10-dimensional feature space related to 10 Sentinel-2 bands.

Classes (categories) are assigned to each cluster using the following algorithm taking the image, clusters and reference at the input.

Let n be the number of categories, $R_i, i \in [1, n]$ reference of the category i within the image, m - number of clusters and $C_j, j \in [1, m]$ set of pixels in cluster j. We will use notation $\|S\|$ for number of pixels in set S.

Classification happens in 2 steps. For some cases where additional analysis is required (Section 3.3), only the first step is used. We will use terms *initial* classification (step 1) and *full* classification (step 2).

If a cluster overlaps with reference set of exactly one category, it is obvious choice to assign that category to all pixels in such cluster.

If a cluster overlaps more than one set of reference, the category it overlaps the most should be assigned: $\arg\max_j \frac{\|C_i \cap R_j\|}{\|C_i\|}$. If the reference set is untrustworthy (imprecise and/or outdated), this category is only assigned if at least a certain fraction of a cluster overlaps the reference of the category (initial classification threshold is used).

For the second step we are left with clusters that do not overlap any reference at all. We use pixels from already assigned clusters to assign temporary category labels to all unassigned pixels based on spectral similarity. Any supervised algorithm can be used for this step; we used KNN [5]. Afterwards we treat those temporary labels as reference set for clusters that didn't overlap actual reference: $\arg\max_j \frac{\|C_i \cap L_j\|}{\|C_i\|}$, where L_j are temporary labels of category j.

3.2. Clustering step analysis

The Dynland clustering algorithm starts with forming a cluster around each pixel and operates with the cluster-pixel matrix indicating which pixels are put in which cluster(s). Such approach requires huge memory resources limiting the number of pixels to be clustered. Although it is using sparse matrices to operate with such data (implementation in MATLAB [6]), memory limitations strongly affect the applicability of the algorithm. To deal with this problem, we propose to use a sampling procedure taking each *n*-th pixel from the image on both axes, perform clustering of these pixels and distribute other pixels within the formed clusters afterwards. The procedure used for distribution was KNN looking for one nearest neighbor using the Euclidian distance. It is understandable that the nonparametric nature of the Dynland clustering is compromised in this way and that results in lower accuracy of classification.

The approach described above was used to create 9 different clustering results – each obtained using a different clustering (overleap) step (using every pixel, using every 2nd, 3rd, …, 9th). Each of these clustering results was then assigned categories using a classifier described in Section 3.1. Classification was performed using 80% of reference pixels from each category of the reference (remaining 20% were used for validation).

Classification results were compared to reference, evaluated visually and by calculating the Cohen's kappa coefficient. Visual analysis shows that classification is degraded even with the clustering step value of 2 (see Figure 6). If n=4, one of the smallest categories (transition mires and quaking bogs) is completely lost in the label image.

Kappa values for classification decrease with each increase of the clustering step (see Figure 7). The image of Dobele has the least decrease of kappa values for different clustering steps; for other 2 images, this decrease is more significant.

