AutoCoast

Trustworthy Unsupervised ML Model for Drawing Coastlines and Creating Benchmark Dataset

Helmholtz Project: Autocoast

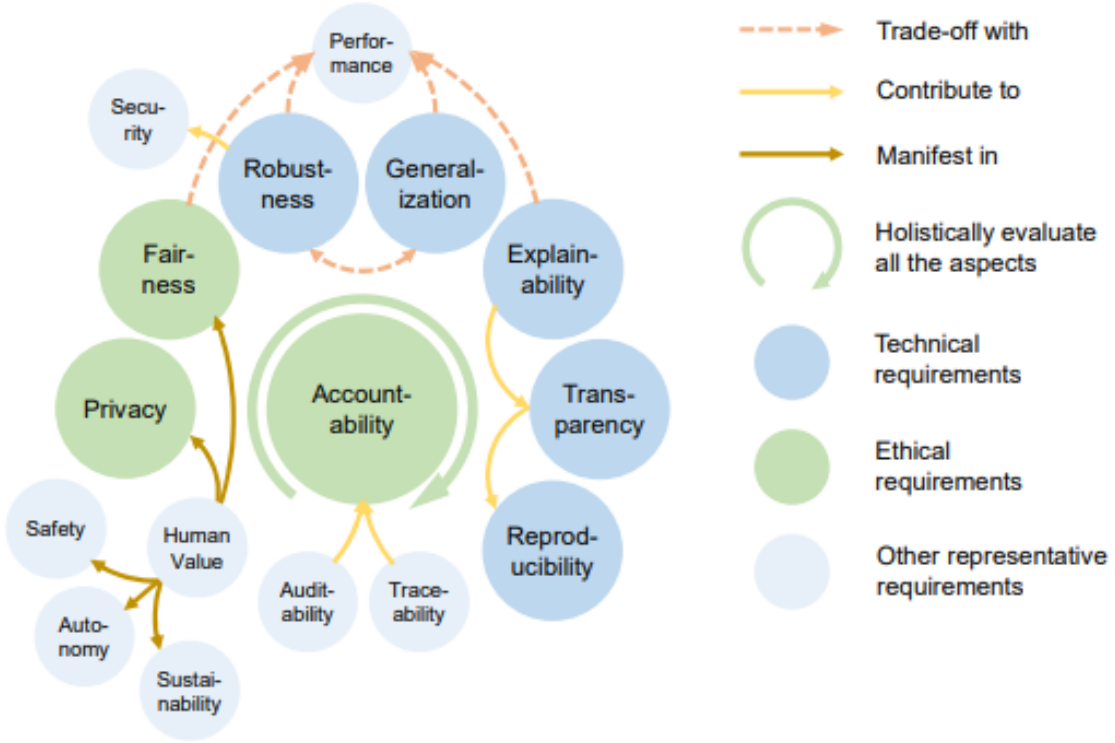
Chandrabali Karmakar[1], David Pogorzelski[2], Peter Arlinghaus[3], Andres Camero[4], Wenyan Zhang[5]

[1,4] German Aerospace Center (DLR)

[2,3,5] Institutes of the Helmholtz-Zentrum Hereon

Trustworthy AI – a coastal use case

- Coastal line detection and change monitoring is a vital question to ensure safety of lands, humans
- Not yet an established trustworthy AI model for the detection of coastal changes
- The Autocoast project aims to have a consolidated framework for drawing coastlines as well as detecting coastline changes
- Openly available satellite data (Sentinel-1, Sentinel-2, Landsat) are used



The problem at hand – Automatic drawing of coastlines for large datasets

- We need to draw the coastline in Sentinel-2 dataset at the Baltic sea
- We do not know the dataset
- How many classes are there
- Given n classes, how many are of interest while drawing the coastline ?
- How to reduce human labor in labelling classes using machine learning?
- At the same time, can we also improve the model from human inputs?

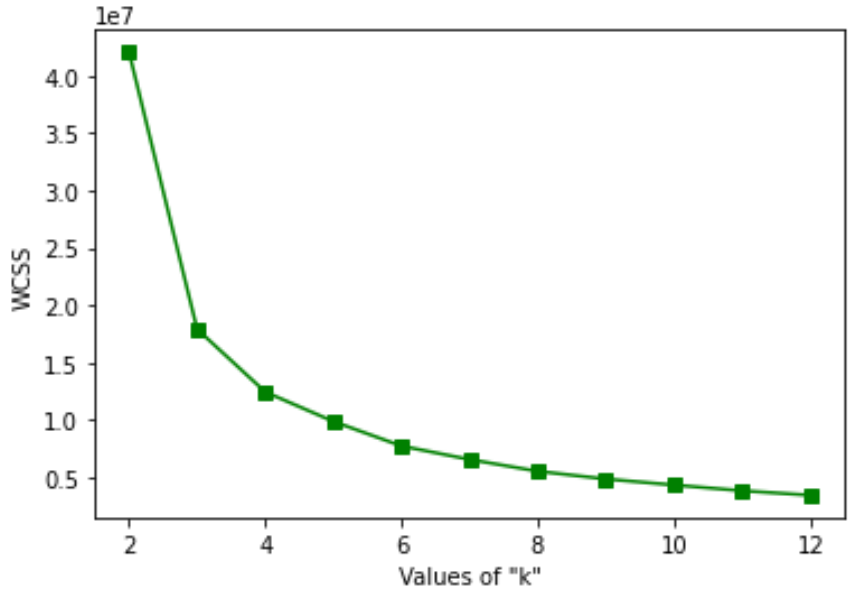
Finding the optimal number of clusters in the dataset

