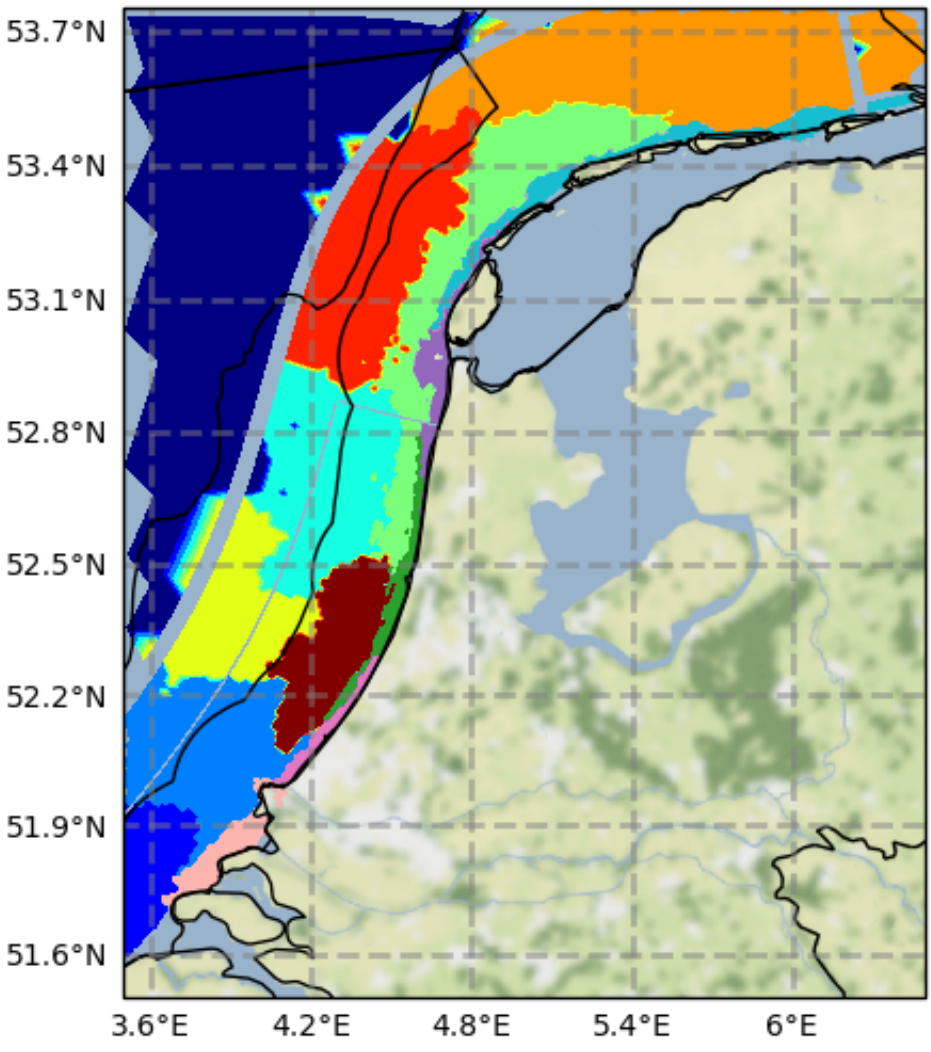
Clustering satellite data to define eutrophication monitoring zones

based on chlorophyll-a concentration

Laura Veerhoek

Bachelor End Project 2020



Deltares
Enabling Delta Life

TU Delft

Clustering satellite data to define eutrophication monitoring zones

based on chlorophyll-a concentration

by

Laura Veerhoek

to obtain the degree of Bachelor of Science
at the Delft University of Technology,
to be defended publicly on Monday July 13th, 2020 at 09:30 AM.

Student number:	4972112	
Project duration:	April 27, 2020 – July 13, 2020	
Thesis committee:	Dr. ir. G. Y. H. El Serafy,	TU Delft, Deltares, supervisor
	Prof. dr. ir. A. W. Heemink,	TU Delft
	Dr. ir. G. F. Nane,	TU Delft

Deltares
Enabling Delta Life

TU Delft

Abstract

OSPAR's Commission has been battling eutrophication since the problem was first established in the 1950s. To battle eutrophication, an important factor is to monitor it. Five indicators are used together to assess the status of eutrophication, determined by the Common Procedure. These are the chlorophyll-a concentration, the turbidity, the nitrate and phosphorus concentration, the oxygen levels and the biological water quality. All five indicators need to be known to obtain the final eutrophication status. However, just looking at the chlorophyll-a concentration on its own is also a good measure. This thesis focuses only on the chlorophyll-a concentration as an indicator for eutrophication.

To monitor the North Sea, the OSPAR's Commission has established eutrophication monitoring zones. The aim of this study is to determine eutrophication monitoring zones based on available satellite data of the chlorophyll-a concentration in the Dutch part of the North Sea. The zones are defined using four clustering algorithms: K-means clustering, Hierarchical clustering, Random Forest clustering and HDBSCAN. The results from these clustering algorithms are compared to both each other and to the previously defined eutrophication zones.

First, the case study region is split into two areas: the coastal area, which lies closer to the shore, and the offshore area, which lies farther away from the shore. The best result for this separation was generated by K-means clustering with two clusters.

Afterwards, the eutrophication zones are determined separately in the offshore area and the coastal area. The clustering results are ranked based on four criteria. The first criterion is correspondence to OSPAR's previously defined eutrophication monitoring zones. The second criterion is the similarity of the clusters to the zones that are visible in the data. The third criterion is the performance determined by validation metrics. This criterion was considered less important because of the lack of ability to capture the goals of the research. The last criterion is confirmation through the HDBSCAN clustering. This was added later during the study when it was found that HDBSCAN yielded very accurate results. Due to how HDBSCAN works these accurate results were not usable directly, as the number of clusters this yields it too high, but they were usable for verification. The best results were found through random forest clustering with respectively nine and five clusters for the offshore and coastal areas.

Subsequently, the zones derived from clustering were compared to other data to see whether the determined monitoring zones also hold over time. This appeared to be the case. Moreover, the distribution of the chlorophyll-a concentration for each zone is determined.Additionally, the trend of the chlorophyll-a concentration of one determined monitoring zone is analysed over time. Lastly, the defined eutrophication monitoring zones are compared to other defined zones within the Dutch North Sea coast. These other zones were fishery policies, marine protected areas, spatial planning, and bathymetry. The comparison validated the defined monitoring zones.

Preface

With this bachelor thesis, I present my scientific research results for determining eutrophication monitoring zones using clustering algorithms on preprocessed satellite data of the chlorophyll-a concentration in the Dutch North Sea coast.

I would like to thank my supervisor, Ghada El Serafy from Deltares, for her time, valuable input, and amazing support during my internship at Deltares. I also wish to thank Lőrinc Mészáros for all his input, both during the meetings and via email, and for the feedback throughout the process. Even though physical meetings were impossible during the period of my internship, our meetings were always very helpful and enjoyable.

Finally, I would like to thank Tina Nane and Arnold Heemink for taking a seat in my thesis committee.

Laura Veerhoek
Delft, July 2020

Contents

List of Figures

1 Introduction

Eutrophication is the process of water being enriched by nutrients. Eutrophication causes an increase in the growth of algae, such as phytoplankton. Higher growth of algae leads to there being less oxygen for other fish and plants. If this happens for a longer period, it contributes to the poor health of the marine environment and can even result in dead zones. Dead zones are areas in bodies of water where no marine life is possible. Eutrophication peaks naturally every year during a so-called spring bloom period, from mid-February to mid-June. However, mankind has increased the amount of eutrophication severely by, for example, discharges from a sewage treatment plant, industrial plants, fish farms, and agriculture. Thus, eutrophication and more specifically the human increase in eutrophication is something that needs to be combated (OSPAR, 2020c).

The OSPAR commission is a convention between 15 countries that protects and preserves the North-East Atlantic (OSPAR, 2020a). The North-East Atlantic is divided into five regions: the Arctic Waters, the Greater North Sea, the Celtic Seas, the Bay of Biscay and Iberian Coast, and the Wider Atlantic. This thesis will zoom in on eutrophication in part of the Greater North Sea, which is the worst in terms of the status of the OSPAR regions (OSPAR, 2013). To combat eutrophication, it first needs to be monitored and analysed. OSPAR (2013) has constructed a Common Procedure framework to analyse and combat eutrophication. Previously, eutrophication monitoring zones have been established in the North Sea by OSPAR based on the Dutch monitoring programme MWTL. The observations within the zones are supposed to behave similarly so that the monitoring is made easier. Near the coast are the highest peaks in eutrophication in the bloom period.

There are five main indicators for eutrophication: (Ferreira et al., 2011)

1. Chlorophyll-a concentration
2. Turbidity
3. Nitrate and phosphorus concentration
4. Oxygen levels
5. Biological water quality

This thesis will focus on the chlorophyll-a concentration as an indicator for eutrophication. Chlorophyll-a is only one part in the multi-step method defined by the Common Procedure to identify eutrophication. However, it does provide reliable insight into the trends of eutrophication (OSPAR, 2020b). To monitor the chlorophyll-a concentrations satellite observations can be used.

This thesis will aim to determine eutrophication monitoring zones where the chlorophyll-a concentration behaves similarly based on the statistical history, found in satellite data. This will be done using clustering algorithms to cluster the samples together that are close to one another spatially and behave similar over time. A secondary goal is to compare the clustering algorithms and their functionality in

determining the monitoring zones. Clustering algorithms group data into clusters. Data inside a cluster is more similar to each other than to data outside the cluster. Since this is also what is desired for creating eutrophication zones, clustering algorithms are used in this thesis.

1.1. Area and data

The data used in this case study to define eutrophication monitoring zones, is pre-processed satellite data of the chlorophyll-a distribution in the North Sea. This data was recorded by the MERIS instrument (Medium-Spectral Resolution Imaging Spectrometer) on board of the ENVISAT. This instrument is not operational anymore. The satellite data was interpolated onto a a numerical grid in order to be able to compare it with numerical model results. The gridded data was then interpolated in both space and time using the DINEOF (Data Interpolating Empirical Orthogonal Functions) algorithm from Beckers and Rixen (2003).

More specifically, the case study area is the region 50 km from the Dutch coast line, excluding the waters with different system dynamics, such as the Waddensea, the 'Nieuwe Waterweg', the Westerschelde, the Grevelingenmeer, and the Oosterschelde. This data set contains around 33 million data points over the course of 7 years. To start, this thesis focuses on the analysis of the year 2003.

The eutrophication monitoring zones which the clustering results will be compared to are defines by OSPAR, which can be found in Figure 1.1. These zones have been used many times for previous research, examples are Blaas (2013), Baretta-Bekker and Prins (2013), and obviously OSPAR itself periodically. There are officially five zones. However, some zones overlap each other, such as the coastal waters and the southern bight. This overlapping part will be considered as its own zone during this study.

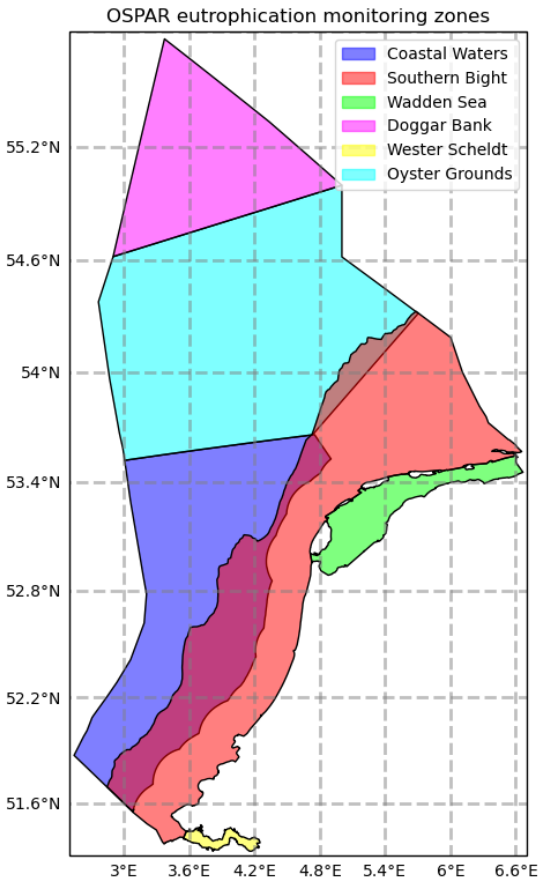


Figure 1.1: Eutrophication monitoring zones defined by OSPAR

1.2. Structure

This thesis contains an analysis of clustering algorithms as a way of defining eutrophication monitoring zones from the chlorophyll-a concentration in the Dutch North Sea coast. In Chapter 2 the methodology of defining the eutrophication zones is described. The clustering algorithms, their procedures, their assumptions, and their possible advantages and disadvantages are discussed in Chapter 3. Chapter 4 contains the results and their interpretation and validation. The results are then also compared to the existing eutrophication monitoring zones determined by OSPAR. Chapter 5 gives a conclusion of the analysis and a recommendation for further uses. A discussion on what could be improved in this study is contained in Chapter 6.

2

Methodology

This chapter describes the research methodology used for determining the eutrophication monitoring zones. This is done by describing which part of the available data is used and how it is used to determine the monitoring zones.

2.1. Data for zone definition

The analysis will be performed on the data from the year 2003. However, the choice should be made what part of the data is going to be used to define the monitoring zones in the Dutch coastal area. Within 2003 the options are the following:

- Run the clustering algorithms on all the data from 2003.
- Run the clustering algorithms only on the data from the spring bloom period since the peaks in chlorophyll-a concentration only take place in this period. The bloom period runs from mid-February to mid-June.
- Run the clustering algorithms on one day in the bloom period to simplify the computations. This would be a good option if there is one day that captures all distinct zones in the data. A possible choice for this day could be a day in the bloom period when there is a peak in chlorophyll-a concentration.

The objective of this study is to define eutrophication monitoring zones that capture the peak concentrations and the zones that occur when leading up to this peak. Next to this, the computational runtime is also a factor that plays a role in selecting which option is best.

Running the clustering algorithms on all the data from 2003 would be too slow. Another reason not to use all of the data from 2003 is that any peaks in chlorophyll-a concentration would be evened out by the large amount of ordinary days. As these peaks are important to the clustering, and running on such a large dataset was not feasible, an alternative had to be found. Secondly, using only the data from the spring bloom period could be a better option, since that decreases the number of days where no peak is visible. This means that the peaks have more importance. Nevertheless, it would still result in long computational runtimes. The third option, where one day in the bloom period is used, covers both the objective to capture the peak concentrations and the desire for a short runtime. Consequently, the clustering algorithms will be run on one day in the bloom period.

There is high variability in the concentration of chlorophyll-a in the Dutch coastal area over the year, even in the spring bloom period. Comparing a representative day for the start of the bloom period, and a bloom peak in respectively Figure 2.1 and Figure 2.2 shows an example of this. In Figure 2.1, a clear distinction in at least two zones can be seen, one more close to the shore and one farther away from the shore, named offshore. The offshore area itself does not contain very distinct zones.
In Figure 2.2, the difference between the offshore and coastal zones is barely visible. However, in the

offshore area, multiple distinct zones are visible. If the definition of zones would be based on the 22nd of March, the distinction between the coast and offshore would be lost, but if the zones are to be defined based on the 17th of February, the zones in the offshore area would be lost. The choice for the day is thus very important to the outcome of the definition of the zones.

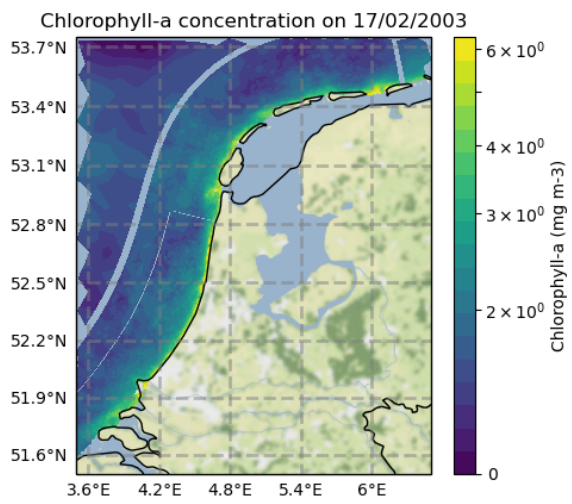


Figure 2.1: Chlorophyll-a concentration on a representative day for the start of the bloom period (17-02-2003).

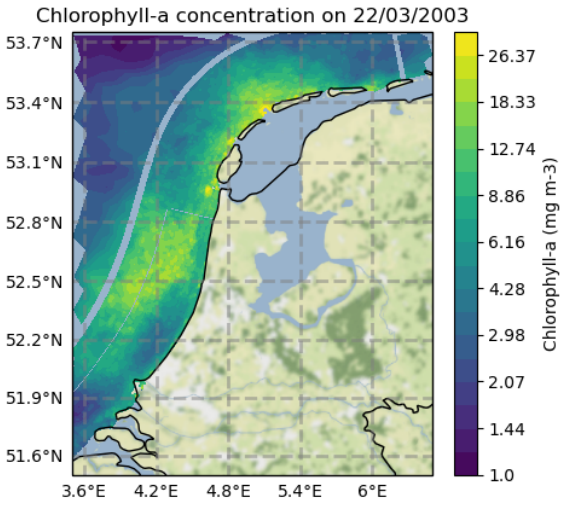


Figure 2.2: Chlorophyll-a concentration on a representative day for a bloom peak (22-03-2003)

2.1.1. Day(s) for zone definition

Going through the visualisation of all available days in the bloom period of 2003 it became clear that several distinct regions need to be found in the eventually chosen day. These are the following:

- The distinction between the coastal area and the offshore area,
- Regions in the offshore area,
- Regions in the coastal area.

In the data available in the bloom period of 2003, it is not possible to see all regions at once for one specific day. This is because the variability in different regions is not comparable. Therefore, the need arises for assessing regions separately. This results in using data from more than one day to cluster on. By first separating the coast and offshore on one day and then defining the zones separately for both the coastal and offshore regions, results improved significantly. It was possible for all visible zones to be clustered together.

Firstly, the coastal area and the offshore area will be separated by clustering on a day that states a clear distinction between the two. This is the case when the spring bloom begins and therefore it shows the coastal area, where the conditions for algae growth are more optimal. A representative day is the 17th of February, Figure 2.1. After the coast-offshore separation was made, the offshore zones were defined on a day where there are higher concentrations of chlorophyll-a and zones are visible. The highest in chlorophyll-a concentration would be on the 16th of April. However, on this day the zones in the offshore area were not visible. Therefore, a day representative for leading up to the peak was chosen, namely the 22nd of March (Figure 2.2). The zones within the coastal area were defined based on a day representative of the start of the bloom, but with visible zones. The 25th of February was chosen for this.