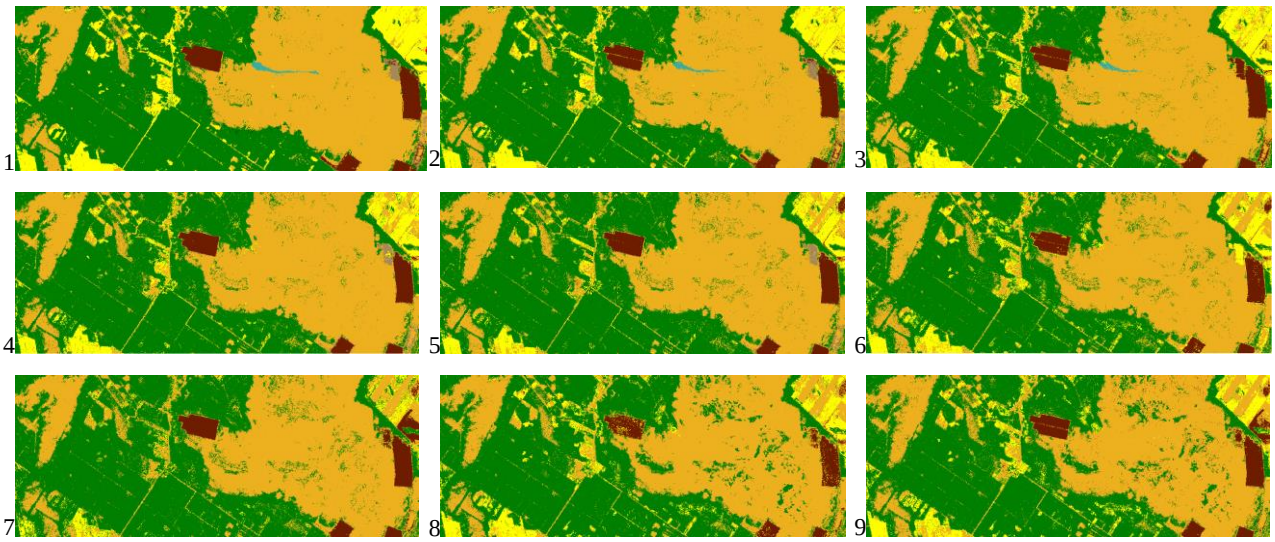

Figure 6. Clustering step analysis results- classified image for the Cena bog area (yellow: agriculture, green: forests; orange: active raised bogs; light green: degraded raised bogs; light blue: transition mires and quaking bogs; brown: licensed peat extraction sites; tan: abandoned peat extraction sites). Clustering step is indicated to the left of each image.

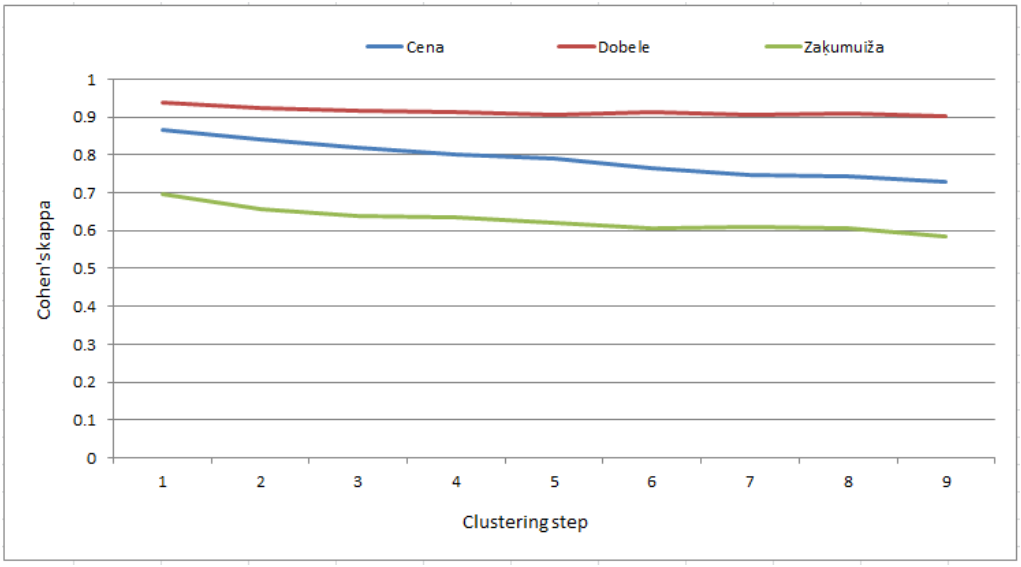

Figure 7. Clustering step analysis results.

3.3. Methods for compensation of missing reference

If the classification is performed for an image fragment where the reference data is missing for some category, the resulting label image will not include pixels of that category. If that happens, and the category is actually present within the fragment area, reclassification is needed to obtain the properly classified image.

Two reclassification methods are proposed and were analyzed. They are based on combined processing of two images: 1) the target image is the one which is classified now and for which some reference data are missing; 2) the

supplementary image is the one which can be classified properly because the reference data are available in full (for all categories).