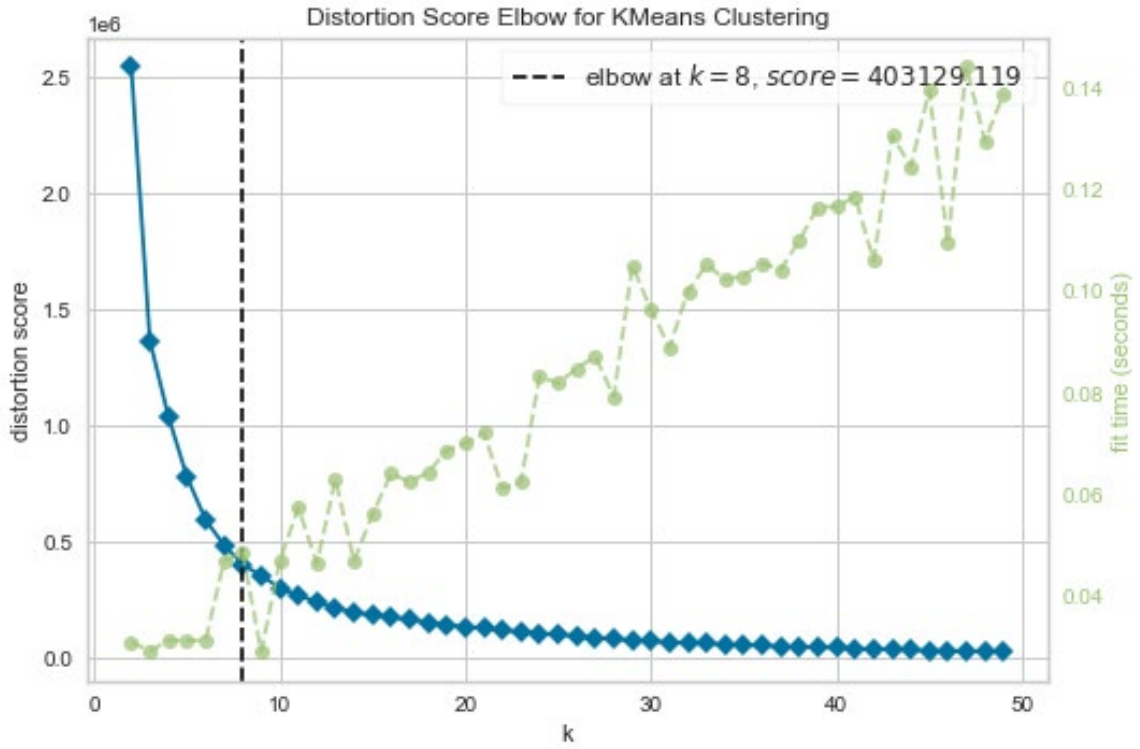
- A general objective
- Several clustering algorithms are to be compared
- A balance between accuracy and resource consumption is desirable
- Starting with dummy data(dynamicworld) for coastal region, will switch to real dataset (coasts around Baltic sea) in future