The choices for these specific days were hand-picked by going through the visualisations of the available data and finding which days exhibit distinct zones in each area. The choice for the day that will be used for the coastal area was only made after the separation of the coastal area and the offshore area

was made. The reason for this is that the variations in this area are overshadowed by the variations in the offshore area. A day representative for the variability of chlorophyll-a concentration in the coastal area is the 25th of February, given in Figure 2.3.

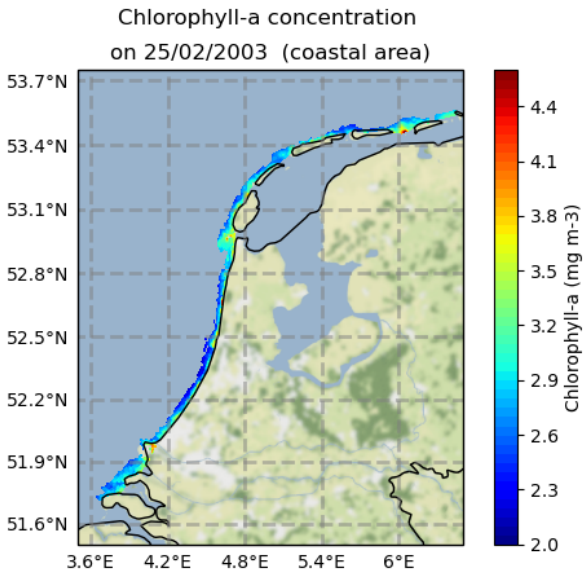


Figure 2.3: Chlorophyll-a concentration in the coastal area on 25-02-2003

After the selection was made for the data, the algorithms were run to produce the results in Chapter 4. The implementation can be found in Appendix D. The clustering algorithms that were run on the data are K-means clustering, hierarchical clustering, Random Forest Clustering, and HDBSCAN. How these clustering algorithms work and what their advantages and disadvantages are can be found in the next Chapter 3.

Subsequently, the results were compared to one another through four criteria:

- They resemble OSPAR's already existing eutrophication monitoring zones.
- The zones that were expected to be clustered visually were clustered by the algorithm.
- The clustering result yielded the best values for the validation metrics.
- Their result was confirmed through the accurate HDBSCAN clustering result.

The last criterion was added after it had been established that the HDBSCAN algorithm gave a very accurate clustering, which was unusable because the number of clusters was too high.

The general methodology used for this study is visualised in Figure 2.4.

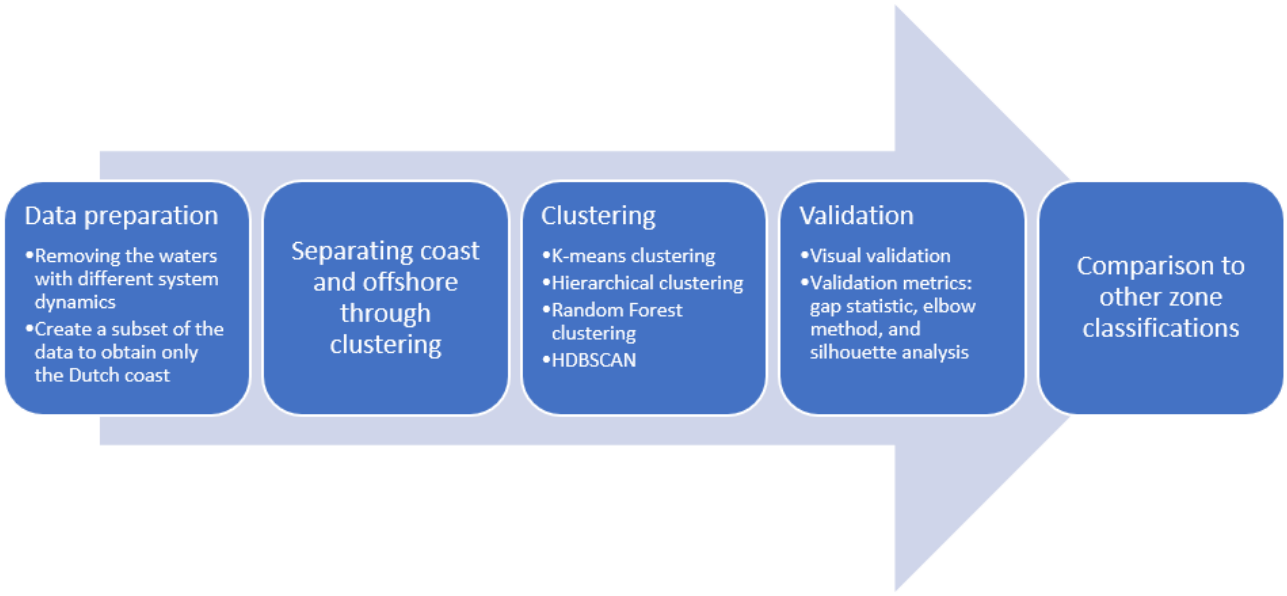


Figure 2.4: General methodology used for this study

3

Clustering algorithms

Inside one eutrophication region, the chlorophyll-a concentration should behave similarly over time and also be close spatially. This can be interpreted mathematically as clusters, where each cluster is a zone where the chlorophyll-a concentration and location coordinates are similar. The problem clustering algorithms try to solve was well formulated by Aggarwal and Reddy (2014, p. 2):
"Given a set of data points, partition them into a set of groups which are as similar as possible."

To determine the clusters there are many existing clustering algorithms. All exhibit different features, and this thesis aims to compare their functionality in determining the eutrophication monitoring zones. The algorithms that will be used for comparison in this thesis are K-means clustering, hierarchical clustering, Random Forest clustering, and HSBSCAN.

3.1. K-means clustering

The K-means clustering algorithm was first introduced by Lloyd (1982), it is also known as Lloyd's algorithm or naive K-means. K-means consists out of the following steps.

1. Initialise k cluster centres, so-called centroids. This can be done by a random selection of data points. However, to speed up the process the kmeans++ can also be used, more on this in Section 3.1.1.
2. Each data point is assigned to the nearest cluster centroid. The most commonly used distance metric is the Euclidean distance. It is important to note that the data should be scaled since distances are being used.
3. The centroids are updated to the mean of all data points inside the cluster.
4. Step 2 and 3 are repeated until the algorithm converges and the centroids are no longer updated in step 3 or a maximum number of iterations is achieved.

K-means partitions the data into k clusters while minimising the distance between the points of a cluster and the centroid of that cluster, the Within-Cluster Sum of Squares (WCSS) see equation 3.1.

$$WCSS = \sum_{i=1}^{k} \sum_{x_j \in c_i} (x_j - \bar{x}_i) \tag{3.1}$$

where c_i is the i^{th} cluster, and $\bar{x}_i$ is the mean of all the objects assigned to cluster c_i.

A problem with K-means clustering is that it is an NP-hard problem (Dasgupta, 2008). This means that you can end up in a local minimum. Thus, the outcome may not be optimal. Whether a local minimum is reached or not depends a lot on the initialisation of the centroids. Luckily, there are ways to improve the initialisation. One of the most commonly used solutions is to run the algorithm multiple times. This

results in multiple random initialisations. The WCSS can then be compared and the best clustering, with the lowest WCSS, can be chosen. Another solution would be to use K-means++ initialisation, see Section 3.1.1. These two solutions can also be used simultaneously.

Another disadvantage of the K-means clustering is that the number of clusters, k, needs to be defined beforehand. Not only the quality and usefulness of the results but also the space and time complexity are impacted. To approximate the number of clusters, there are a few common methods: the elbow method, the gap statistic analysis and the silhouette analysis. All three of these methods require running the algorithm with multiple values for k. If there is some knowledge of the data, this can be reduced to only testing the most likely values. More on these methods can be found in Section 3.6.

3.1.1. K-means++

To improve the initialisation process, Arthur and Vassilvitskii (2006) suggested the K-means ++ initialisation. The intuitive idea is that the initial centroids are chosen far apart. The fist centroid is chosen randomly. The next centroid is then chosen as the data point that is the farthest away from the first centroid. The third centroid is chosen as the data point farthest away from the previously chosen centroids. This is continued until you have k centroids. After initialisation, the standard K-means algorithm is executed. The K-means++ initialisation has a longer computation time than the standard K-means initialisation because the distanced between objects and centroids need to be computed. However, the convergence increases so much that the total runtime is reduced significantly.

3.2. Hierarchical Clustering

Hierarchical clustering is based on creating clusters to optimise an objective function (Ward Jr, 1963a). The objective function is called the linkage criterion which is the distance between clusters based on the distance between data samples. The type of hierarchical clustering that will be used in this study is agglomerative clustering. Agglomerative clustering starts with n clusters, where n is the number of samples in the data set, and iteratively merges clusters based on the objective function. The hierarchy that is created through this algorithm can be visualised in a dendrogram. An example of a dendrogram can be found in Figure 3.1

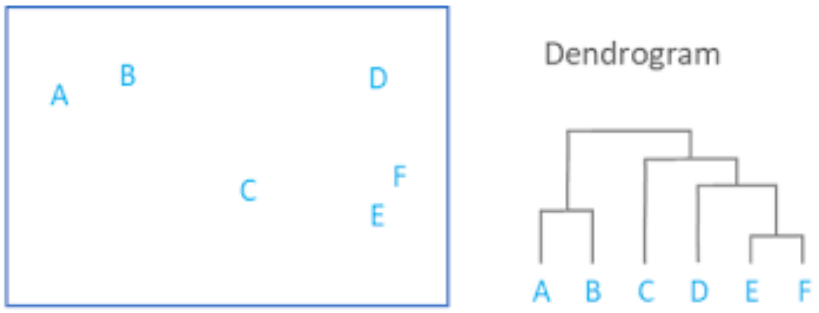


Figure 3.1: A dendrogram example (Bock, nd)

After the hierarchy is built, the clustering is given by choosing the number of clusters and slicing the dendrogram/hierarchy at the height that results in that many clusters.

The input parameters for hierarchical clustering are the distance measure, the linkage criteria and the number of clusters. For the distance measure, each possible distance can be chosen. The most used distance measure is the Euclidean distance. Another possibility is that an already precomputed distance matrix is used as input instead of a distance measure and the data. However, this does limit the options for linkage criteria. Some linkage criteria need the actual values of the data points to calculate the distance. Ward's criterion is an example.

For linkage criteria, there are a lot of options. The two most common ones are Ward's linkage criterion and average linkage criterion.

3.2.1. Ward's Linkage Criterion

Ward's linkage criterion (Ward Jr, 1963b) is based on minimising the total within-cluster sum of squares, equation 3.1. This is the same objective as K-means has.

3.2.2. Average Linkage Criterion

The average linkage criterion states that the distance between two clusters is equal to the average distance from any member of one cluster to any member of the other cluster (Sokal, 1958). This can be expressed in formula form as the following:

$$\frac{1}{|A| \cdot |B|} \sum_{a \in A} \sum_{b \in B} d(a, b) \tag{3.2}$$

3.3. Random Forest

A random forest is not originally a clustering algorithm. To be specific, a random forest is an ensemble classifier. An ensemble classifier is a collection of weak classifiers that are combined to increase their strength but maintain simplicity. Nevertheless, a random forest can be used to derive a distance or similarity measure of the data. This distance measure has many advantages over euclidean distance. Some of these advantages are given in Table 3.1 (Shi and Horvath, 2006).

3.3.1. Supervised Random Forest

A random forest is a collection of n decision trees (Breiman, 2001). To classify an object, it is run down each decision tree. This gives n classification votes. Taking the majority vote gives the final classification from the forest. This procedure is called bagging.

The building of one decision tree can be separated into two steps. The first step is to retrieve the data that the tree will be using by bootstrapping the original data. Bootstrapping means pulling a subset from the original data with replacement. The second step is to build a decision tree. At each node, only a subset of the columns of the data is used to determine the optimal split. To determine the optimal split the GINI index is used (Gini, 1912). The number of columns used is a parameter. The tree is grown as large as possible and is not pruned.

3.3.2. Validation

Since a bootstrapped data set and not all data is used for building each tree there is an opportunity to validate the tree. The data that did not end up in the bootstrapped data is called the out-of-bag (OOB) data. Running this data down the respective trees and noting the number of times they are correctly classified gives an OOB-error.

3.3.3. Unsupervised Random Forest

Determining which split is the best using the GINI index is only possible is you have labelled data. Clustering is an unsupervised procedure and therefore the random forest should be adapted to an unsupervised procedure. This procedure is explained by Breiman and Cutler (1912).

The main idea is to create a synthetic data set drawn from the original data. The synthetic data and real data are artificially labelled based on if they are real or synthetic, respectively 0 or 1. To create the synthetic data the empirical distribution of each feature is determined. m new data samples are drawn from those distributions.

Now that the data is essentially labelled, the original random forest algorithm can be used.

3.3.4. Proximity matrix

After building the random forest it can be used to derive a similarity measure. To derive this for two objects, say object a and object b, the objects are both run down each tree in the forest. They are similar if they end up in the same leaf nodes. The measure is thus the number of times in the same leaf node divided by the number of trees. The values for each combination of a and b in the data can be represented in a proximity matrix, which can be converted to a distance matrix.

This distance matrix can then be used by a different clustering algorithm such as hierarchical clustering to obtain the clustering. This thesis will use the hierarchical clustering algorithm with the average linkage criterion. From this point forward that combination will be referred to as random forest clustering.

3.4. Hierarchical Density-Based Spatial Clustering for Applications with Noise (HDBSCAN)

Hierarchical Density-Based Spatial Clustering for Applications with Noise, HDBSCAN, was presented by Campello et al. (2013) as an improvement to the DBSCAN algorithm (Ester et al., 1996). The algorithm takes two input parameters, namely the minimum cluster size (m_{clsize}) and the minimum points (m_{pts}). What these entail will be explained later in this section.

The HDBSCAN algorithm can be separated into 5 steps.

1. Transforming the space to separate sparse and dense data by defining a new metric.
2. Build the minimum spanning tree based on this new metric.
3. Create cluster hierarchy tree.
4. Condense the cluster hierarchy tree.
5. Extract the clusters from the tree.

3.4.1. Step 1: Transforming the space

The first step is to transform the space to separate sparse and dense data. In the HDBSCAN algorithm, the core distance measure is used to determine the denseness or sparsity.

Def. Core distance$_k$ of a point a, denoted as $core_k(a)$, is the distance to the m_{pts}-nearest neighbour of a.

If the core distance is high, the data point is sparse since there are not a lot of data points around it and vice versa. To separate the sparse and dense data a new distance measure is introduced, called the mutual reachability distance.

Def. Mutual reachability distance$_k$ between a point a and a point b is the maximum of their Euclidean distance, and both of their core distances.

$$reach\_d_k(a,b) = \max\{core_k(a), core_k(b), d(a,b)\} \tag{3.3}$$

The data will be transformed based on this new distance measure.
To clarify how this transformation separates dense and sparse data an example is given. If the distance between a sparse point and a dense point is smaller than the core distance of the sparse point, the distance in the new space will be further apart. Thus separating the dense and sparse data.

3.4.2. Step 2: Build a minimum spanning tree

The new space of transformed distances can be considered as a weighted graph where each edge between vertices has the weight of the mutual reachability distance, the so-called mutual reachability graph (G_{mpts}). To eventually obtain clusters, the data points that are close to each other should be connected. Data points far apart do not need a direct connection. Therefore, the edges with a high weight, and thus high distance apart, should be removed. An alternative way of looking at it would be to only keep the edges with low weight while the graph remains connected. This can be done by building a minimum spanning tree. This can be done with Prim's algorithm (Jarník, 1930)(Korte and Nešetřil, 2001).

A data point is considered noise if the density level is below the core distance of that point. When considering a density-based cluster hierarchy this is an important feature. To maintain this in the HDBSCAN algorithm, an edge to itself will be added to the minimum spanning tree for each data point. Thus expanding the minimum spanning tree.

3.4.3. Step 3: Construct hierarchical tree

From the minimum spanning tree constructed in the previous step, we can derive a hierarchical tree. At the root node, all objects are assigned to the same cluster. Removing the edges of the minimum spanning tree iteratively in descending order and clustering the remaining connected components together gives a hierarchical tree.

3.4.4. Step 4: Condense hierarchical tree

The hierarchical tree from the previous section can be visualised using a dendrogram. However, if the data being considered is relatively large, it results in the dendrogram not being easily readable. Zooming in on what happens in the tree gives an interesting observation, namely that most of the times only a few points are detaching from a bigger cluster at each step. Campello et al. (2013) classified three possible evolutions of a cluster.

1) Less than m_{clsize} points split off from a cluster, therefore only shrinking a cluster. This component is then clustered as noise.

2) A cluster splits into two clusters if the remaining connected components both have more than m_{clsize} objects.

3) A cluster disappears is is is split into two and the remaining connected components both have less than m_{clsize} objects.

The m_{clsize} is thus a threshold for the cluster size which smoothes out the tree. Any cluster below the threshold is deemed noise. To make HDBSCAN more similar to other density-based clustering algorithms and to simplify it, the default setting for the m_{clsize} and the m_{pts} parameters is to make them equal to each other. Whether or not this is the best approach completely depends on the data at hand

3.4.5. Step 5: Extract clusters

Extracting clusters from the condensed hierarchical tree can be done in multiple ways. A vastly used course of action before HDBSCAN was taking one single horizontal cut at some level of the tree. However, the density level of a certain level of the tree is equal in each cluster when taking one single horizontal cut. Therefore, clusters with different density levels are impossible. To solve this, HDBSCAN uses a new cluster extraction mechanism: choosing the clusters based on their stability.
The intuitive idea is that a prominent cluster exists for a long time and for that reason can be called stable. ”Long time” corresponds to a lot of density levels in the dendrogram.

Def. λ is a density threshold that corresponds to the density levels in the hierarchical dendrogram.

Def. $\lambda_{min}(C_i)$ is the density threshold at which cluster C_i emerged.

Def. $\lambda_{max}(C_i)$ is the density threshold at which cluster C_i either splits or disappears.

Def. $\lambda_{max}(x_j, C_i) = \min\{f(x_j), \lambda_{max}(C_i)\}$ is the density level of which the object x_j no longer belong to the cluster C_i, either because the points falls out or the cluster's end of existence is reached.