In the first approach, initial classification is performed for the target image. The result depends on the threshold value setting the minimum overlap of the cluster with reference for assigning a category to the cluster (we will call it "initial classification threshold"). After that, remaining clusters of the target image are assigned categories based on their similarity with classified pixels of the supplementary image.

The second approach uses initial classification of the supplementary image (again obtained using different thresholds), classified pixels of categories for which the reference data are missing in the target image are added to the target image to be classified along with their classification result to be used as the reference). Newly constructed image is then clustered and the full classification is obtained.

In order to effectively analyze reclassification performance, both images (with full reference) were tested by artificially removing one reference category from the target image and using the other as the supplementary image, repeating that for every category. After reclassification, its results were compared against the full reference of the image to calculate the kappa coefficient. The results are illustrated in Figures 8 and 9.

For the first method, increase of the initial classification threshold mostly affected reclassification due to missing reference data for the forest class. For most other class exclusions classification result is nearly constant (Figure 8). There is a significant variation between exclusion of different classes.

Classification results for the second method are virtually unaffected by the initial classification threshold and having very little variation between different class exclusions (Figure 9).

3.4. Discussion

As expected for the clustering step analysis, best classification results were obtained if clustering of every pixel was used, gradually declining with every increase of the clustering step. The best kappa values and the least decrease for the image of Dobele are explained by only having 3 categories in the reference of that image (other images had 6 – 7 categories).

Accuracy increase with increasing initial classification threshold for the first reclassification method is explained by outdated reference available for this image. It mostly affects the forest class; it is assumed that clear cuts of the forest have occurred so that the reference data are outdated.

4. CONCLUSIONS

The main outcomes of this study can be summarized as follows:

- The Dynland clustering algorithm can be applied to large images using the clustering (overleap) step for skipping pixels on both axes. Classification accuracy is decreasing gradually with every increase of this step. It is not recommended to use a clustering step higher than 2.
- The second reclassification approach (adding classified pixels from supplementary image to the target image to be classified) is a preferable way for solving the missing reference problem when the Dynland classification technology is applied. Care should be taken to ensure that both target and supplementary images are obtained in the same sensing conditions. Actually they should be fragments of the same Sentinel-2 tile image to ensure proper reclassification.

5. ACKNOWLEDGMENT

This study was supported by the ERDF-funded project entitled "Remote sensing based system for forest risk factor monitoring (Forest Risk)" (Project No. 1.1.1.1/21/A/40).