- Predictions are trustworthy, UQ methods to be applied

Initial checks –ELBOW plots

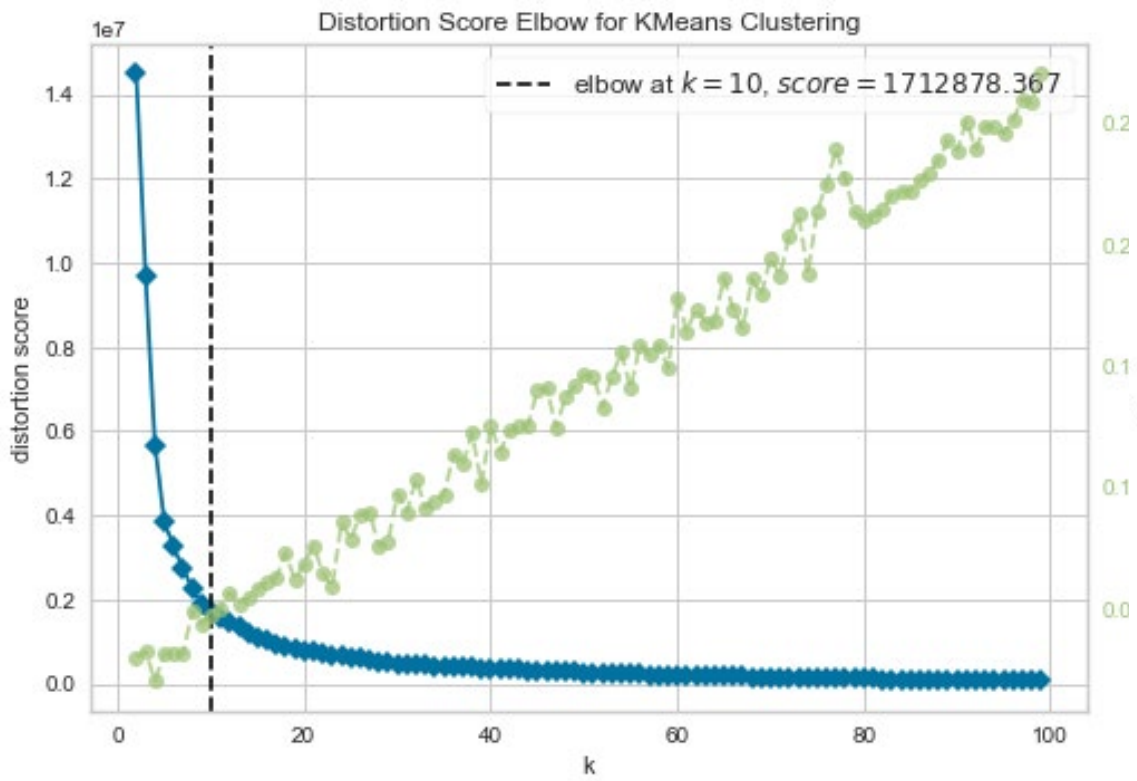
$$WCSS = \sum_{i \in n} (X_i - Y_i)^2$$



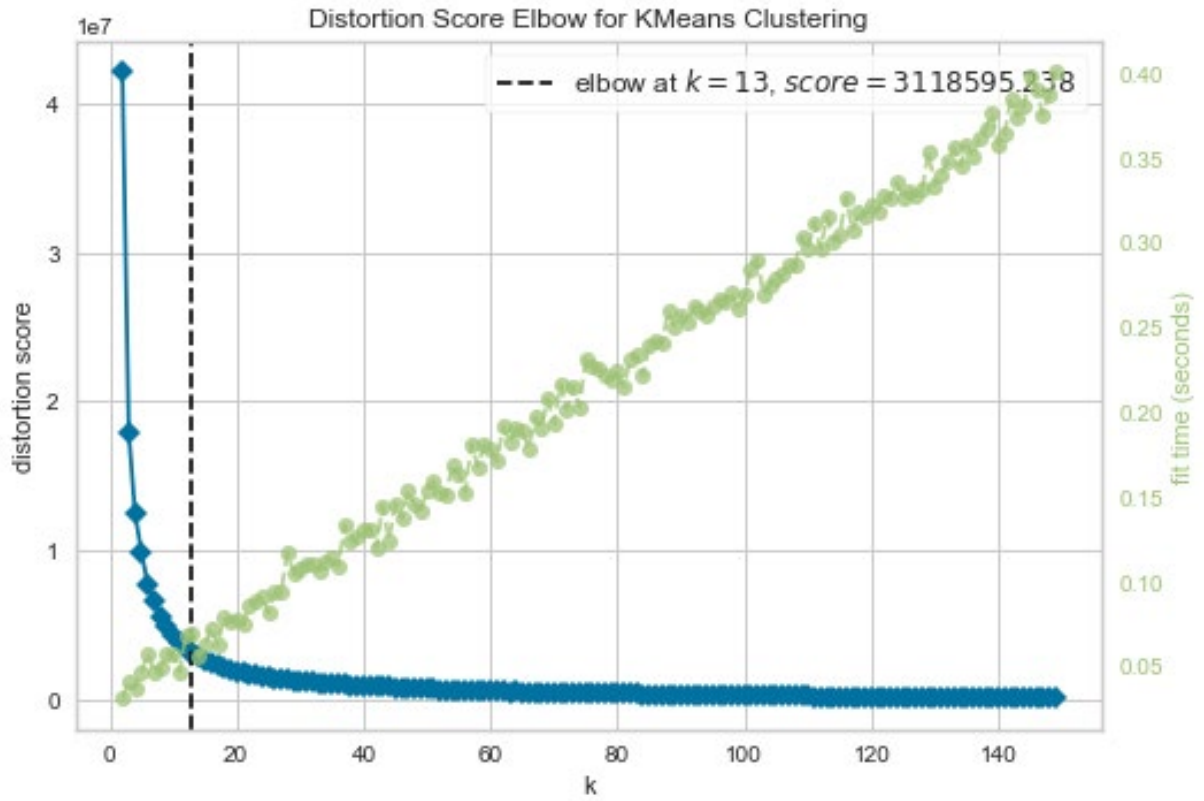


100 samples, max number of clusters: 50 (n/2)

Initial checks –ELBOW plots



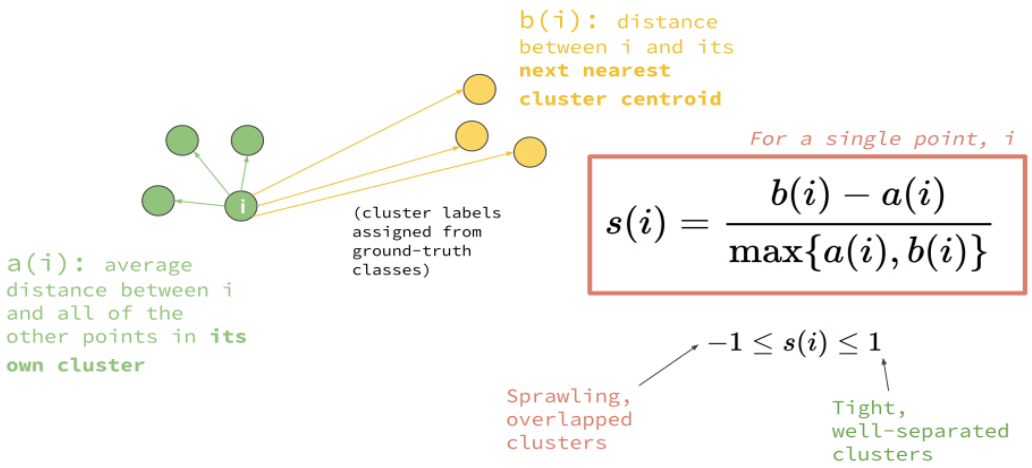
200 samples, max number of clusters: 100 (n/2)



300 samples, max number of clusters: 150 (n/2)

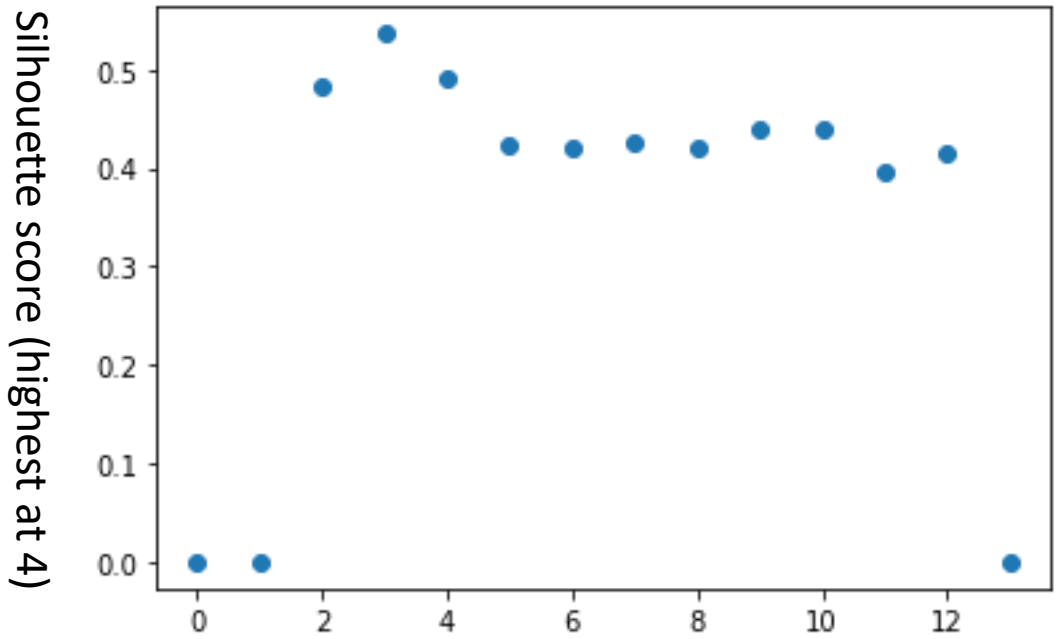
Silhouette score

The Silhouette Coefficient is calculated using the mean intra-cluster distance (a) and the mean nearest-cluster distance (b) for each sample. The Silhouette Coefficient for a sample is (b - a) / max(a, b). To clarify, b is the distance between a sample and the nearest cluster that the sample is not a part of. Note that Silhouette Coefficient is only defined if number of labels is 2 <= n_labels <= n_samples - 1.
A high value indicates that the object is well matched to its own cluster and poorly matched to neighboring clusters.