Stability of a cluster C_i can be expressed in terms of the density thresholds where the cluster emerges and vanishes.

Def. Stability is a measure of how long each point in the cluster stays in the cluster.

$$S(C_i) = \sum_{x_j \in C_i} (\lambda_{max}(x_j, C_i) - \lambda_{min}(C_i)) \tag{3.4}$$

For each cluster, the stability can be calculated. Selecting the few clusters with the highest stability does not work for two reasons. The first is that it could be possible that one cluster is the ”child” or stems from another cluster being selected. Thus some objects would be clustered twice. Secondly,

there may be objects not being clustered at all.

The problem of selecting the clusters can be expressed as an optimisation problem. Let C_1 be the cluster at the root and let $\{C_2, \dots, C_k\}$ be the collection of all other clusters in the condensed density-based hierarchical tree.

$$\max_{\delta_2,\dots,\delta_k} J = \sum_{i=2}^{k} \delta_i S(C_i) \tag{3.5}$$

$$\text{subject to } \begin{cases} \delta_i \in \{0, 1\}, & i = 2, \dots, k \\ \sum_{j_h} \delta_j = 1, & \forall h \in L \end{cases} \tag{3.6}$$

The solution to this optimisation problem is the following:

1. Initialise $\delta_2 = \dots = \delta_k = 1$, and set $\hat{S}(C_h) = S(C_h)$ for all leaf nodes.
2. starting from deepest level, traverse the tree from the bottom-up and do:
   (a) if $S(C_i) < \hat{S}(C_{il}) + \hat{S}(C_{ir})$ set $\hat{S}(C_i) = \hat{S}(C_{il}) + \hat{S}(C_{ir})$ and set $\delta_i = 0$
   (b) else: $\hat{S}(C_i) = S(C_i)$ and set $\delta_\cdot = 0$ for all clusters in C_i's subtrees.

3.5. Advantages and disadvantages of each clustering method

	Advantages	Disadvantages
K-means clustering	• Fast • Simple • Intuitive input parameter	• May end up in local minimum • Sensitive to outliers • Number of clusters is an input parameter
Hierarchical clustering	• Fast • Simple • Linkage criterion can be chosen to suit the goal	• Number of clusters is an input parameter • Linkage criterion choice not always intuitive
Random Forest clustering	• Relatively intuitive input parameters • Handles highly skewed variables well, which is the case for chlorophyll-a concentration. • Provides a difference distance metric, especially good if euclidean distance is not necessarily obvious. This is the case for spacial temporal data such as in this study. • Robust to outliers in the data. There are both high and chlorophyll-a concentrations and the ones that are high can be seen as outliers. Outliers in this instance does not mean falsely measured, rather far off the mean.	• Slow • The random forest only creates dissimilarity matrix, a different clustering algorithm is still needed to obtain clusters • Number of clusters is an input parameter
HDBSCAN	• Fast • Noise detection. However, for this research, this is not per se a big advantage since there does not seem to be real noise in the data just peaks in the concentration of chlorophyll-a. • Number of clusters is not an input parameter, but determined by the algorithm. • Clusters with different sizes and densities	• Parameter choice not intuitive • Having the algorithm determine the number of clusters could lead to a lot of clusters, whereas this research aims to keep the number of clusters as small as possible while still capturing all information. • Data can be clustered as noise while it is not necessarily noise, just an outlier.

Table 3.1: Advantages and disadvantages of each clustering algorithm

3.6. Metrics

The quality of a clustering algorithm can be measured using certain metrics. These metrics can in return also be used to determine the best input parameters for the algorithm, such as the number of clusters in the K-means clustering algorithm. They can also be used to compare different clustering algorithms as validation metrics. Possible options for metrics are the elbow method, the silhouette method and the gap statistic. All three have their advantages and disadvantages which will be discussed in the following sections.

3.6.1. Elbow method

The elbow method was first established by Thorndike (1953). It uses the WCSS, as stated in equation as a measure of the quality. Simply minimising the WCSS would not be beneficial since that would result in choosing the number of clusters as high as the number of samples in the data set. The goal is to have a low WCSS while keeping the number of clusters to a minimum. The elbow method makes use of the fact that as soon as a certain number of clusters has been added, adding another cluster only decreases the WCSS slightly. This point is called the elbow point. Ideally, this elbow point would be distinguishable in a graph. However, this is not always the case. Thus it is beneficial to use the multiple metrics.

Methods such as K-means and hierarchical clustering minimise the WCSS using the numbers of clusters given as an input parameter. For those algorithms, the elbow method is especially useful in determining the number of clusters.

When using the WCSS as a validation metric for comparing clustering algorithms it is enough to minimise the WCSS value, since it is a measure for the error.

3.6.2. The silhouette method

The silhouette method presented by Rousseeuw (1987) is a way to measure the tightness and separation of the clusters. The silhouette method starts by computing a silhouette score for each observation. Two measures are defined to create the silhouette score: one for the dissimilarity of the cluster and one for the separation from the closest cluster.

Def. a(i) is the average dissimilarity or distance between the points inside the cluster that i is assigned to, say A.

$$a(i) = d_{avg}(i, A) \tag{3.7}$$

where $d_{avg}(i, C)$ is the average dissimilarity or distance between i and the points inside cluster C.

Def. b(i) is the minimum of all $d_{avg}(i, C)$'s over the collection of clusters i is not assigned to.

$$b(i) = \min_{C \neq A} d_{avg}(i, C) \tag{3.8}$$

For each observation the silhouette score is calculated by the following equation:

$$S(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \tag{3.9}$$

All of these scores are averaged together to create the average silhouette score. The average silhouette score ranges from -1 to 1. If it is close to one the clusters are perfectly separated, near zero the clusters are likely to be overlapping, and below 0 the data points are most likely inaccurately clustered.

3.6.3. Gap statistic

The gap statistic was invented by Tibshirani et al. (2001). It compares the total within-cluster variation with the expected value under a null reference distribution of the data. The null reference distribution is a distribution where there is no clustering, which coincides with the uniform distribution.

Def. Gap statistic for k number of clusters:

$$Gap_n(k) = E_n^* log(W_k) - log(W_k) \tag{3.10}$$

where W_k is the total within-cluster variation, E_n^* is the expectation under a sample size n from the reference distribution defined by bootstrapping by generating B copies of the data set and computing the average.

There are two ways to choose the number of clusters using the gap statistic. One is choosing k where the gap statistic is the highest. The second one is choosing k the smallest where

$$gap(k) \geq gap(k+1) - s_{k+1} \tag{3.11}$$

where s_k is the standard error.

3.6.4. Use the metrics with caution

Although it seems ideal to choose the input parameters solely based on the results from the metrics, there are two main downfalls to it. The first being that different metrics could yield different results and thus not giving a definitive answer. The second downfall is that sometimes the clusters that should be obtained do not satisfy the objective function the metrics try to minimise.

4

Results

In this chapter, the results of defining eutrophication monitoring zones using different clustering algorithms will be presented and discussed.

First, the coastal area and the offshore area were separated. Then, the zones in the offshore area were defined. Subsequently, the zones in the coastal area were defined. Afterwards, the defined zones were compared to other days within the bloom period, and also compared to other years. Lastly, the defined monitoring zones were compared to other previously defined zones in the North Sea coast.

4.1. Separating the coastal area and the offshore area

To separate the coast and offshore, the chlorophyll-a data from a representative day from the start of the bloom period was used. The 17th of February was used (Figure 2.1). Before results could be produced using the clustering algorithms, the number of clusters needed to be chosen as this is an input parameter in three of the four clustering algorithms. To separate the coast and offshore an obvious choice for the number of clusters would be two. However, it could also be possible to choose more than two clusters and combine clusters by hand to obtain the best clusters. To determine the optimum number of clusters the elbow method, silhouette and gap statistic were used. The reasoning behind the statistics are explained in Section 3.6. All three statistics were computed for each clustering algorithm. For K-means clustering, the results are visualised and explained. For the other algorithms, the results are stated and visualisations are given in Appendix A.1.

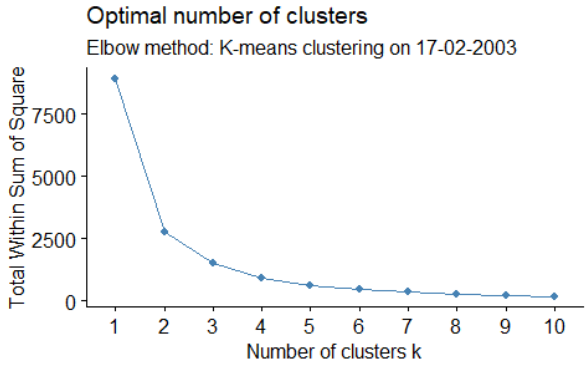


Figure 4.1: The elbow method for K-means clustering to separate the coast and offshore

The WCSS for different values of the number of clusters based on the K-means clustering method are given in Figure 4.1. The elbow method gives an indication here that it could be either 2 or 3, but has no definitive answer.

In Figure 4.2 the Gap statistic is given. Here, the optimal number of clusters is the lowest number

of clusters where the gap statistic is higher than the next gap statistic minus the standard error from this next value. This is the case for one cluster. Thus, the gap statistic declares that there should not be any clustering to the data. This is not realistic. The second best option is two number of clusters, because this has the second highest gap statistic.

In Figure 4.3 the silhouette average scores are given. These simply need to be maximised, which is the case for number of clusters equal to two.

This prompts the conclusion that the assumption of two clusters is correct for the K-means clustering algorithm.

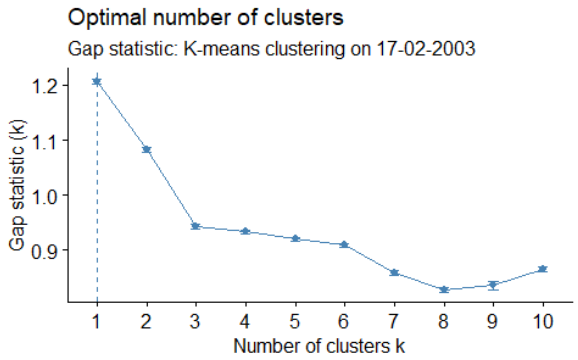


Figure 4.2: The gap statistic for K-means clustering to separate the coast and offshore

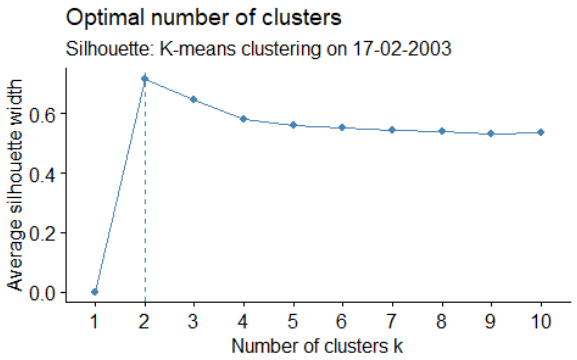


Figure 4.3: Silhouette analysis for K-means clustering to separate the coast and offshore

For the other clustering algorithms the results are the following:

Clustering method	Elbow method	Silhouette average	Gap statistic
Hierarchical clustering	2, 3, or 4	2	1
Random Forest clustering	inconclusive	10	1

Table 4.1: Values of the goodness of clustering metrics for separating the coast and the offshore.

The values in Table 4.1 do not give a very clear answer for the number of clusters that should be chosen. Therefore, the number of clusters will be equal to two.

4.1.1. K-means

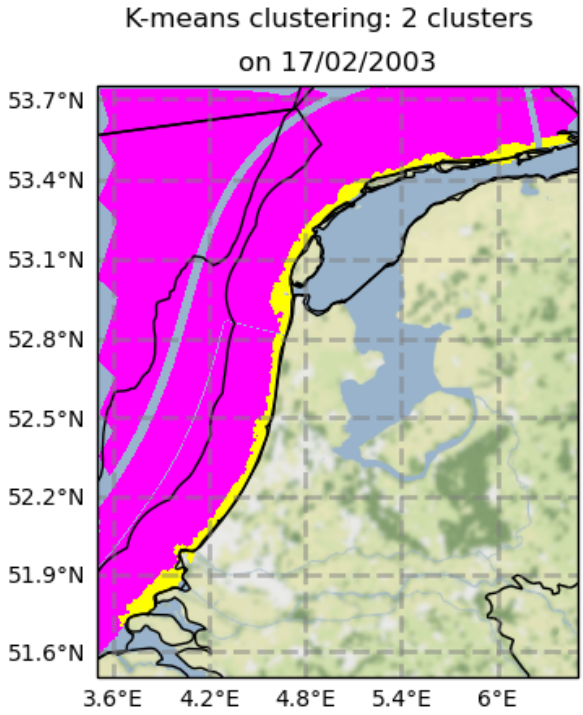


Figure 4.4: K-means clustering with 2 clusters on 17-02-2003

The results produced by K-means clustering with a predefined number of clusters of two can be found Figure 4.4. The result is a separation between the coast and offshore which is almost parallel to the Dutch coast line.

The black lines in the figure are the previously defined eutrophication zones. The clustering based on the K-means algorithm represents the line closest to the coast in parallel well. The only difference is how far into the sea it lies. The previously defined zone border lies a lot deeper into the sea than the one obtained from the K-means clustering.

4.1.2. Hierarchical clustering

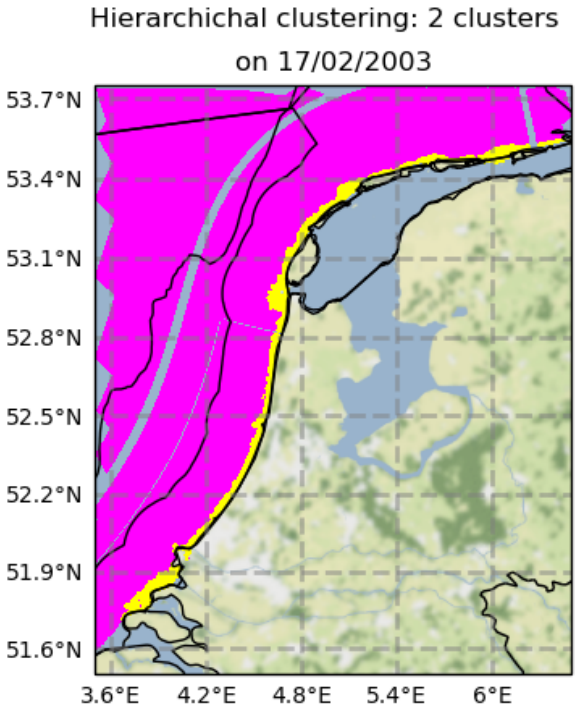


Figure 4.5: Hierarchical clustering with 2 clusters on 17-02-2003

The results from the K-means clustering, and hierarchical clustering with euclidean distance and Ward's linkage criterion are very similar. The results from hierarchical clustering are seen in Figure 4.5. A possible reason for the similarity is that the linkage criterion chosen in the hierarchical clustering technique is Ward's criterion. As explained in section 3.2.1, the ward linkage minimises the total within-cluster variance. This is the same objective as K-means clustering. There is one small difference and that is the width of the coastal area, which is slightly smaller for hierarchical clustering.

4.1.3. Random Forest

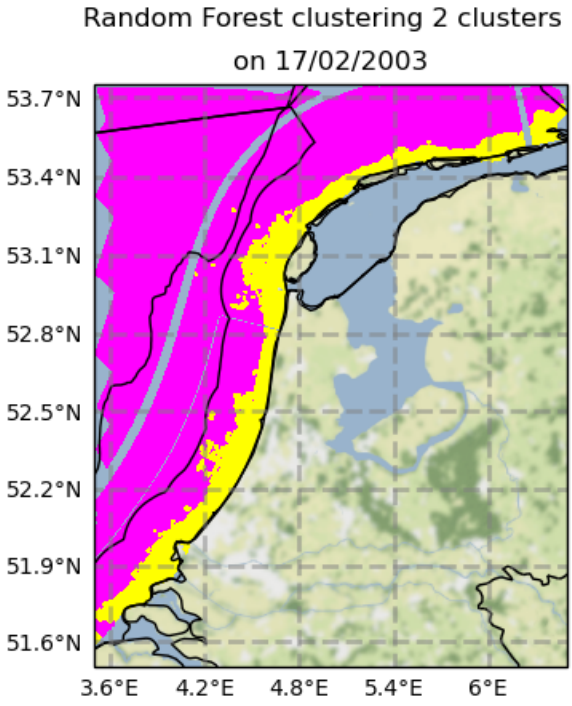


Figure 4.6: Random Forest clustering with 2 clusters on 17-02-2003

To recap, random forest clustering is the random forest proximity measure combined with the hierarchical clustering using the average linkage criterion. Random forest clustering with two clusters yielded the results in Figure 4.6. There are a few remarkable observation that can be made.

The first is that the border protrudes a lot more into the sea compared to K-means and hierarchical clustering. It almost reaches the original eutrophication zone given by the black line closest to the shore.

Another observation, is that there are small regions in the offshore area that are clustered together with points near the coastline. These can be seen in the figure as yellow dots surrounded by pink. This is reasonable near the border between the pink and the yellow part, but the farther away from this border, the less realistic it becomes. A possible reason for this is that the distinction between the coast and offshore is based solely on the height of the concentration of chlorophyll-a and not on also the location. In the offshore areas, multiple areas exhibit slightly higher concentrations of chlorophyll-a. The distance between those offshore areas and the coast is apparently very different when comparing euclidean and random forest distances.