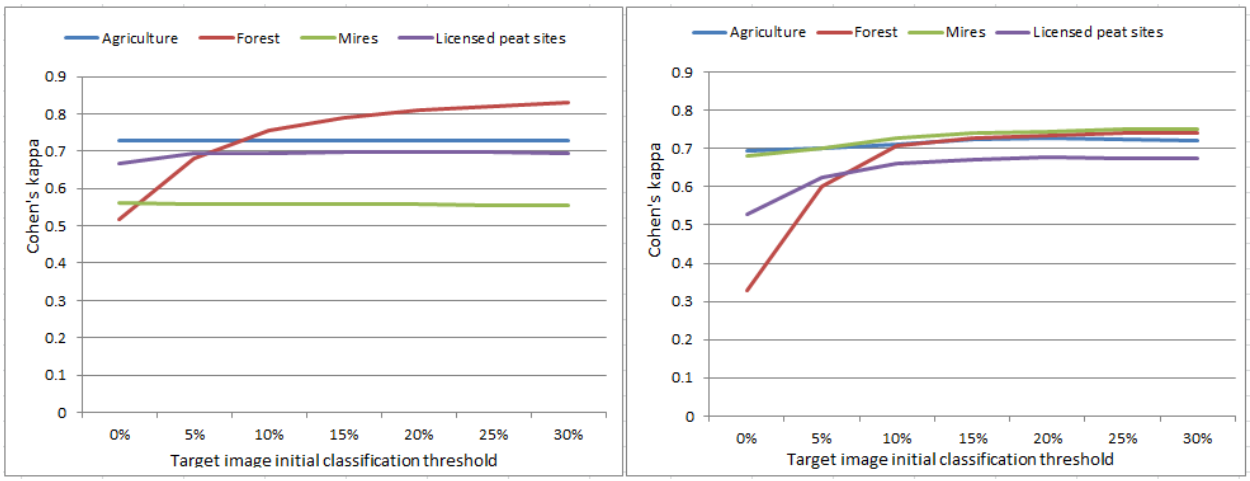

Figure 8. Classification accuracy of the first method for west (left) and east (right) images.

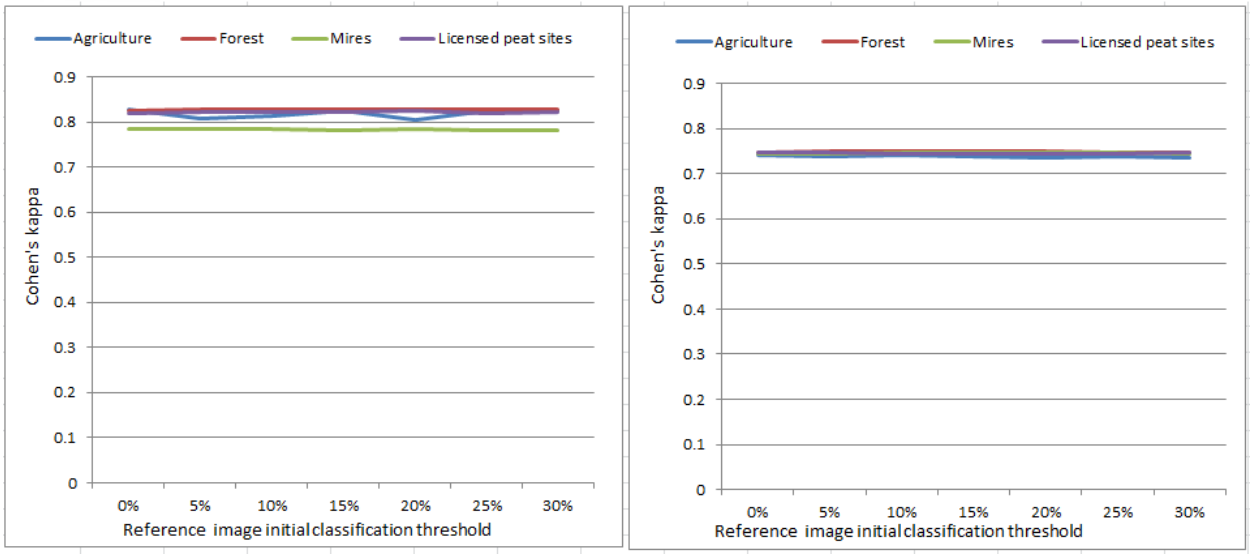

Figure 9. Classification accuracy of the second method for west (left) and east (right) images.

6. REFERENCES

[1] R. Dinuls and I. Mednieks, “Nonparametric classification of satellite images”, in Proc. of the 2018 International Conference on Mathematics and Statistics, Porto, Portugal, July 15-17, pp. 64-68.

[2] “Sentinel-2”. Accessed March 31, 2023. https://eos.com/find-satellite/sentinel-2/ .

[3] “CLC2018”. Accessed March 31, 2023. https://land.copernicus.eu/pan-european/corine-land-cover/clc2018

[4] “Forest Type 2018”. Accessed March 31, 2023. https://land.copernicus.eu/pan-european/high-resolution-layers/forests/forest-type-1/status-maps/forest-type-2018

[5] Cover, Thomas M.; Hart, Peter E. (1967). "Nearest neighbor pattern classification". IEEE Transactions on Information Theory. 13 (1): 21–27.

[6] “MATLAB”. Accessed March 31, 2023. https://www.mathworks.com/products/matlab.html