Peter J. Rousseeuw (1987). “Silhouettes: a Graphical Aid to the Interpretation and Validation of Cluster Analysis”. Computational and Applied Mathematics 20: 53-65.

Silhouette score



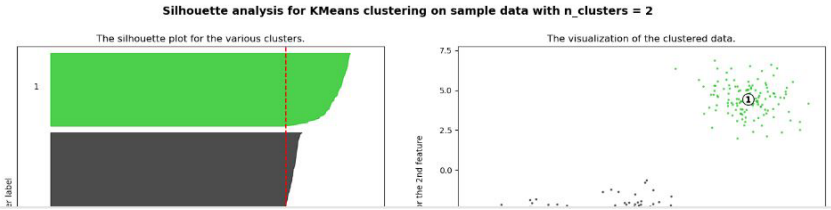
Selecting the number of clusters with silhouette analysis on KMeans clustering

Silhouette analysis can be used to study the separation distance between the resulting clusters. The silhouette plot displays a measure of how close each point in one cluster is to points in the neighboring clusters and thus provides a way to assess parameters like number of clusters visually. This measure has a range of [-1, 1].

Silhouette coefficients (as these values are referred to as) near +1 indicate that the sample is far away from the neighboring clusters. A value of 0 indicates that the sample is on or very close to the decision boundary between two neighboring clusters and negative values indicate that those samples might have been assigned to the wrong cluster.

In this example the silhouette analysis is used to choose an optimal value for `n_clusters`. The silhouette plot shows that the `n_clusters` value of 3, 5 and 6 are a bad pick for the given data due to the presence of clusters with below average silhouette scores and also due to wide fluctuations in the size of the silhouette plots. Silhouette analysis is more ambivalent in deciding between 2 and 4.

Also from the thickness of the silhouette plot the cluster size can be visualized. The silhouette plot for cluster 0 when `n_clusters` is equal to 2, is bigger in size owing to the grouping of the 3 sub clusters into one big cluster. However when the `n_clusters` is equal to 4, all the plots are more or less of similar thickness and hence are of similar sizes as can be also verified from the labelled scatter plot on the right.



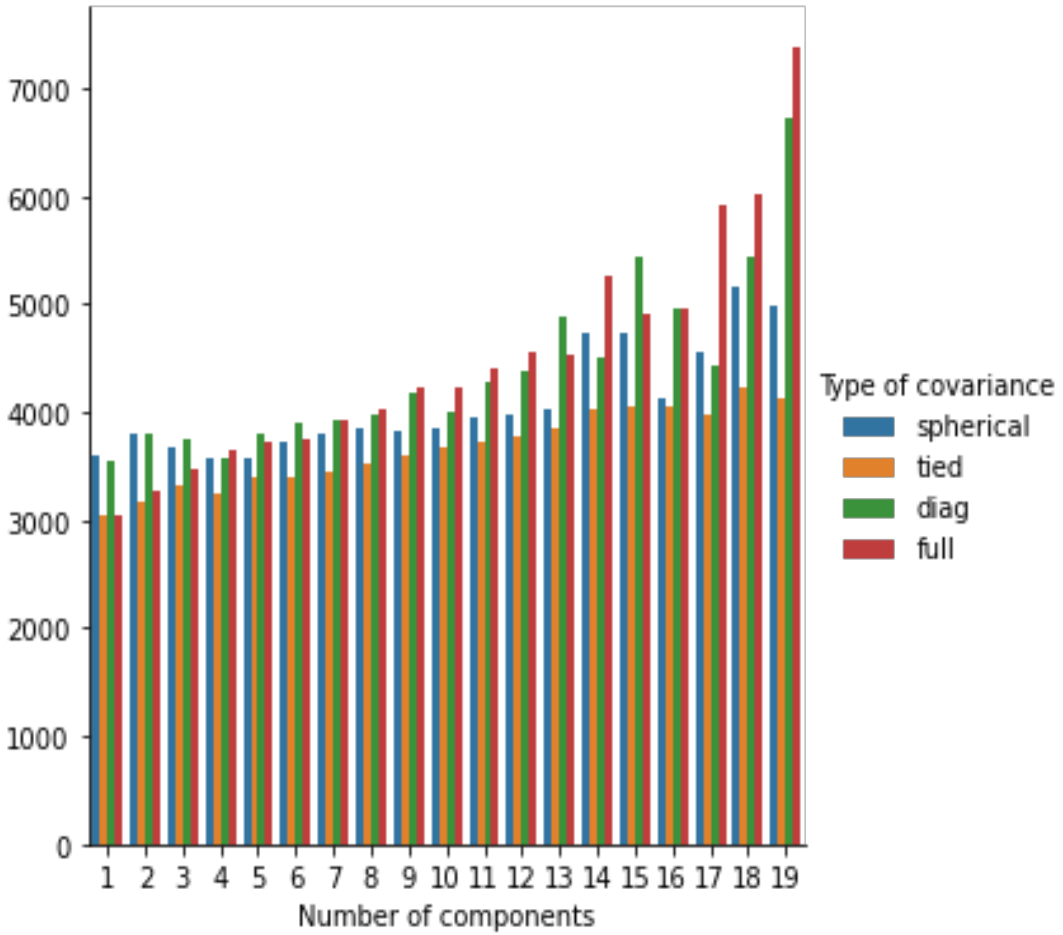
The silhouette ranges from −1 to +1, where a high value indicates that the object is well matched to its own cluster and poorly matched to neighboring clusters.