4.1.4. HDBSCAN

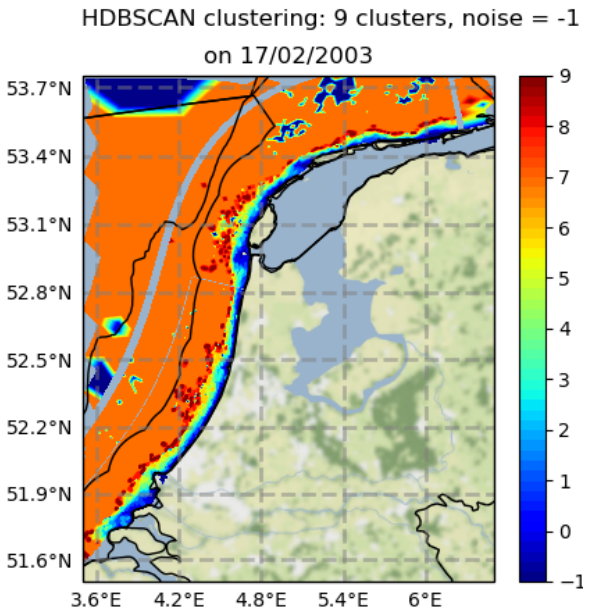


Figure 4.7: HDBSCAN with noise on 17-02-2003

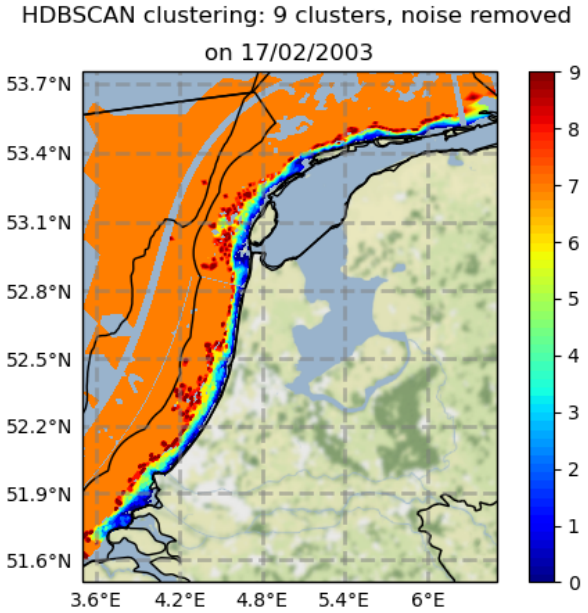


Figure 4.8: HDBSCAN without noise on 17-02-2003

The HDBSCAN algorithm, where the number of clusters is determined by the algorithm, resulted in 21 clusters, excluding noise. This can be seen in Figure 4.7. Noise is classified as -1 and Figure 4.8 shows the results where the noise is removed. To recap, a sample is classified as noise if there are not a lot of similar samples surrounding it. A more in depth definition can be found in Section 3.4

HDBSCAN recognises that there are multiple levels in the coastal region. This gives a more accurate representation of the real situation. The levels more closely to the shore, the clusters with number zero to two resemble the clusters as given in hierarchical clustering. If cluster numbers 3 through 6 are added the result is the same as for K-means clustering. If cluster number 8 and 9 are added, the result resembles the Random Forest clustering. This more accurate representation is thus also a good way to confirm the results from the previous algorithms. The levels visible in the HDBSCAN clustering corroborate the shape and size of the coast cluster from the previous clustering algorithms.

For the purpose of creating a separation between the coastal area ans the offshore area there must be only two clusters. This can be achieved by combining clusters. The choice of what clusters should be combined is done visually. Namely, the offshore area is probably cluster number 7. If the cluster 7 is considered as the offshore area and the rest is combined into the coast, this would give the split as seen in Figure 4.9.

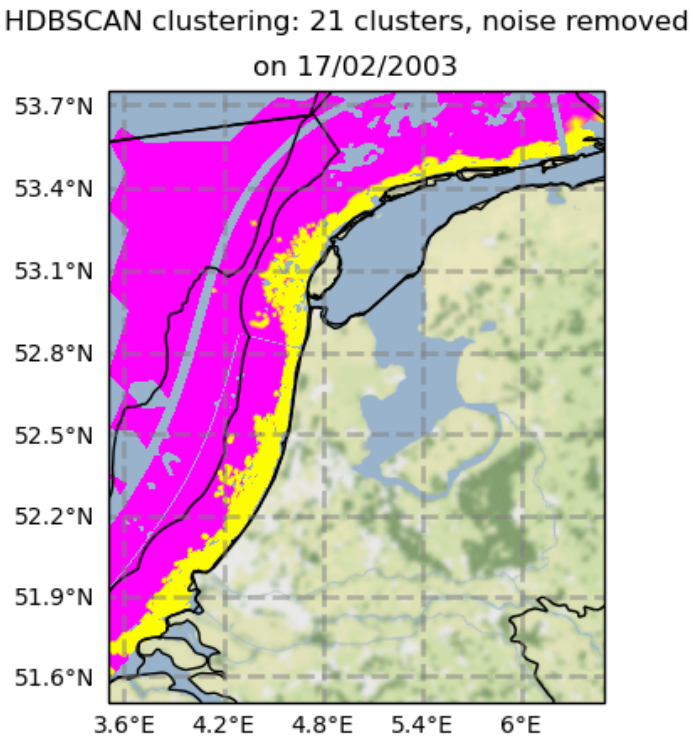


Figure 4.9: HDBSCAN combined to 2 clusters on 17-02-2003

This result gives a wider coastal area that still also follows the previously defined zones. There are areas that protrude out quite some more such as at $53.1°$N. A very interesting observation is that the HDBSCAN clustering is almost identical to the random forest clustering, apart from the noise. There is no clear explanation for this.

4.1.5. Comparing the clustering algorithms

All clustering algorithms led to similar results, the only big variation was in the width of the coastal area. The width was increasing in the order of hierarchical clustering, K-means clustering, random forest clustering, and HDBSCAN.

From these results, the idea that there should be a split between the coastal area and the offshore area is confirmed. Since all clustering algorithms yield very similar results and no clustering algorithm is specifically better than others in this case, the clustering algorithm closest to the average will be used to separate the coast and the offshore areas. In this case, average corresponds to the average width of the coastal area. The algorithm closest to this average is K-means clustering. After the coast is separated from the offshore area, cluster analysis can be done on the coast and the offshore regions separately.

4.2. Defining the zones offshore

After removing the coastal area clustered by K-means in the previous section, the zones in the offshore area can be defined. This will be done using the location and chlorophyll-a concentration data from the 22nd of March, see Figure 2.2. The reasoning behind this choice can be found in the methodology in Chapter 2.

Within the offshore area there are multiple zones that are clear and should be uncovered in the clustering result. A visualisation of these zones can be found in Figure 4.10. The number of clustered should intuitively thus be at least more than the eight seen in that figure.

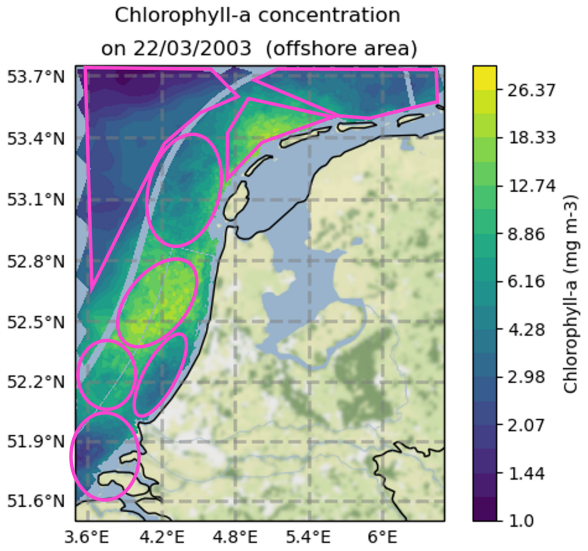


Figure 4.10: Visible zones in the offshore area on 22-03-2003

One of the zones that is visible in Figure 4.10 holds the Rhine ROFI. ROFI stands for region of freshwater influence. The Rhine ROFI is the region in the Dutch part of the North Sea where the ”Nieuwe waterweg”, which originates from the Rhine, ends. Because of the rotation of the earth the water that stems from the river is directed to the right, which is North in this case. The dimensions of the Rhine ROFI can be up to 20-40 kilometre wide and up to 100 km long (de Ruijter et al., 1997). However, on average the size is what is also seen in Figure 4.10 (Van der Giessen et al., 1990).

The number of clusters can also be determined by the metrics defined in Section 3.6. The results from these can be found in table 4.2. The visualisations of the metrics are given in Appendix A.2.

Clustering method	Elbow method	Silhouette average	Gap statistic
K-means clustering	4, 5, or 6	4	8
Hierarchical clustering	4, 5, or 6	6	10
Random Forest clustering	inconclusive	9	1

Table 4.2: Values of the goodness of clustering metrics in the offshore area

The results from the different metrics are very different and also differ a lot over the different methods and thus overall inconclusive. For that reason the number of clusters will be chosen visually, based on Figure 4.10. There are at least eight zones in the offshore area. To account for a small clustering between the larger clusters the number of clusters will be chosen equal to 9.

4.2.1. K-means clustering

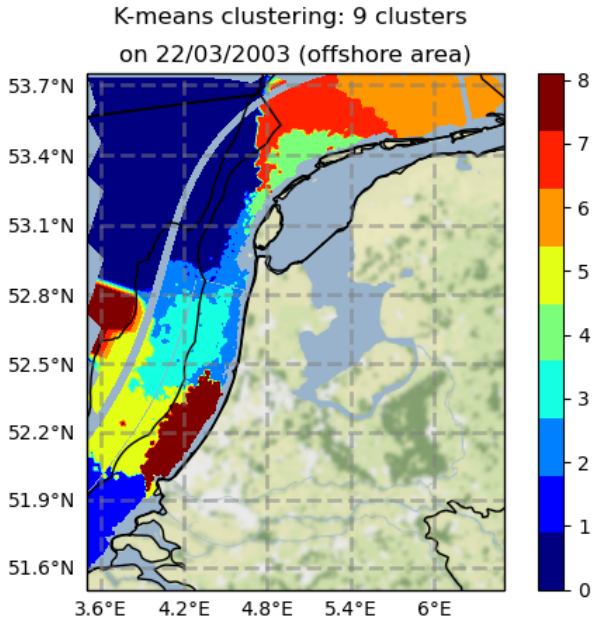


Figure 4.11: K-means clustering with 9 clusters on 22-03-2003

Using K-means clustering on the 22nd of March with nine clusters gives an accurate presentation of the zones that were desired to be uncovered. In dark brown, cluster number 8, the Rhine ROFI is visible. In bright green, cluster number 4, the region to the north-west of the Waddensea is visible.

The clusters slightly follow the previously defined zone borders. However, not fully since the bright blue cluster, cluster number 3, is separated into two by the previous zone border. This could mean two things: the previously defined zone was not perfect and there should not be a zone border there, or there only seems to be a zone on this day but it is split into two if there other days are considered.

Another interesting observation is that the bright blue zone in the middle, cluster number 3, encapsulates observations that are classified to a different cluster, namely, cluster number 2. This can be seen by the darker blue dots in the bright blue cluster. This is not realistic, but is most likely the case since there are also lower chlorophyll-a concentrations in a zone and the clustering algorithm works with a euclidean distance. The distance in concentration and the location coordinates have equal weights. This is both an advantage and a disadvantage. It is good, since anything other than equal weight would be unnatural. The bad part is that the clustering algorithm cannot sense the fact that there also can be samples with a slightly different concentration inside a cluster. If a different day was chosen to fit the clustering on, and there was less of a peak in chlorophyll-a concentration, there may not be this difference in the zone.

4.2.2. Hierarchical clustering

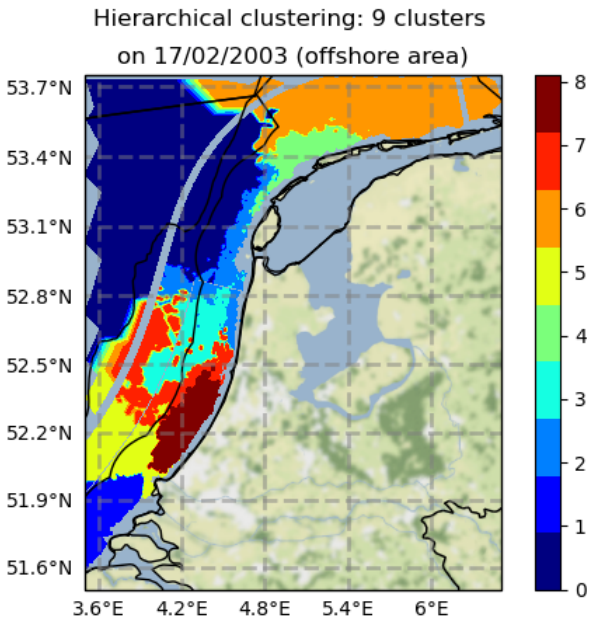


Figure 4.12: Hierarchical clustering with 9 clusters on 22-03-2003

Using hierarchical clustering on the 17th of February resulted in very similar results to K-means clustering. There were only two differences in shapes of the zones.

One is that the the previously yellow cluster, number 5, has split into two, creating the red cluster, number 7, around the bright blue cluster, number 3.

The other is that the orange and red cluster in K-means clustering, number 6 and 7, merged together to become cluster orange number 6. This cluster covers a large area north of the Waddensea.

It is interesting to note that almost the exact same samples were encapsulated in the bright blue zone as for k-means clustering.

4.2.3. Random Forest clustering

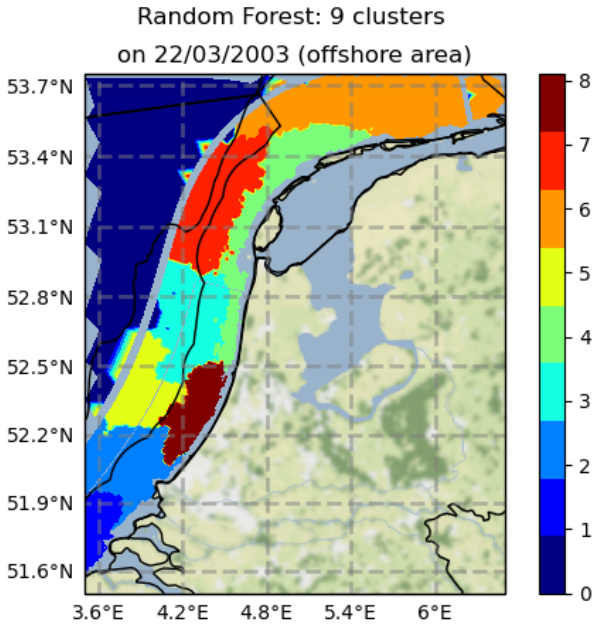


Figure 4.13: Random Forest clustering with 9 clusters on 22-03-2003

The random forest clustering with 9 clusters yielded the results in Figure 4.13. There are a few aspects in this clustering that were different from k-means and hierarchical clustering.

The first being the shape of the zone that contains the Rhine ROFI, cluster number 8. In K-means and hierarchical clustering this zone had an oval shape. In the random forest clustering it is an oval with an indent on the side from the yellow zone. cluster number 5.

Another aspect that is interesting is that the the green cluster, number 4, comes far more down to the south, whereas this part was previously clustered separately.

A third difference is that there are far less samples being encapsulated by a different zone.

Finally, when comparing the random forest clustering zones with the existing eutrophication zones it seems to be resembling them more than previous clustering results. The dark blue zone, cluster number 0, almost completely follows the previously defined eutrophication zones that lies more off-shore. Also, a new zone is defined, the red zone with cluster number 7. This zone resembles that region inside the two previously defined eutrophication zone borders. To recap, this area is the overlap of two of OSPAR's eutrophication zones. The random forest clustering result thus confirms that this overlap should be considered its own zone.

4.2.4. HDBSCAN

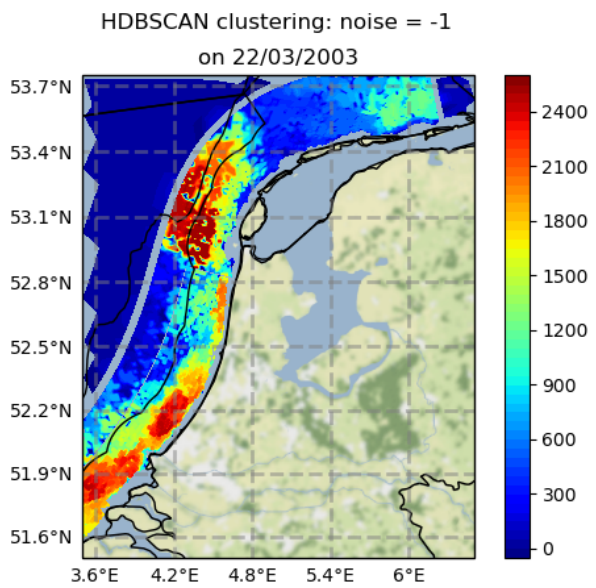


Figure 4.14: HDBSCAN with noise on 22-03-2003

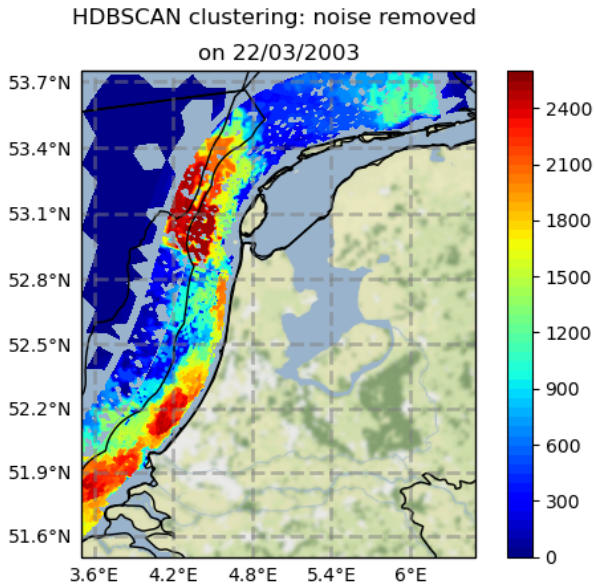


Figure 4.15: HDBSCAN without noise on 22-03-2003

In Figure 4.14 the clustering of the HDBSCAN are given where noise is classified as -1. This noise is removed in Figure 4.15. HDBSCAN resulted in over 2400 clusters. This gives a more accurate clustering, but less realistic. For this thesis the goal was not to perfect the clusters, but to see if clustering algorithms could detect the clusters and compare their performances. Therefore, the HDBSCAN result will be used to validate the other clustering algorithms for the offshore area.

In the HDBSCAN clustering result there are zones visible that substantiate the previously defined zones. These zones are:

- The red zone in random forest clustering, cluster number 7,
- the zone containing the Rhine ROFI, cluster number 8,
- the zone in front of Sealand, cluster number 1,
- the more offshore zone, cluster number 0,
- the zone to the north-east of the Waddensea, cluster number 4.

The last zone mentioned is not as very clearly visible due to the colour scheme, but is certainly there.

All zones considered, HDBSCAN supports random forest clustering with nine clusters as the best.

4.2.5. Comparing the clustering algorithms

The clustering algorithms yielded results in the offshore area were related. However, there were significant differences between the results. The clustering algorithms were compared based on four criteria:

- They resemble OSPAR's already existing eutrophication monitoring zones.
- The zones that were expected to be clustered visually, seen in Figure 4.10, were clustered by the algorithm.
- Their result was confirmed though the accurate HDBSCAN clustering result.
- The clustering result yielded the best values for the validation metrics.

Random forest clustering was the most accurate match to OSPAR's monitoring zones. There were a couple clusters that exceeded the boundaries of the zones, but for the most part the shapes were similar. The main difference was the fact that clustering result contains more clusters than the number of monitoring zones. Using the newly defined monitoring zones would thus yield a more accurate result.

The second criteria was that the zones that were clearly visible when looking at the chlorophyll-a concentration, were clustered by the algorithm. Random forest clustering was the only clustering algorithm that obtained the red zone, cluster number 7, as its own cluster. The other clusters were mostly obtained by all clustering algorithms in some form.

Almost each zone from the random forest clustering is also visible in the clustering result from HDBSCAN. This does not hold for the other clustering algorithms.

In table 4.3 the values for the validation metrics are shown for each clustering algorithm.

	Total within cluster sum of squares	Silhouette average	Gap statistic
K-means clustering	2582	0,427	0,926
Hierarchical clustering	2875	0,423	0,971
Random Forest clustering	4555	0,000	0,600

Table 4.3: The values of the validation metrics for the offshore area

To recap, the objectives are minimising the value for the total within cluster sum of squares and maximising the silhouette average and the gap statistic. Based on these metrics k-means clustering would be the best algorithm, and random forest clustering the worst. This is where a problem arises: all other validation criteria pointed towards random forest clustering being the best, but the validation metrics state that it is the worst. This is a good example of what was touched upon in section 3.6.4. Namely, that metrics do not necessarily have the same objective as the research and thus may not be suitable in all situations. All three metrics are a measure for how similar the data points in a cluster are, where the silhouette average also takes the distances to other clusters into account. Apparently, maximising the similarity within the clusters does not yield the best result as defined per the other criteria. Therefore, the result from criterion will be considered less important.

All in all, the Random Forest clustering with nine clusters is the best at defining monitoring zones in the offshore area, compared to the other clustering algorithms.

4.3. Define zones in coastal area

Now that the zones in the offshore area are defined, the coastal area is analysed. The data from the 25th of February, Figure 2.3, is being used to do this. Reasoning for this can be found in the methodology in Chapter 2. Again, the first step is determining the number of clusters. Looking at the concentration of chlorophyll-a on the 25th of February, there seem to be four clusters. These are visualised in Figure 4.16.

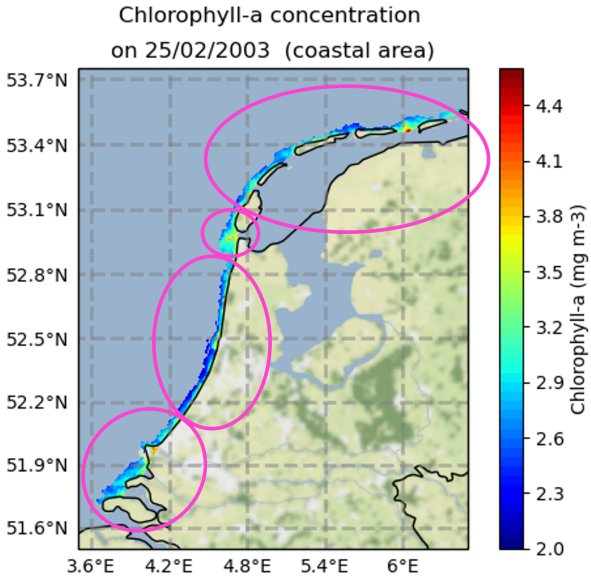


Figure 4.16: Visible zones in the coastal area 25-02-2003

Another way to determine the optimal number of clusters was through metrics, which are explained in Section 3.6. The results from these metrics are stated in Table 4.4.

Clustering method	Elbow method	Silhouette average	Gap statistic
K-means clustering	5, or 6	5	3
Hierarchical clustering	inconclusive	6	1
Random Forest clustering	inconclusive	10	1

Table 4.4: Values of the goodness of clustering metrics in the coastal area

The metrics do not give a cohesive answer. Since visually there are around 4 clusters, that is what will be used as in input parameter for the clustering algorithms.

4.3.1. K-means clustering

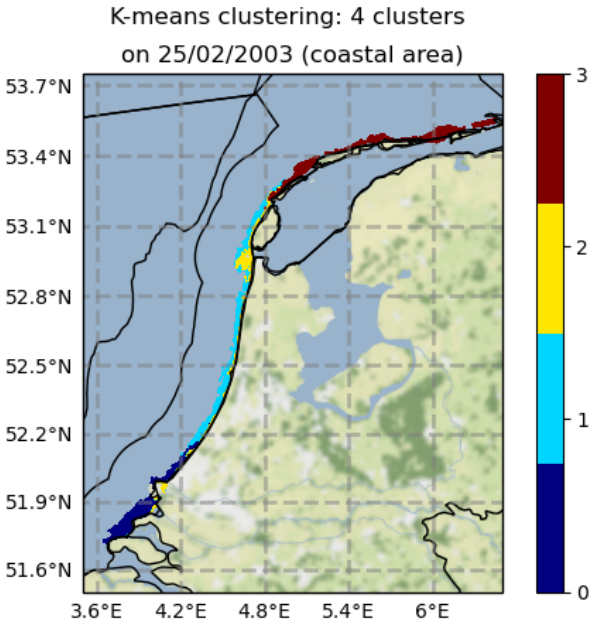


Figure 4.17: K-means clustering with 4 clusters on 25-02-2003

The K-means clustering algorithm gives the result as shown in Figure 4.17. Comparing this clustering to the one that was expected in Figure 4.16, the clusters do not coincide perfectly. However, there are a few interesting observations that were made.

The first being that the edge between the light blue and dark blue cluster is diagonal. This shows that there is a difference between really near coast and a bit farther away.

Another interesting observation can be made on that the yellow zone that protrudes slightly more in the sea around $(4.7°\text{E}, 53°\text{N})$. It is clustered as its own cluster but surrounded on both sides by the light blue cluster. A cluster being surrounded by another cluster is not realistic in this case.

4.3.2. Hierarchical clustering

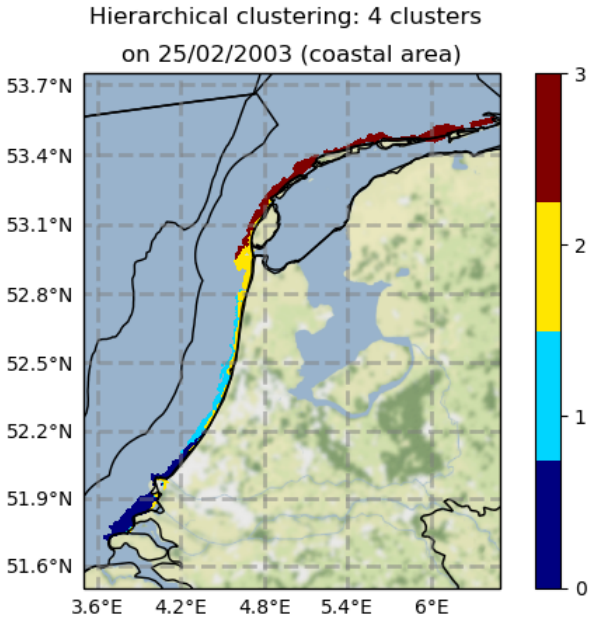


Figure 4.18: Hierarchical clustering with 4 clusters on 25-02-2003

In Figure 4.18 the results from hierarchical clustering with four clusters is given. This result exhibits the same diagonal edge between the light blue and the dark blue clusters. The area that protrudes a bit more in the sea around ($4.7°$E,$53°$N) is clustered completely as its own cluster and is not surrounded by a different cluster. The hierarchical clustering resembles the expected zones almost exactly.

4.3.3. Random Forest clustering

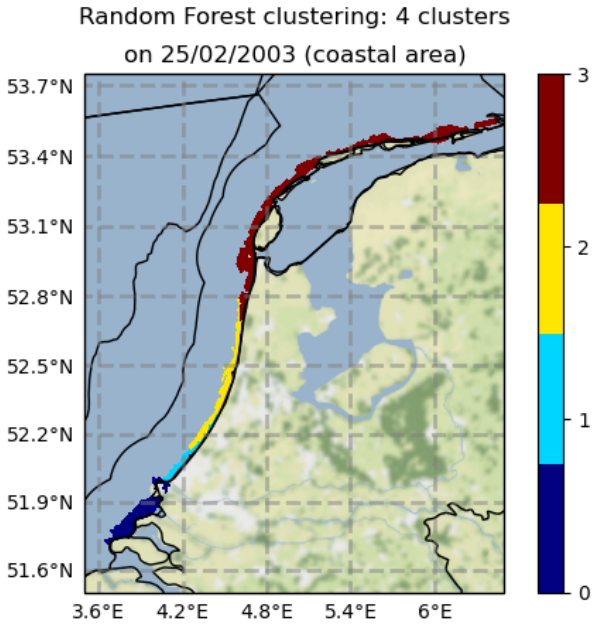


Figure 4.19: Random Forest clustering with 4 clusters on 25-02-2003

The result from random forest clustering is given in Figure 4.19. The random forest clustering is different form the K-means and hierarchical clustering in two ways:

The first way is that the differences in the lower half, below $52.8°$N, are more accentuated using random forest clustering. It separated the area outside of Zealand and the area above the Rhine.

The second way is that the whole upper part, above $52.8°$N, is clustered together. Thus, there is no separate cluster in the area around $(4.7°\text{E}, 53°\text{N})$. From Figure 4.16 that area does seem to be different from the upper part. This prompts the idea that random forest clustering needs more clusters to get the desired clusters. The random forest clustering with five clusters is seen in Figure 4.20.

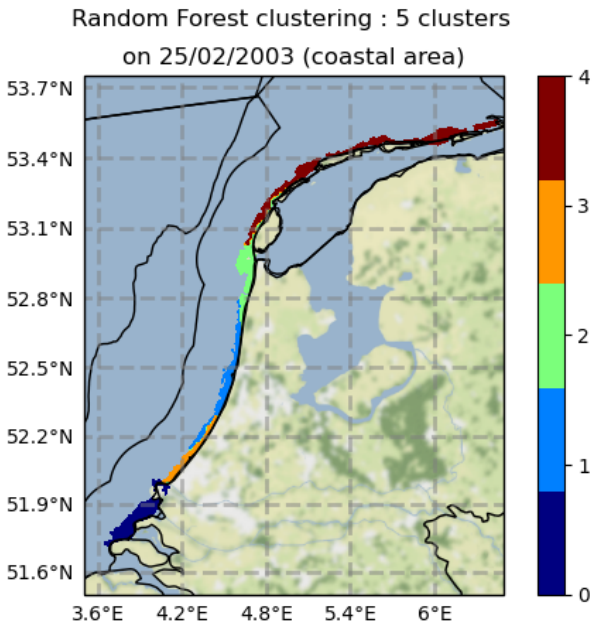


Figure 4.20: Random Forest clustering with 5 clusters on 25-02-2003

If random forest clustering is run with five clusters it separated the upper cluster into two. Resulting in the area around (4.7°E,53°N) being clustered as one cluster. This is exactly what was expected and corresponds to the desired clustering results.

4.3.4. HDBSCAN

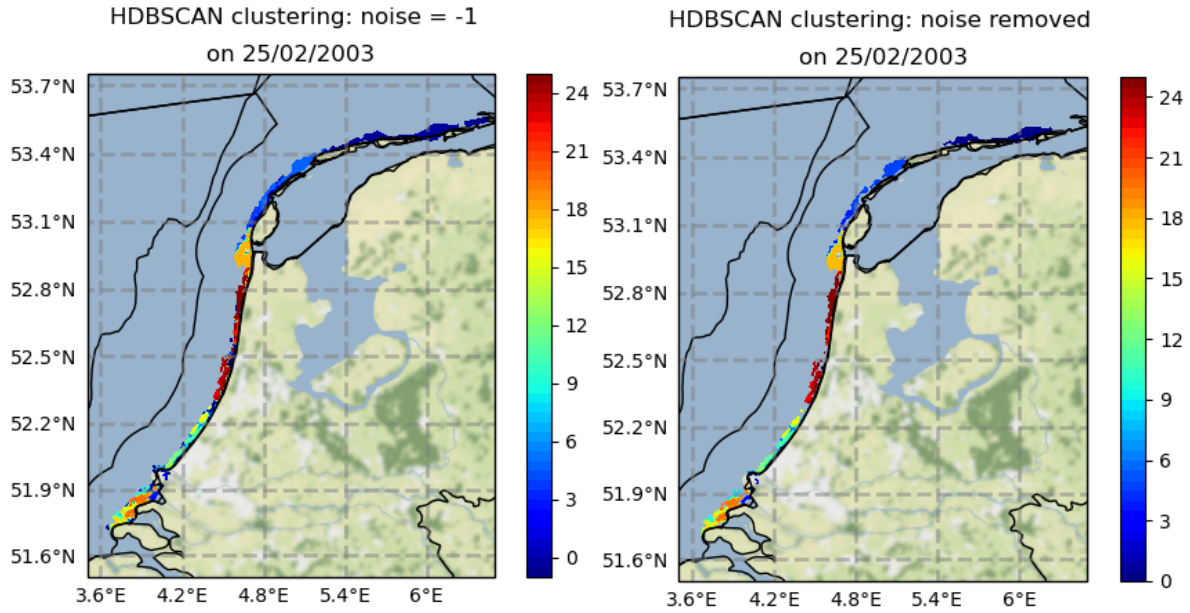


Figure 4.21: HDBSCAN with noise on 25-02-2003

Figure 4.22: HDBSCAN without noise on 25-02-2003

HDBSCAN resulted in 25 clusters visualised in Figure 4.21 with noise, and in Figure 4.22 without noise. In comparing the two figures, it should be pointed out that there is not a lot of noise. The HDBSCAN result can again be used to validate the clusters defined earlier in this section.

The first observation is that the area around $(4.7°\text{E}, 53°\text{N})$ is clustered as its own cluster. This supports the random forest clustering with five clusters and the hierarchical clustering, which also clustered this as its own zone.

Interestingly enough, the diagonal edge between zone one and two in k-means and hierarchical clustering, and one and three in random forest clustering, is not visible in the HDBSCAN result. There is a division, but it is not diagonal.

In the coast outside Zealand there are a lot of clusters in the HDBSCAN result. The reason for this may be that the concentrations of chlorophyll-a varies a lot in this region. Since it is not really connected to the rest of the coastal area because of an indent in the shore, clustering this as a separate region seems to be reasonable. This is the case for random forest clustering with both four and five clusters.

All in all, the HDBSCAN result supports the random forest clustering with five clusters as the best clustering.

4.3.5. Comparing the clustering algorithms

In the coastal area the clustering algorithms yielded significant differences between the results. The clustering algorithms were compared based on three criteria. These are the criteria as mentioned before in Chapter 2 and used in Section 4.2.5, without the first criterion where the result should resemble OSPAR's monitoring zones. The reason for this is that OSPAR does not divide the sea so close to the shore. As an alternative, the results will be compared to other possible divisions of the North Sea in Section 4.7. For clarity the criteria that are used in this section are repeated:

- The zones that were expected to be clustered visually, seen in Figure 4.16, were clustered by the algorithm.
- Their result was confirmed though the accurate HDBSCAN clustering result.
- The clustering result yielded the best values for the validation metrics.

Both random forest clustering with five clusters and hierarchical clustering determined the clusters similar to the zones visibly seen in Figure 4.16.

The result from HDBSCAN suggested that the area in front of Zealand should be its own cluster. Random forest clustering with five clusters has this property. Hierarchical clustering also clustered this together, but extended more to the north.

The last criterion is based on the validation metrics stated in Figure 4.5.

	Total within cluster sum of squares	Silhouette average	Gap statistic
K-means clustering	1858	0,435	0,563
Hierarchical clustering	2123	0,380	0,600
Random Forest clustering (4 clusters)	1042	0,010	0,167
Random Forest clustering (5 clusters)	1025	0,020	0,151

Table 4.5: The values of the validation metrics for the coastal area

The validation metrics ranked the clustering algorithms from best to worst as: K-means clustering, hierarchical clustering, random forest clustering with five clusters, and random forest clustering with four clusters. Seeing as random forest clustering is ranked low and the goal of the research is the same as in the offshore area, the same reasoning can be used as in section 4.2.5. Namely, that the validation metrics do not provide an accurate insight into the situation for the purpose of this research. Therefore, the result from this criterion is considered less important.

Overall, Random Forest clustering with five clusters is the best option when defining monitoring zones in the coastal area.

Combining this result with the one from Section 4.2.5 gives the defined eutrophication monitoring zones. It is the combination of random forest clustering with nine and five clusters respectively in the offshore and coastal area.

4.4. Comparing the zones with other days in the bloom period

The results derived from the clustering algorithms in the previous sections are based on the data of a total of three days. Since there is always movement in the sea, the possibility exists of the zones changing over time. Comparing the defined eutrophication zones to other days in the bloom period may give insight into the correctness of the zones. This is done by overlaying the contours of the clusters over the chlorophyll-a concentration on different days. Two of these comparisons can be found in Figure 4.23 and Figure 4.24.