Bayesian information Criteria – a more reliable estiamate

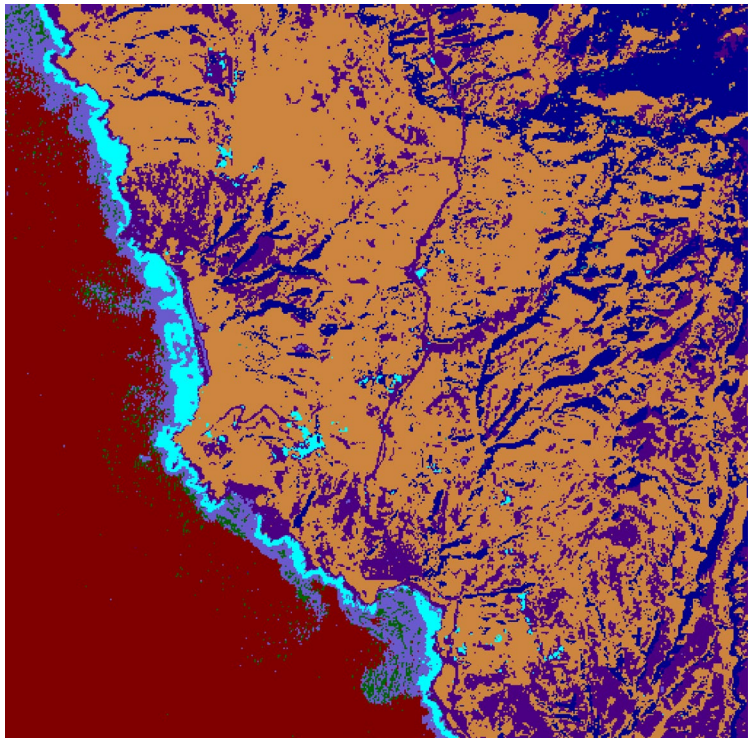
$$\mathrm{BIC} = k\ln(n) - 2\ln(\widehat{L}).$$

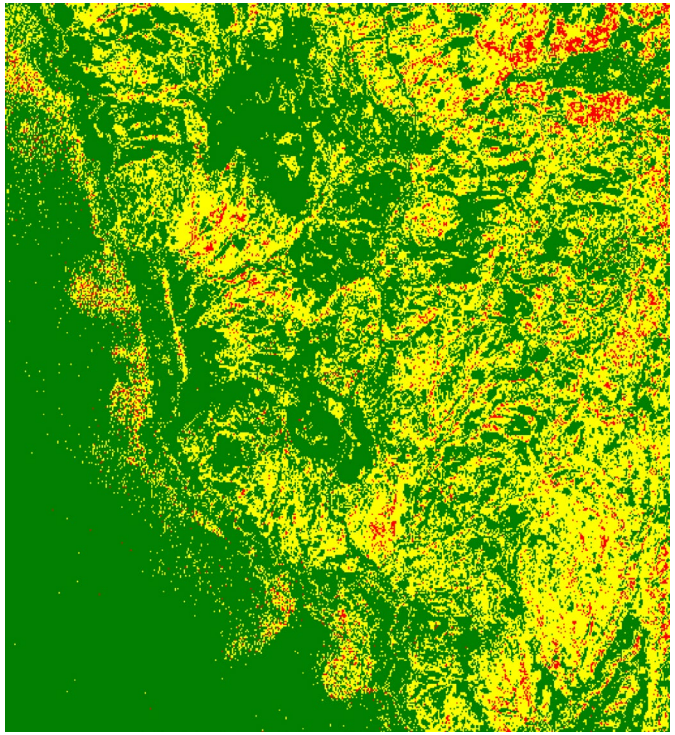
where

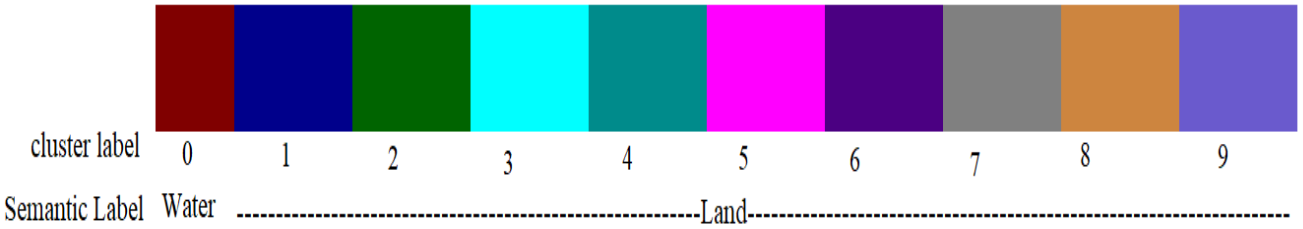
- $\hat{L}$ = the maximized value of the likelihood function of the model M, i.e. $\hat{L} = p(x \mid \hat{\theta}, M)$, where $\hat{\theta}$ are the parameter values that maximize the likelihood function;
- x = the observed data;
- n = the number of data points in x, the number of observations, or equivalently, the sample size;
- k = the number of parameters estimated by the model. For example, in multiple linear regression, the estimated parameters are the intercept, the q slope parameters, and the constant variance of the errors; thus, $k = q + 2$.

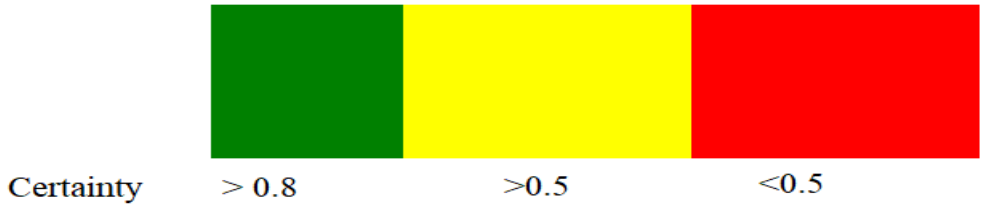


Uncertainty aware segmentation with soft clustering

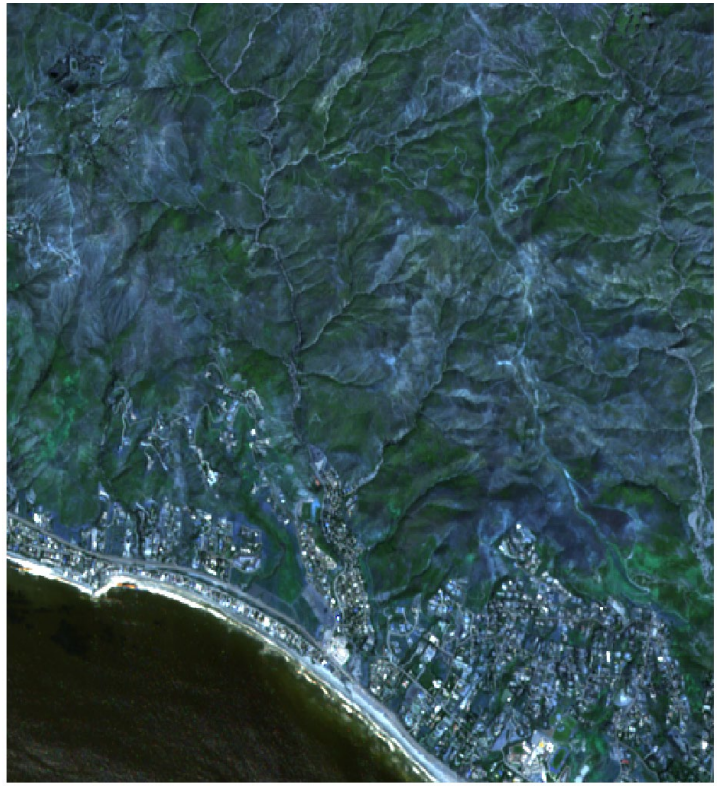


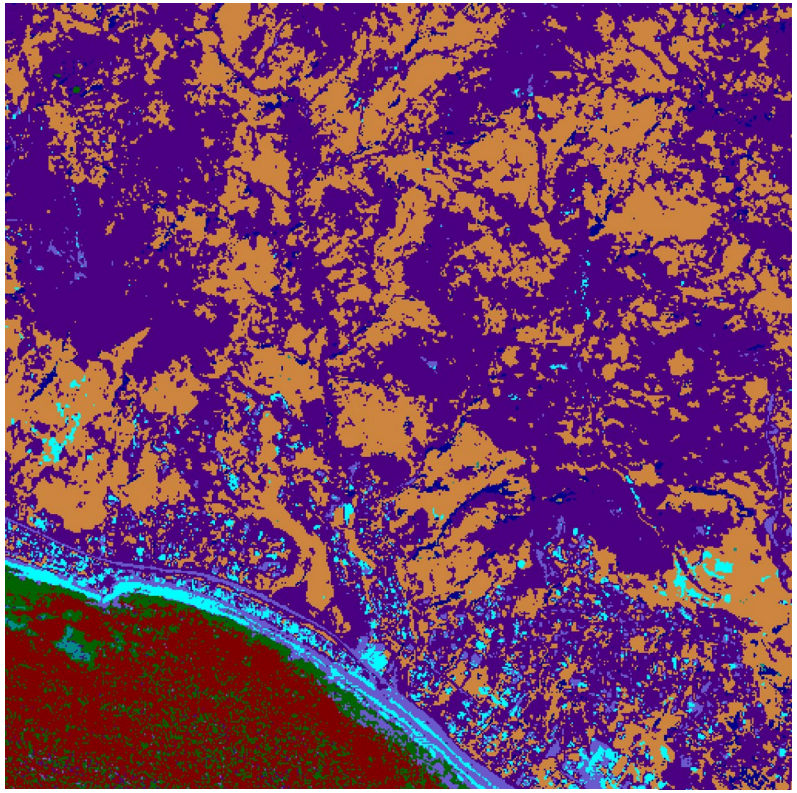


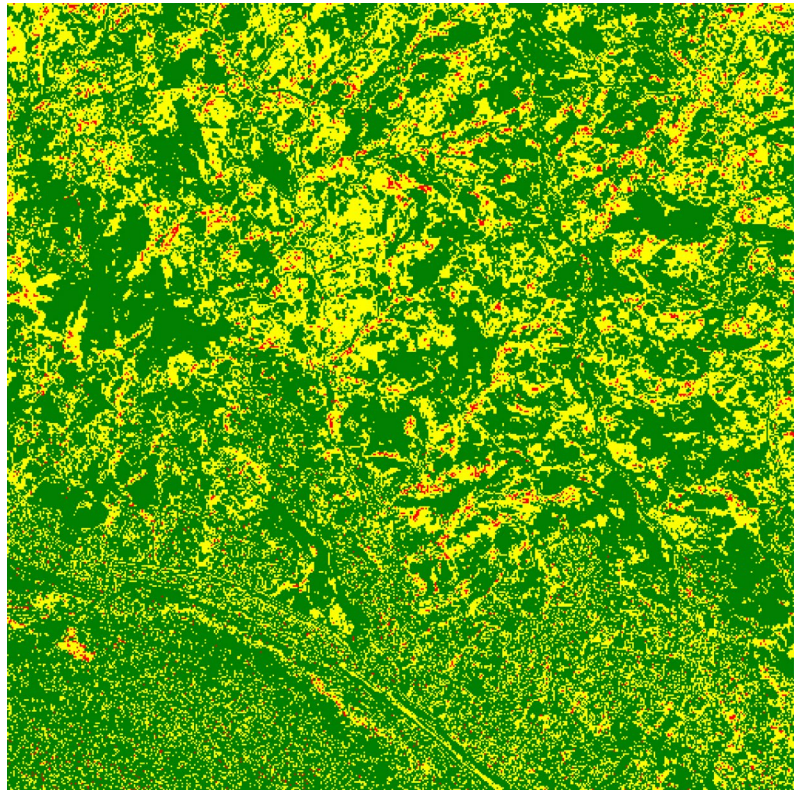


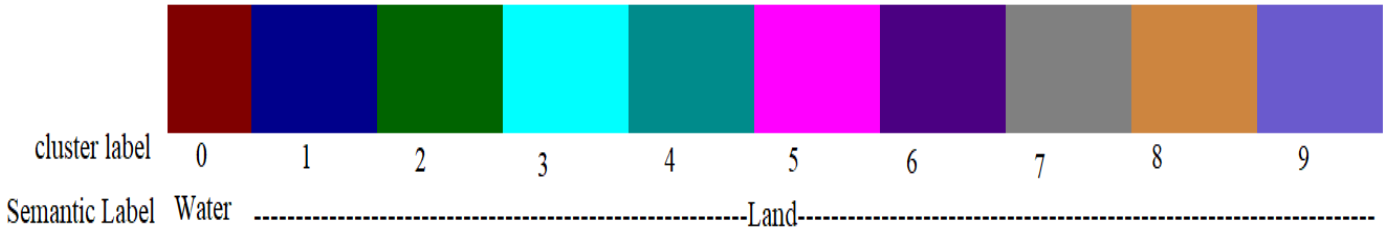


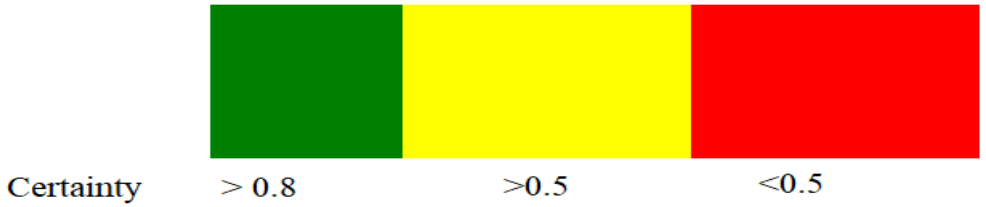
Uncertainty aware segmentation with soft clustering