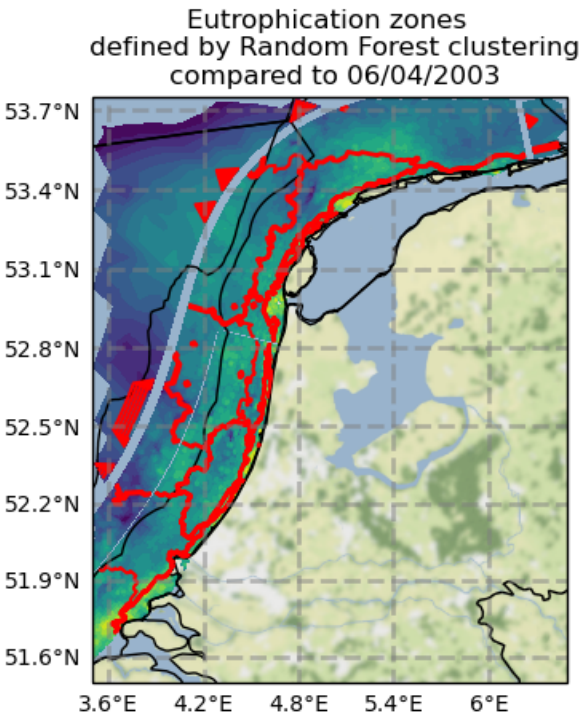


Figure 4.23: Comparison of the defined eutrophication monitoring zones to the chlorophyll-a concentration from 6/4/2003

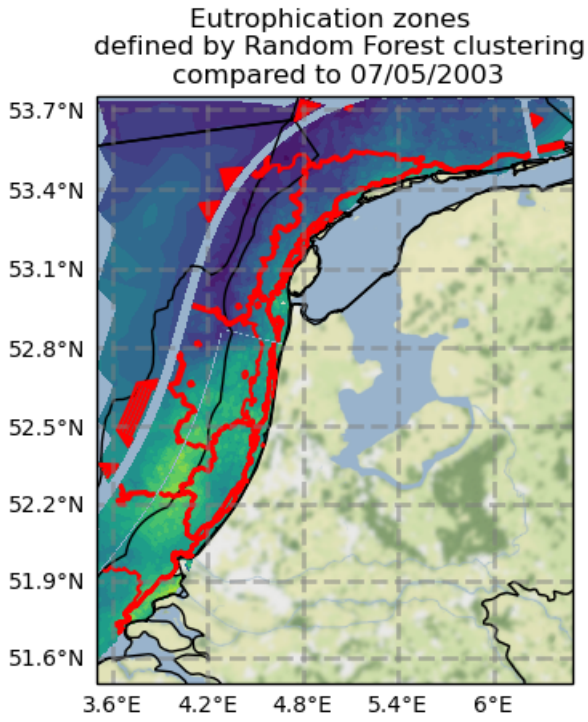


Figure 4.24: Comparison of the defined eutrophication monitoring zones to the chlorophyll-a concentration from 7/5/2003

The days that are used to compare the defined eutrophication monitoring zones to are chosen randomly from the bloom period of 2003, but after the days that were used to fit the data.

Figure 4.23 shows some interesting observations.
First, the separation between coast and offshore is considered. The higher concentration of chlorophyll-a are near the shore and seem to lie well inside the defined coastal area. The only improvement would be that the coastal zone in front of Zealand should be slightly extended towards the south. Since the separation between the coast and offshore area follows a line parallel to the shore, it is intuitive that it should be extended to cover the whole shore.

In the offshore area there are definitely variations in concentrations but there do not seem to be many clear zones. Therefore, it does not confirm nor deny the validity of the defined zones. The only zone really well visible is the zone north-west of the Waddensea. That seems to be slightly more extended to the North.

In Figure 4.24 there are a few remarkable comparison results to be seen.
The first is that coast-offshore separation is still able to capture the differences between the two areas.

The second is that the Rhine ROFI zone well covers the visible decrease in chlorophyll-a concentration in the Rhine ROFI. Another observation is that the zone in the overlapped part of OSPAR's eutrophication zones is visible. This is the previously bright red zone with cluster number 7 in the offshore random forest classification. This zone seems to be slightly extended to the south on this specific day.

Finally, the peak in the concentration around ($4.1°$E,$52.3°$N) is divided over two defined monitoring zones.

All in all, the defined eutrophication monitoring zones have been mostly confirmed. The zones did change somewhat, though not as much to discredit their validity. Another interesting thing to note is that the chlorophyll-a concentrations within the clusters vary. This is likely since there are different conditions at each location. Not accepting this and separating the clusters accordingly would yield a lot of clusters. This would not be feasible as a monitoring zone. Therefore, this is accepted.

4.5. Analysis of the chlorophyll-a concentration of the defined zone

The chlorophyll-a concentration of each zone can be represented in terms of a probability distribution. The concentration of chlorophyll-a generally follows a log-normal distribution (Campbell, 1995). Therefore, that is the distributions that was used for each zone. It is interesting to observe whether and how the distributions for the chlorophyll-a concentration differ from zone to zone. The expectation is that they do differ since they are clustered to different zones. The zones used in this analysis are the best results from the previous sections. Namely, random forest clustering with nine and five clusters for respectively the offshore and the coast.

The distributions of each zone in the offshore area can be found in Figure 4.25. For the coastal area the empirical distributions of the zones are given in Figure 4.26. It is important to note that the Figures have been zoomed in to clearly see what happens, in reality all distributions run to really high x-values. The full images can be found in Appendix C. The colours in the figures correspond to the colours used to define the zones in Section 4.2 and 4.3. The values for the parameters corresponding to the fitted log-normal distribution can be found in Appendix B.

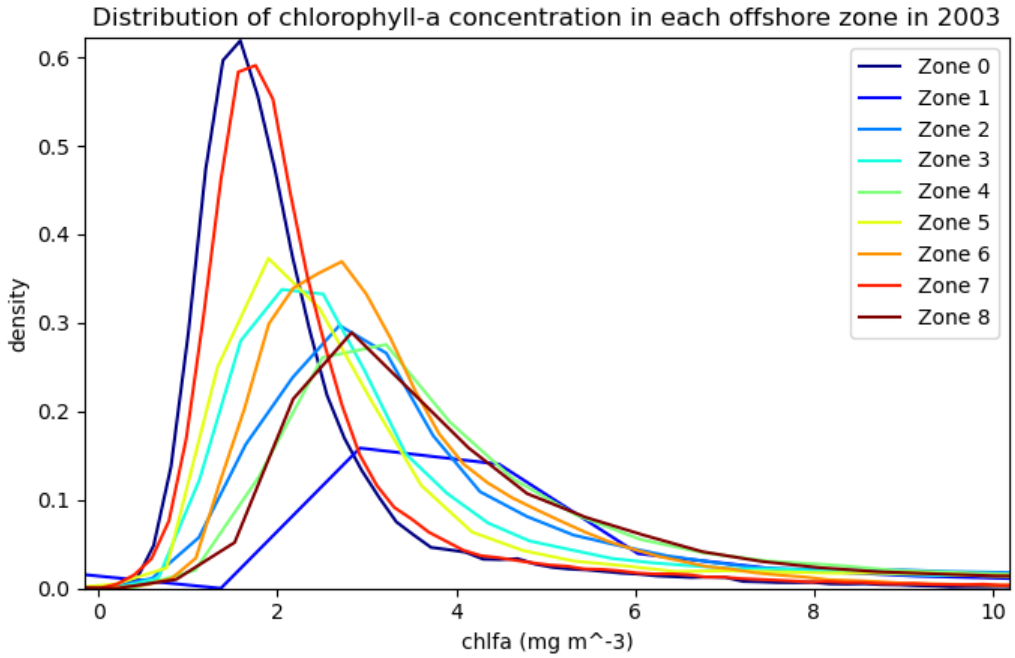


Figure 4.25: Distribution of the chlorophyll-a concentrations in the offshore zones in 2003.

Since Figure 4.25 is a zoomed in image, it is important to note what the shapes of the lines mean. A high peak in density near a low chlorophyll-a concentration on the x-axis means that overall the concentration is low, and high values of chlorophyll-a concentration have a very small probability. This is the case for zone number 0 and zone number 7. These two monitoring zones have similar distribution and more or less lie next to each other, recall from Figure 4.13. This would spark the idea that maybe these clusters should be combined, as is done in K-means clustering and hierarchical clustering. However, there is an important remark to be made about densities. Two zones can have the exact same density, but still be completely different on a day to day basis. Keeping this in mind, zone 0 and zone 1 should not necessarily be combined purely based on the fact that they are situated next to each other and have the same distribution.

Zone 1, the Rhine ROFI zone, has the lowest peak in distribution with the lower values of chlorophyll-a concentration. In other terms, zone 1 contains less chlorophyll-a concentrations that are low and more that are higher. The other zones all have a distribution with similar shape.

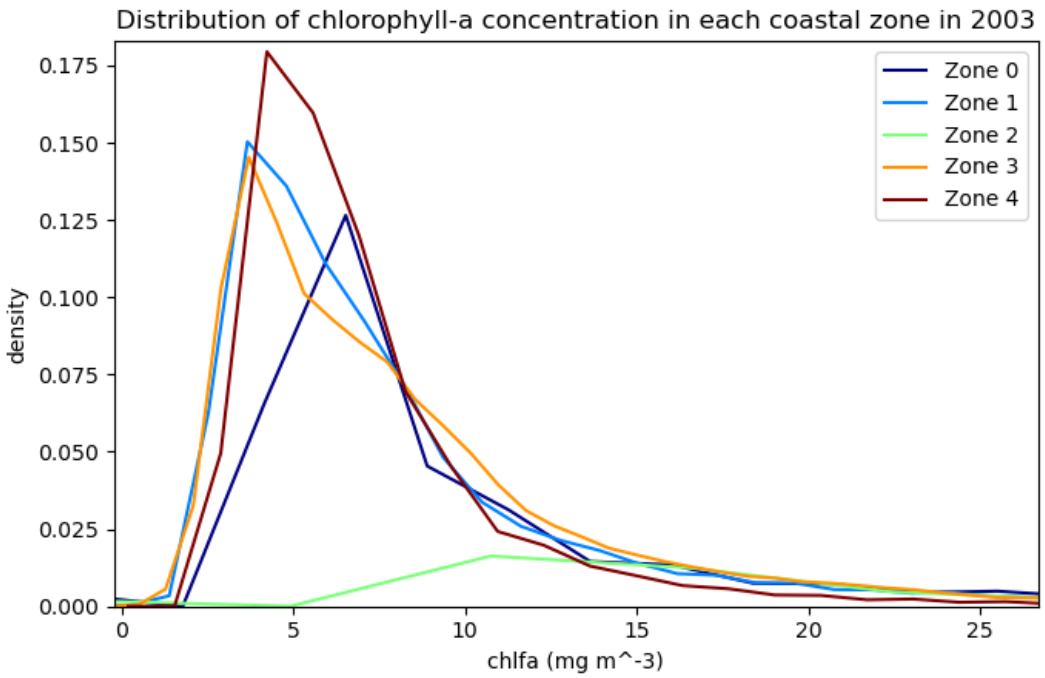


Figure 4.26: Distribution of the chlorophyll-a concentrations in the coastal zones in 2003.

Figure 4.26 portrays the empirical distributions of the zones in the coastal area.

Zone 2 exhibits a really low density for lower values of chlorophyll-a, thus higher values of chlorophyll-a are more common.

For zone 1 and 3 a similar conclusion can be drawn as for zone 0 and 7 in the offshore area.

Comparing the densities from the offshore area to the ones from the coastal area it becomes clear that for the coastal zone overall higher concentrations of chlorophyll-a are more likely. This is easily seen from the y-axes in the figures. For the offshore area the density of chlorophyll-a concentration below 10 reaches a maximum of 0.6. In the coastal areas the maximum is 0.175.

4.6. Distribution of one zone over time

All the analysis up till now have been conducted on the year 2003. An interesting analysis that can be done using the defined eutrophication monitoring zones is how the distribution of a certain zone changes over time. For the zone that contains the Rhine ROFI, zone number 8 in the offshore area, this is given in Figure 4.27. This figure is zoomed in for clarity, the full image is given in Appendix C

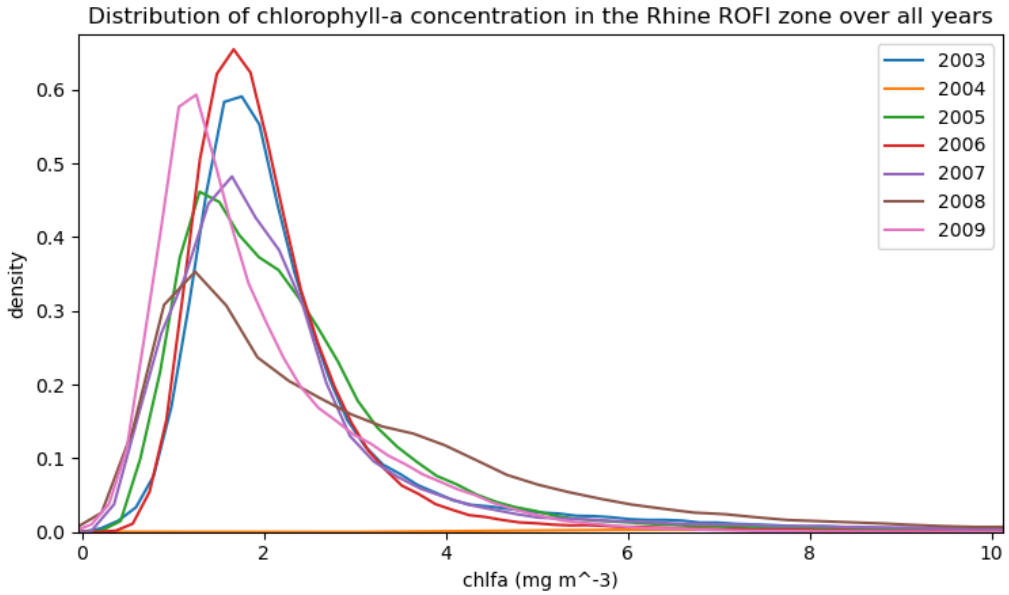


Figure 4.27: Distribution of the chlorophyll-a concentrations in the Rhine ROFI zone over the years.

From Figure 4.27 it can be concluded that the distribution of the Rhine ROFI zone does not stay the same over the years. This is also what was expected. Higher concentrations of chlorophyll-a concentration are as previously mentioned an indicator for eutrophication, which after a longer period results in poor marine health. The fluctuations are as follows: it starts out well in 2003, but in 2004 the chlorophyll-a concentrations are extremely high. Afterwards, the amount of high concentrations decrease in the years 2005, 2006, and 2007. In 2008 there is a small increase in concentrations again, but in 2009 this restored itself.

4.7. Comparison to other sets of zones within the Dutch North Sea coast

The previously defined eutrophication monitoring zones by OSPAR are not the only way that zones are defined within the Dutch part of the North Sea. It would be interesting to see whether other zones have any correlation to the eutrophication monitoring zones defined through clustering. The monitoring zones defined by this study can also serve as a recommendation to possibly reconsider the other identified zones. The types of maps that will be compared to the monitoring zones are selected because of their possible connection to eutrophication and thus the chlorophyll-a concentration. The maps considered in this section are:

- Maritime Zones and Common Fisheries Policy
- Marine Protected Areas
- Marine Spatial Planning
- Bathymetry

4.7.1. Marine limits and Common Fisheries Policy

The maritime limits and the Common Fisheries Policy are first set of the zones that were compared to the clustering results, see Figure 4.28. Chlorophyll-a concentration is an indicator for the amount of phytoplankton in the ocean. Normally, the phytoplankton is eaten by fish higher up in the food chain. When these fish are removed from the water through commercial fishing, the natural enemy of phytoplankton disappears. This could lead to more growth of the phytoplankton, which corresponds to eutrophication. Therefore, overfishing can increase eutrophication (Jackson et al., 2001).

The restrictions of where fishing is allowed and in what quantities are described through zones defined by the Common Fisheries Policy. Within three miles of the shore, the restrictions are the toughest, thus less fishing occurs here. This can also be seen in the clustering of the chlorophyll-a concentration. The border of the separation between the coast and offshore, created in section 4.1, resembles the three-mile zone in the common fishery policy. The different levels in the coastal area that were found when separating the coastal area and the offshore area somewhat resemble the different zones of the Common Fisheries Policy.

Besides the zones depicted in Figure 4.28, fishing is also prohibited in certain areas described by spatial planning. For example, fishing is not allowed in Marine Protected areas, or in wind turbine farms (Noordzeeloket, nd).

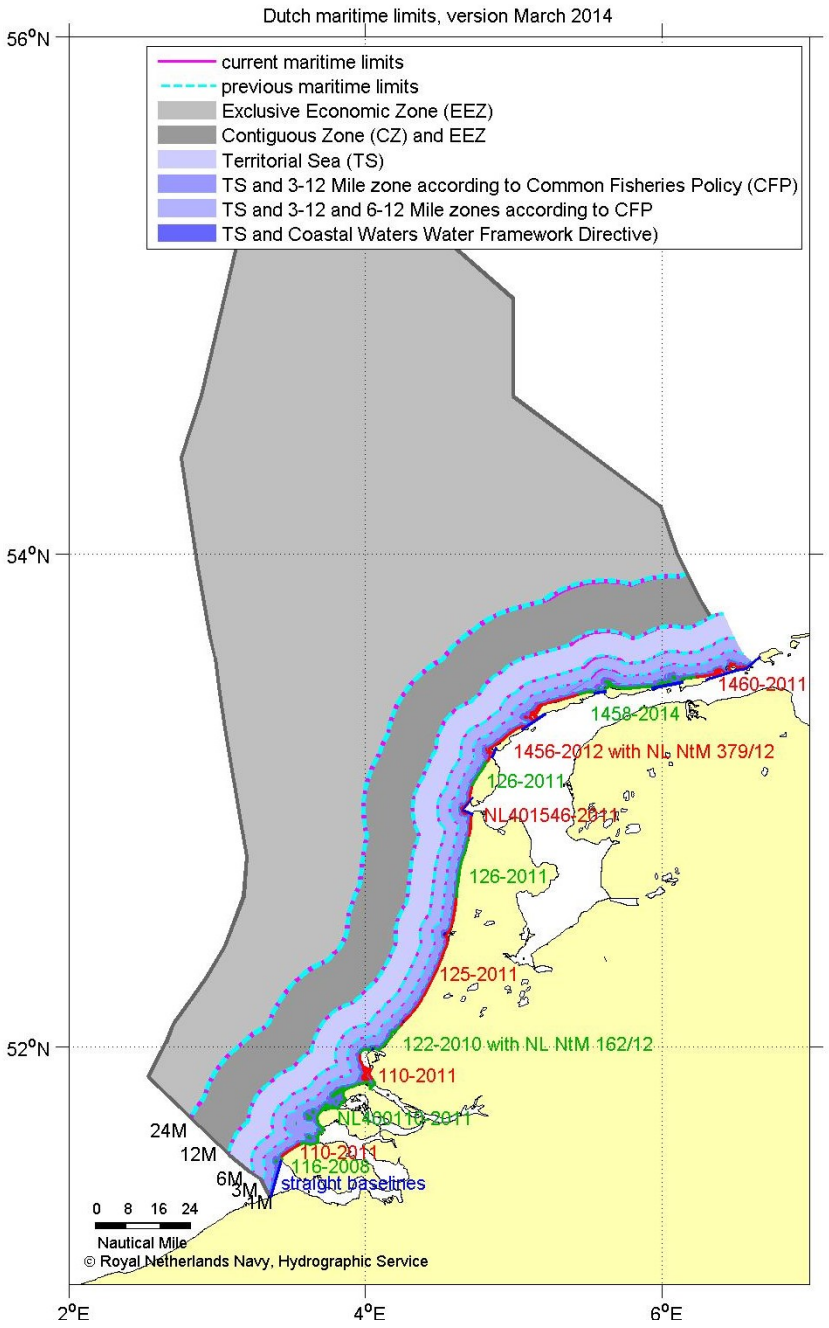


Figure 4.28: Division of the North Sea into maritime zones with the Common Fisheries Policy (Navy, 2014)

4.7.2. Marine Protected Area's

Another set of zones in the Dutch continental shelf is the legally designated marine protected areas, also known as Natura 2000, see Figure 4.29. Marine protected areas are shielded from bottom-impacting fishing with the goal to improve marine health in these areas. This means that the chlorophyll-a concentration in these areas is either high because of poor marine health, or low because they are already recovering. Therefore, it is likely that there is a correlation between the chlorophyll-a concentration and the marine protected area's.

Specifically the areas "Noordzeekustzone", denoted by E, and "Voordelta", denoted by G are zones that correspond to the monitoring zones defined using clustering. The "Noordzeekustzone" is almost exactly equal to zone number 2 and 4 combined in random forest clustering in the coastal area. Zone number 0 matches with the "Voordelta". Thus, the zones defined within the coastal area are also apparent in the marine protected areas. Since the "Noordzeekustzone" is split into two zones by the result from the clustering algorithm, it may be beneficial to redefine the marine protected area's. The results of the clustering of the coastal area from Section 4.3 may be an indication that one area in the "Noordzeekustzone" has a better marine health than the other.

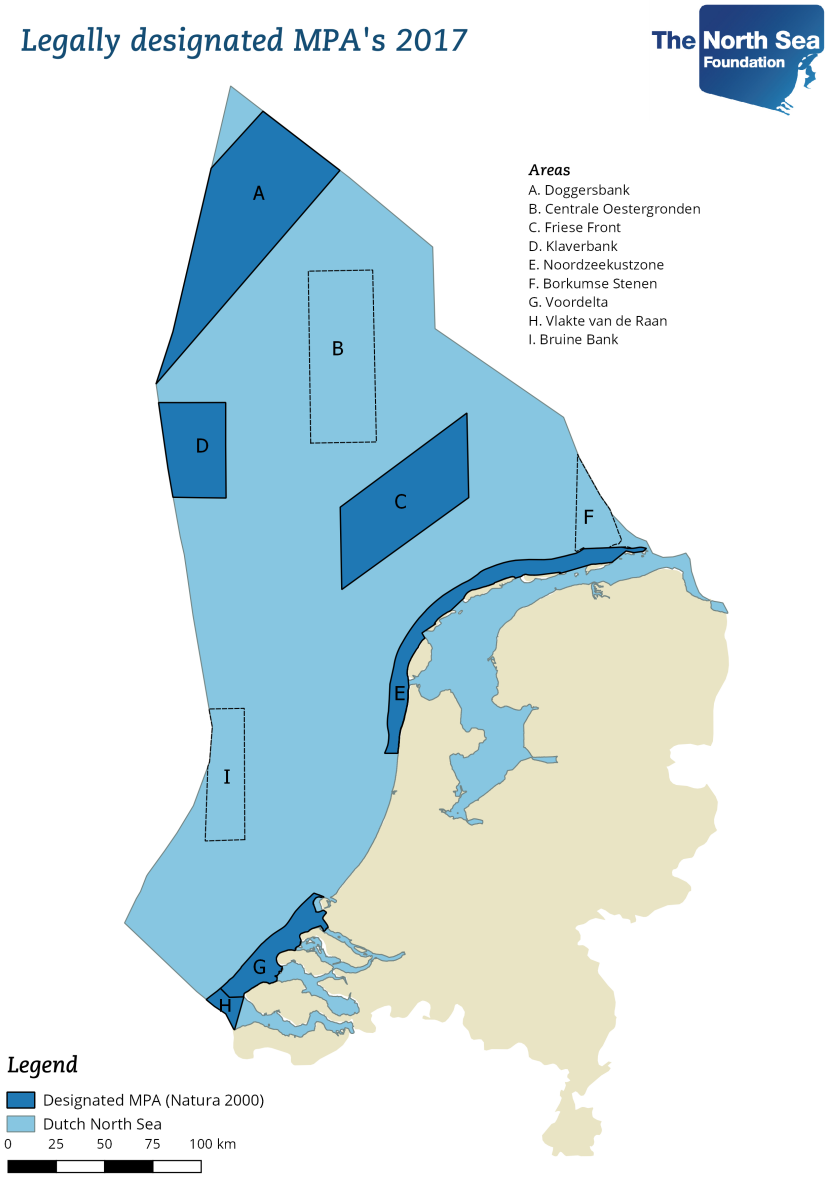


Figure 4.29: Marine Protected areas in the Dutch North Sea (The North Sea foundation, 2017)

4.7.3. Dutch government's spatial policy

The Dutch government has created a marine spatial planning policy that contains rules as to where, for example, windmill parks can be built. This policy can be found in Figure 4.30. This figure also contains the marine protected areas from the previous section. After comparing the policy to the monitoring zones defined through clustering, the only correlation that can be verified is with the green and yellow zone: Natura 2000 and potentially ecologically valuable area. Even though the windmill parks have an effect on the fisheries and through that may affect eutrophication, this is apparently not the case. The rest of the spatial planning attributes do not seem to be connected to eutrophication.

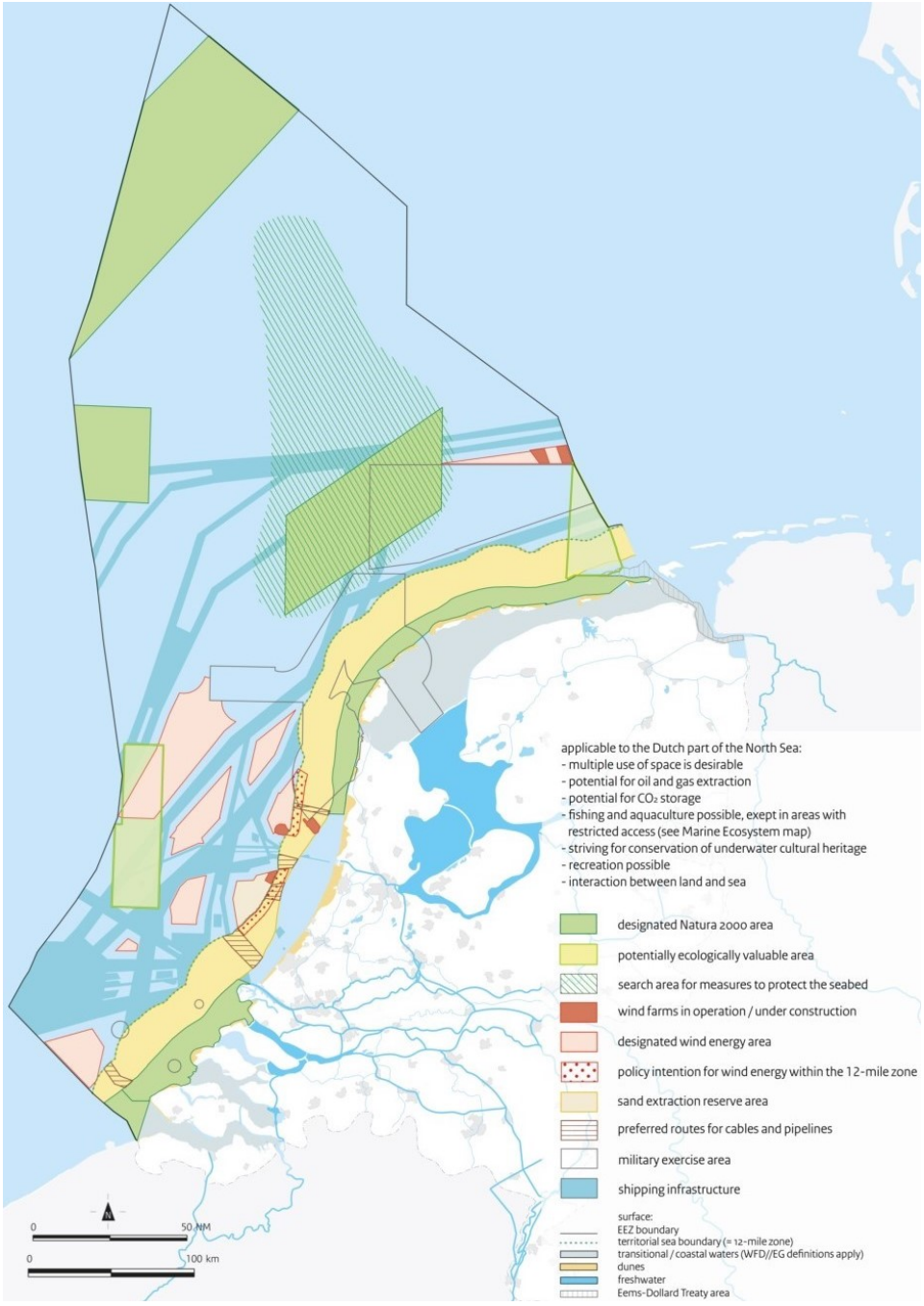


Figure 4.30: Marine spatial planning of the Dutch Continental Shelf (Ministry of Infrastructure and Water Management, 2015)

4.7.4. Bathymetric chart

In Figure 4.31, a bathymetric chart containing data on the height of the seabed of the Dutch North Sea is shown. Previously, a correlation between the barthymetry and chlorophyll-a was established in the southern ocean by Song and Ke (2015). Comparing the bathymetry of the Dutch North sea to the monitoring zones defined by clustering, yields a few mentionable observations.

Firstly, the Sand Wave Field region North of the Waddensea corresponds to zone 4 in the random forest clustering offshore. This zone was also clearly visible when looking at the chlorophyll-a concentration.

Secondly, the distinction between the coastal area and the offshore area from Section 4.1 is also clear in Figure 4.31 through the difference between the red and yellow parts.

All in all, the defined monitoring zones by clustering are partly validated by the bathymetry of the Dutch North Sea.

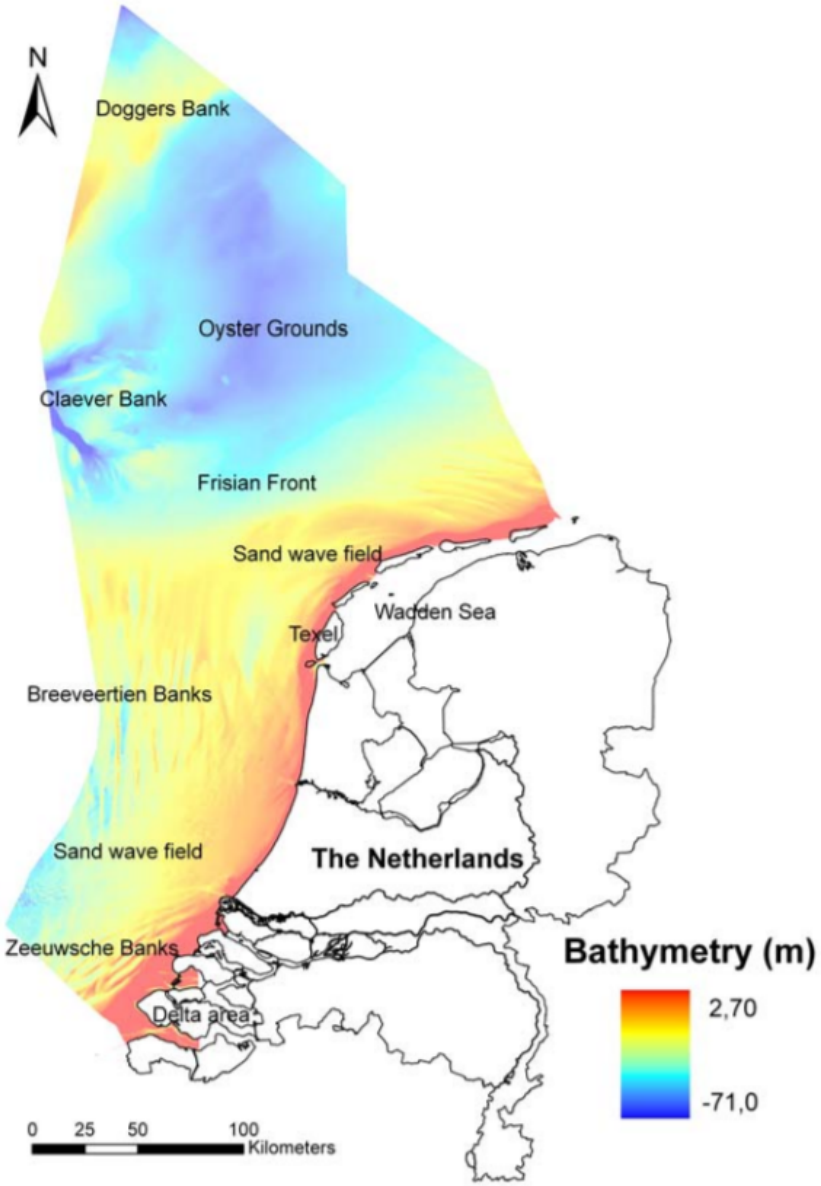


Figure 4.31: Dutch North Sea with frequently used names plotted on the bathymetry (Doornenbal and van Heteren, nd)

5

Conclusion

This report aimed to determine eutrophication zones with similar behaviour in chlorophyll-a concentration based on the statistical history found in satellite data of the Dutch North Sea via clustering algorithms. A secondary goal was to look at which of the four selected clustering algorithms yielded the best results. The term 'best' was defined using four criteria:

- They resemble OSPAR's already existing eutrophication monitoring zones.
- The zones that were expected to be clustered visually were clustered by the algorithm.
- Their result was confirmed though the accurate HDBSCAN clustering result.
- The clustering result yielded the best values for the validation metrics.

Both of these goals have been reached within this thesis.

The coastal area and the offshore area were first separated by K-means clustering with two clusters on only the chlorophyll-a concentration, and not the location. After this partition was made, the offshore area and coastal area were both clustered separately. The clustering algorithms used were K-means clustering, Hierarchical clustering, Random Forest clustering, and HDBSCAN.

Out of the four clustering algorithms, the Random Forest clustering outperformed the rest significantly. Random Forest clustering is hierarchical clustering with average linkage and precomputed distances based on the proximity derived from a random forest trained on the data. For the offshore area, nine clusters were necessary to represent the data best. The coastal area was clustered with five clusters. The result of these clustering algorithms combined can be found in Figure 5.1.

An additional conclusion can be drawn with respect to the metrics used to determine the quality of the clustering. Namely, the metrics used in this study, the elbow method, the silhouette average, and the gap statistic, are not good measures when determining monitoring zones. This is because they measure the distances between the data points within a cluster. However, for monitoring zones, similarity is not the only objective.

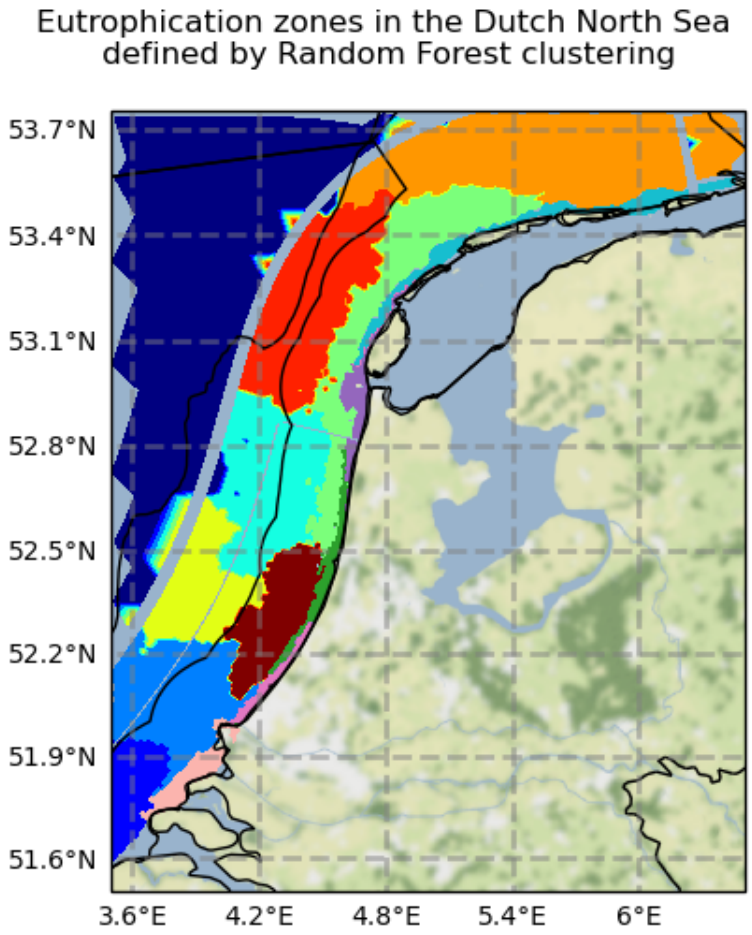


Figure 5.1: Combination of Random Forest clustering with 9 clusters in the offshore area and 5 clusters in the coastal area

The result of the defined monitoring zones can be used to monitor the changes in distribution of chlorophyll-a concentration in the Dutch North Sea. The idea of using clustering algorithms can also be applied to other locations to define more accurate monitoring zones. The monitoring zones defined through clustering give a more accurate representation of the eutrophication status than OSPAR's eutrophication zones.

Based on the result from this study, it is recommended that the idea of using clustering algorithms to determine other types of zones is researched further.

6

Discussion

Even though the research of this thesis completed the goals set at the start, there are always improvements that can be made. These improvements can be carried out through further research.

The first improvements are related to the data. The data used in this thesis was interpolated to create a full image of the North Sea. This interpolation could lead to results being slightly off. Alternatively, raw data could be used.

Another enhancement of this report would be to use all of the data available in the North Sea, or at least all of the data in the Dutch Continental Shelf. Then, the region chosen would be more logical.

Moreover, the analysis of this report were conducted on the year 2003 and the satellite data available was from 2003 up to and including 2009. The choice of using the first year available was made so that the results could be compared to later years. However, this could have been done more extensively. The clustering results could also be refined by including other years as well. Additionally, using more up to date data would yield results that are more functional.

The last improvement to the data could be the precision of the border between the North Sea and areas that have a different different system dynamics. By areas with different system dynamics the Waddensea, the 'Nieuwe Waterweg', the Westerschelde, The Grevelingenmeer, and the Oosterschelde are meant. These were removed since their chlorophyll-a concentration behaviour is different from the North Sea. The precision of the border could be slightly inaccurate due to the fact that is was done by hand following shapefile borders. However, finding a way to do this automatically would be better.