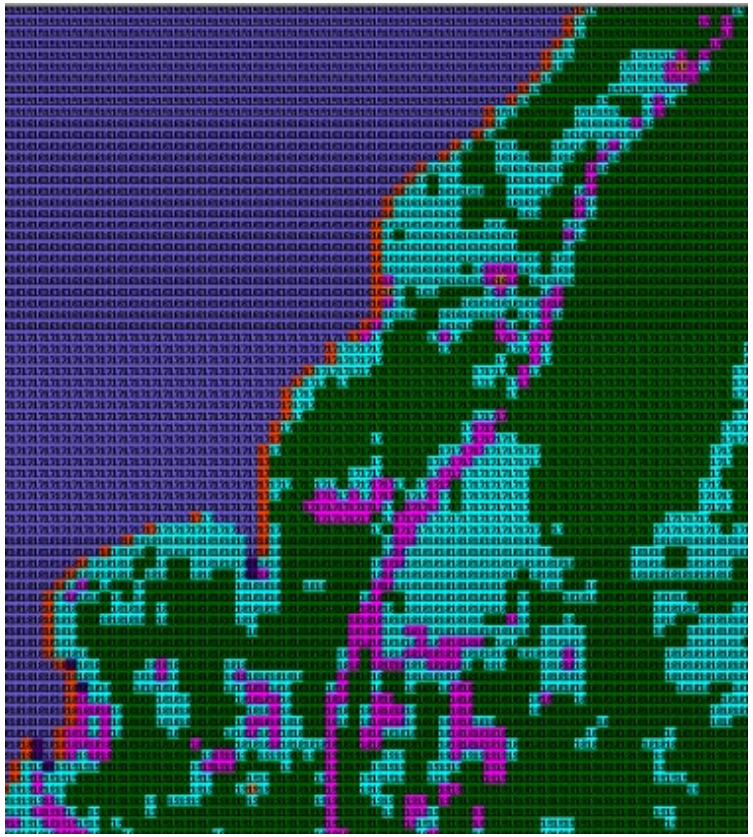


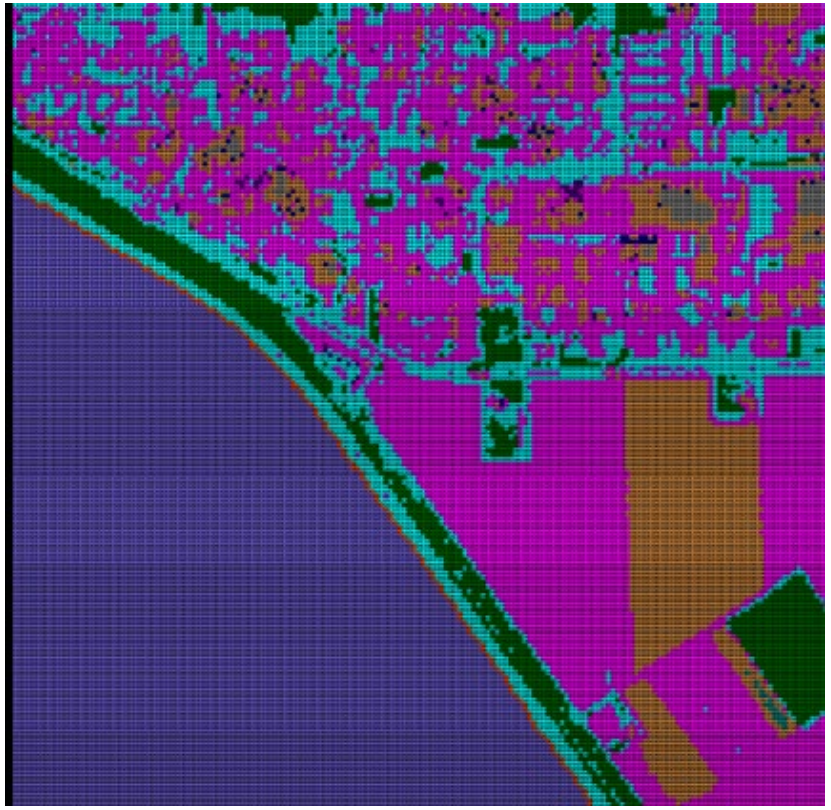


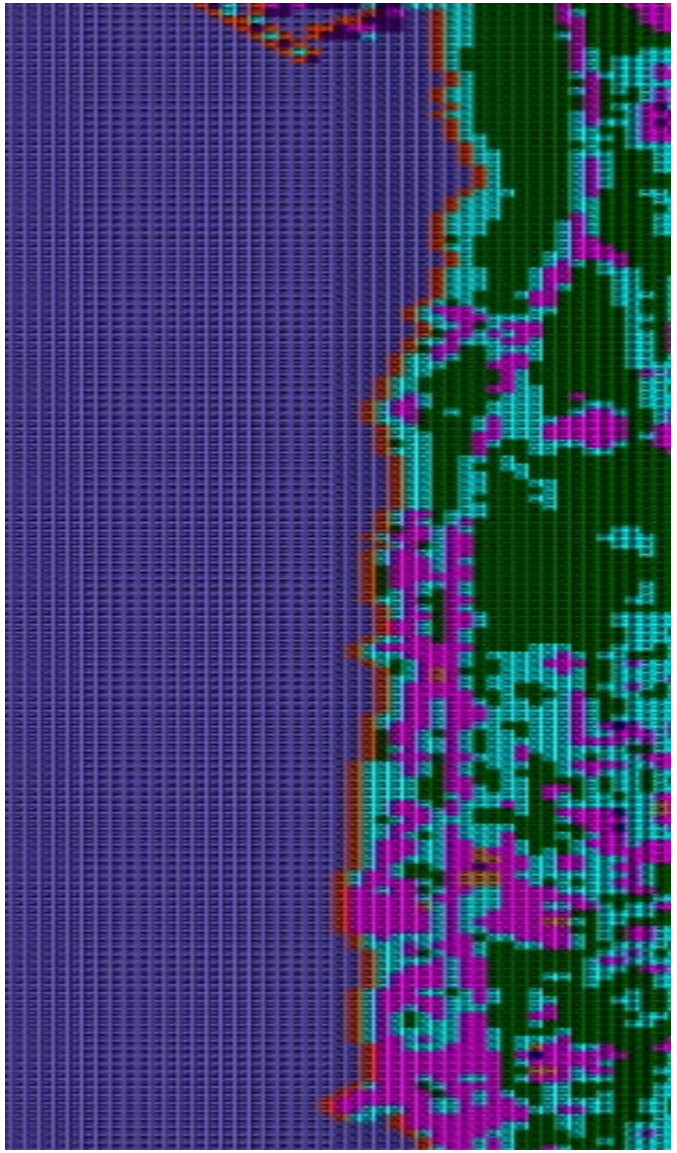




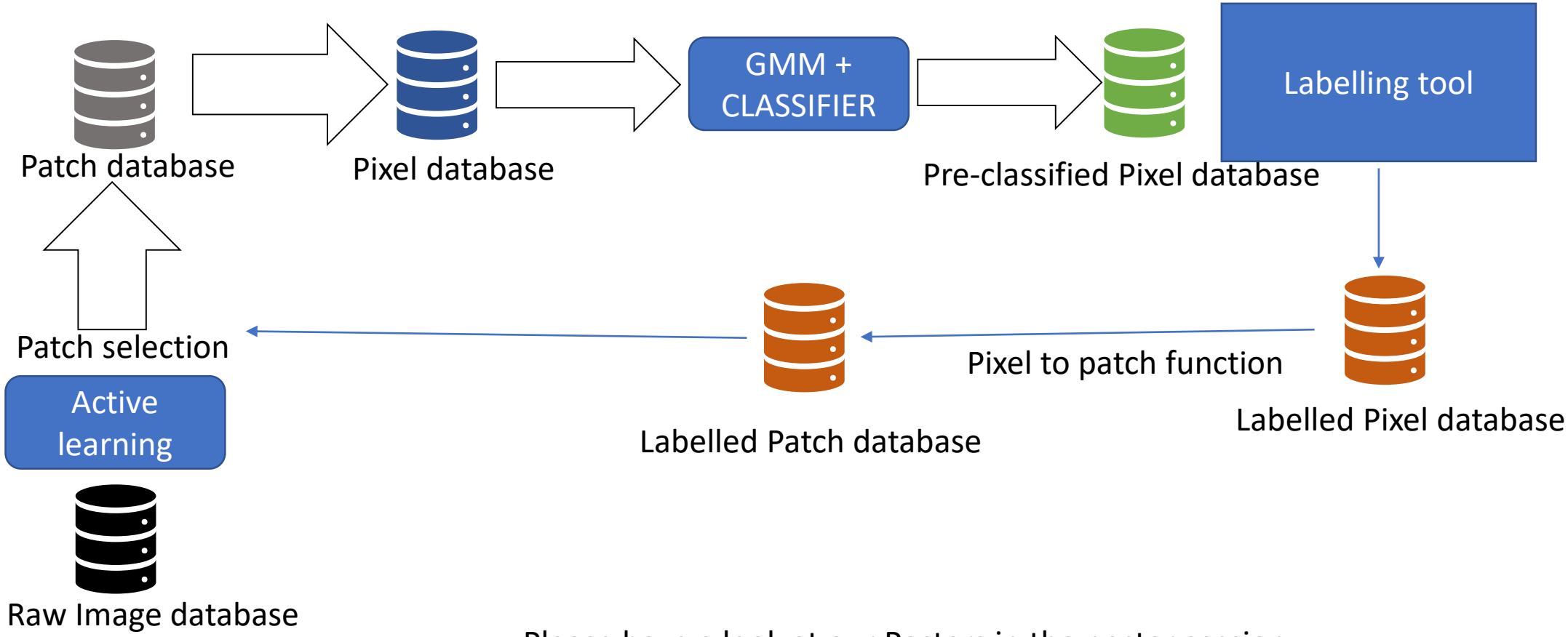
Coastline detection







The total workflow: more details in our Posters



Please have a look at our Posters in the poster session

Thank you

For questions, suggestions please feel free to contact

chandrabali.karmakar@dlr.de