Another part of this study that could use improvement is the thoroughness of the use of the clustering algorithms. For hierarchical clustering, only Ward's Linkage was used. Other linkages, such as single linkage, average linkage, or complete linkage, were not considered in the comparison. Including these would yield a more in-depth comparison analysis.

The choice of the input parameters for HDBSCAN yielded some issues with the number of clusters to the extent that the choice was made to only use HDBSCAN result for validation of the other clustering algorithms. Further research could involve spending more time finding the optimal input parameters, since the algorithm did yield good results. An alternative would be to combine the clusters from HDSBSCAN based on some criteria.

Clustering of zones was done separately in the offshore area and the coastal area because there was not one day that contained all zones clearly distinguishable from each other. A possible solution for this would be to compute the probability distribution of each location, and cluster based on those. A big advantage would then be that a lot more data would be utilised, increasing the reliability. A disadvantage could be that two data points could have the same distribution, but peak on completely different

days thus not belonging together in a zone. This possibility could be explored further.

This thesis only used chlorophyll-a concentration as an indicator for eutrophication. However, there are four other measures needed to be completely sure of the eutrophication status. Further research could include other measures being included in the clustering.

Bibliography

Aggarwal, C. C. and Reddy, C. K. (2014). Data clustering. *Algorithms and applications. Chapman&Hall/CRC Data mining and Knowledge Discovery series, Londra*.

Arthur, D. and Vassilvitskii, S. (2006). k-means++: The advantages of careful seeding. Technical Report 2006-13, Stanford InfoLab.

Baretta-Bekker, J. and Prins, T. (2013). Assessments of phytoplankton in the netherlands and neighbouring countries according to ospar and wfd. (Deltares and BarettaBekker).

Beckers, J.-M. and Rixen, M. (2003). Eof calculations and data filling from incomplete oceanographic datasets. *Journal of Atmospheric and oceanic technology*, 20(12):1839–1856.

Blaas, M. (2013). Eutrophication assessment using remotely sensed and in situ chlorophyll-a data. (Deltares).

Bock, T. (n.d.). What is a dendrogram? https://www.displayr.com/what-is-dendrogram/ (Accessed: 24-06-2020).

Breiman, L. (2001). Random forests. *Machine learning*, 45(1):5–32.

Breiman, L. and Cutler, A. (1912). Random forests manual v4.0. Technical report, UC Berkeley.

Campbell, J. W. (1995). The lognormal distribution as a model for bio-optical variability in the sea. *Journal of Geophysical Research: Oceans*, 100(C7):13237–13254.

Campello, R. J., Moulavi, D., and Sander, J. (2013). Density-based clustering based on hierarchical density estimates. In *Pacific-Asia conference on knowledge discovery and data mining*, pages 160–172. Springer.

Ministry of Infrastructure and Water Management (2015). Policy document on the north sea 2016-2021 (webversie). including the netherlands' maritime spatial plan appendix 2 to the national water plan 2016-2021.

The North Sea foundation (2017). Legally designated marine protected areas. https://www.noordzee.nl/marine-protected-areas-in-the-dutch-north-sea/ (Accessed: 23-06-2020).

Dasgupta, S. (2008). *The hardness of k-means clustering*. Department of Computer Science and Engineering, University of California ….

de Ruijter, W. P., Visser, A. W., and Bos, W. (1997). The rhine outflow: a prototypical pulsed discharge plume in a high energy shallow sea. *Journal of Marine Systems*, 12(1-4):263–276.

Doornenbal, P. and van Heteren, S. (n.d.). Bathymetric range map of the dutch continental shelf (ncp).

Ester, M., Kriegel, H.-P., Sander, J., Xu, X., et al. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In *Kdd*, volume 96, pages 226–231.

Ferreira, J. G., Andersen, J. H., Borja, A., Bricker, S. B., Camp, J., da Silva, M. C., Garcés, E., Heiskanen, A.-S., Humborg, C., Ignatiades, L., et al. (2011). Overview of eutrophication indicators to assess environmental status within the european marine strategy framework directive. *Estuarine, Coastal and Shelf Science*, 93(2):117–131.

Gini, C. (1912). Variabilità e mutabilità. *vamu*.

Jackson, J. B., Kirby, M. X., Berger, W. H., Bjorndal, K. A., Botsford, L. W., Bourque, B. J., Bradbury, R. H., Cooke, R., Erlandson, J., Estes, J. A., et al. (2001). Historical overfishing and the recent collapse of coastal ecosystems. *science*, 293(5530):629–637.

Jarník, V. (1930). O jistém problému minimálním. *Práca Moravské Prírodovedecké Spolecnosti*, 6:57–63.

Korte, B. and Nešetřil, J. (2001). Vojtěch jarník's work in combinatorial optimization. *Discrete Mathematics*, 235(1-3):1–17. translation of (Jarník, 1930).

Lloyd, S. (1982). Least squares quantization in pcm. *IEEE transactions on information theory*, 28(2):129–137.

Navy, R. N. (2014). Division of the north sea into maritime zones with the common fisheries policy. https://english.defensie.nl/topics/hydrography/maritime-limits-and-boundaries/netherlands-boundaries-in-the-north-sea (Accessed: 23-06-2020).

Noordzeeloket (n.d.). Dutch fisheries policy. https://www.noordzeeloket.nl/en/functions-and-use/visserij/ (Accessed: 30-06-2020).

OSPAR (2013). Common procedure for the identification of the eutrophication status of the ospar maritime area. Reference number: 2013-8.

OSPAR (2020a). About page. https://www.ospar.org/about (Accessed: 11-06-2020).

OSPAR (2020b). Chlorophyll-a. https://oap.ospar.org/en/ospar-assessments/intermediate-assessment-2017/pressures-human-activities/eutrophication/chlorophyll-concentrations (Accessed: 15-06-2020).

OSPAR (2020c). Eutrophication. https://www.ospar.org/work-areas/hasec/eutrophication (Accessed: 11-06-2020).

Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65.

Shi, T. and Horvath, S. (2006). Unsupervised learning with random forest predictors. *Journal of Computational and Graphical Statistics*, 15(1):118–138.

Sokal, R. R. (1958). A statistical method for evaluating systematic relationships. *Univ. Kansas, Sci. Bull.*, 38:1409–1438.

Song, C. and Ke, L. (2015). Bathymetrical influences on spatial and temporal characteristics of chlorophyll-a concentrations in the southern ocean from 2002 to 2012 (october to march) using modis. *Geo-spatial Information Science*, 18(4):200–211.

Thorndike, R. L. (1953). Who belongs in the family. In *Psychometrika*. Citeseer.

Tibshirani, R., Walther, G., and Hastie, T. (2001). Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(2):411–423.

Van der Giessen, A., De Ruijter, W., and Borst, J. (1990). Three-dimensional current structure in the dutch coastal zone. *Netherlands Journal of Sea Research*, 25(1-2):45–55.

Ward Jr, J. H. (1963a). Hierarchical grouping to optimize an objective function. *Journal of the American statistical association*, 58(301):236–244.

Ward Jr, J. H. (1963b). Hierarchical grouping to optimize an objective function. *Journal of the American statistical association*, 58(301):236–244.

Appendices

A

Visualisation of metrics to determine number of clusters

A.1. Visualisation of metrics for the separation of the offshore area and coastal area

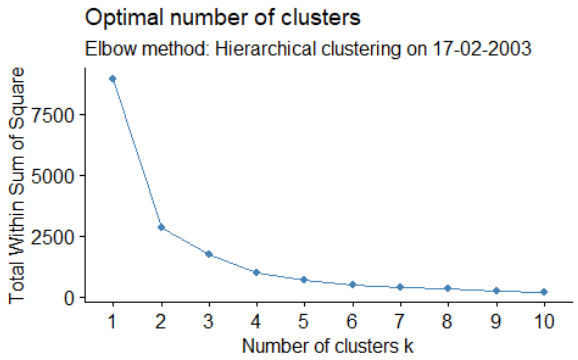


Figure A.1: The elbow method for hierarchical clustering on 17-02-2003

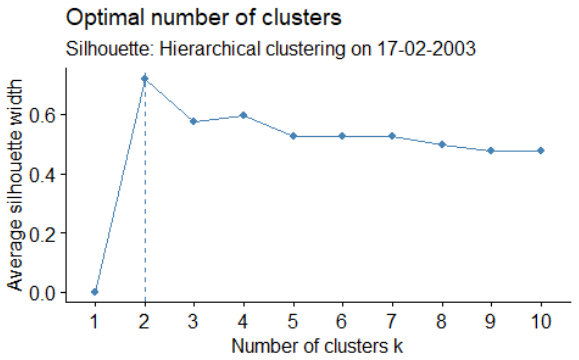


Figure A.2: The silhouette method for hierarchical clustering on 17-02-2003

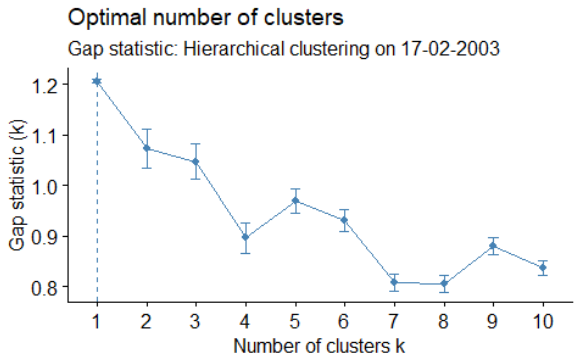


Figure A.3: The gap method for hierarchical clustering on 17-02-2003

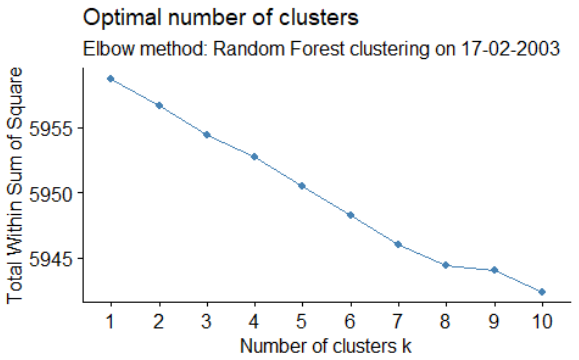


Figure A.4: The elbow method for random forest clustering on 17-02-2003

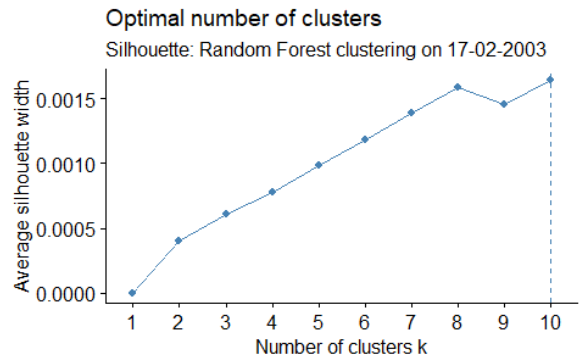


Figure A.5: The silhouette method for random forest clustering on 17-02-2003

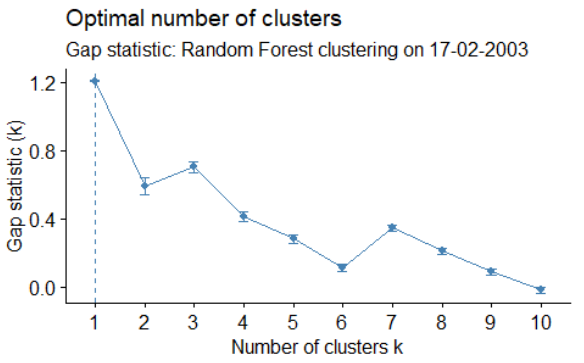


Figure A.6: The gap method for random forest clustering on 17-02-2003

A.2. Visualisation of metrics for the offshore zones

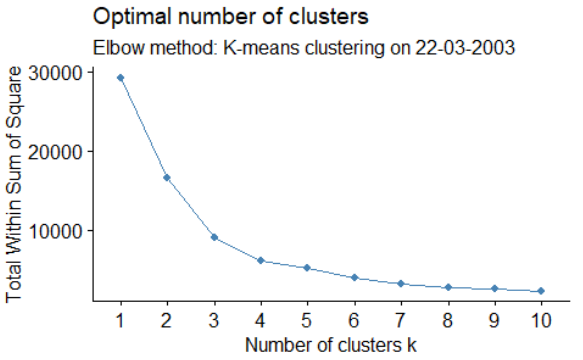


Figure A.7: The elbow method for K-means clustering on 22-03-2003

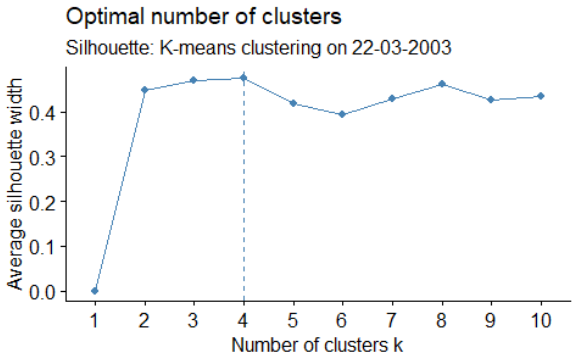


Figure A.8: The silhouette method for K-means clustering on 22-03-2003

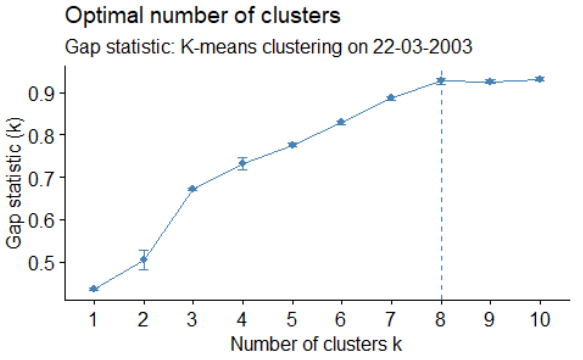


Figure A.9: The gap method for K-means clustering on 22-03-2003

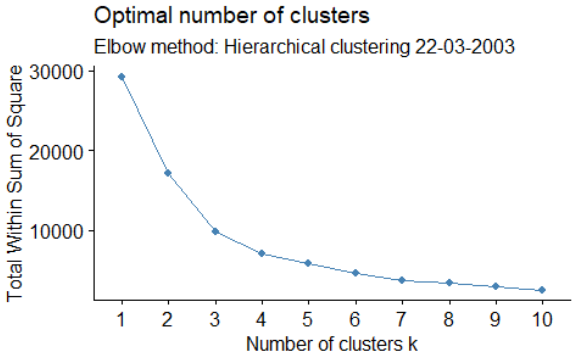


Figure A.10: The elbow method for hierarchical clustering on 22-03-2003

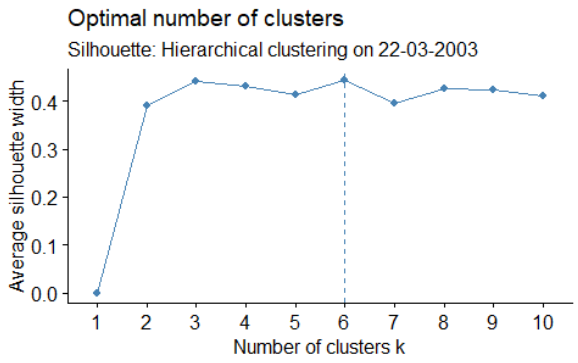


Figure A.11: The silhouette method for hierarchical clustering on 22-03-2003

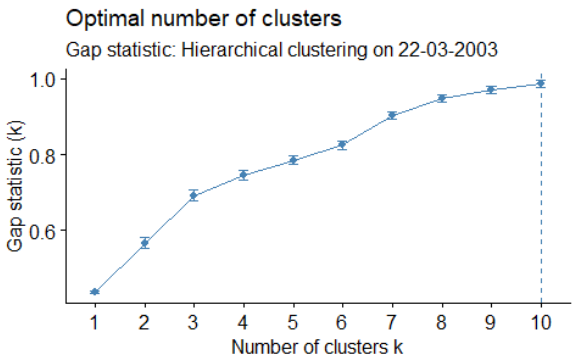


Figure A.12: The gap method for hierarchical clustering on 22-03-2003

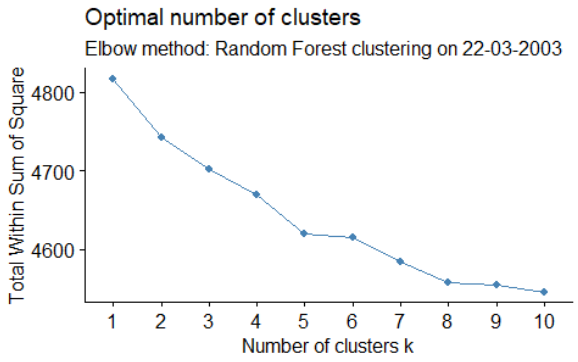


Figure A.13: The elbow method for random forest clustering on 22-03-2003

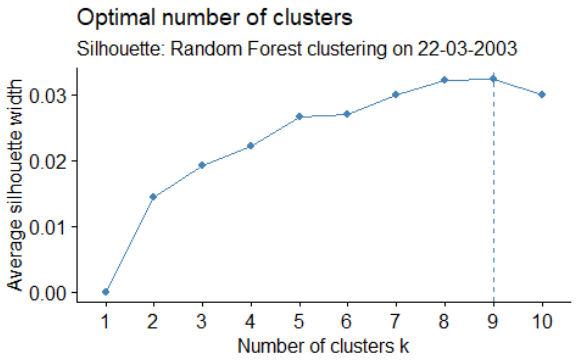


Figure A.14: The silhouette method for random forest clustering on 22-03-2003

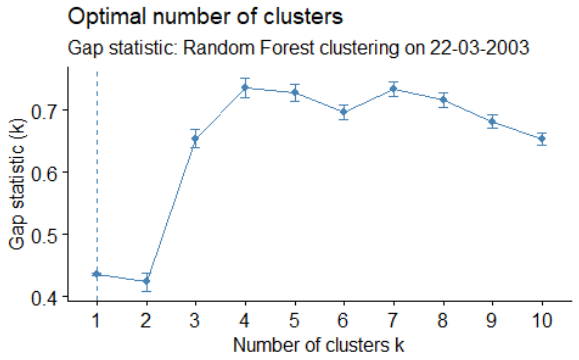


Figure A.15: The gap method for random forest clustering on 22-03-2003

A.3. Visualisation of metrics for the coastal zones

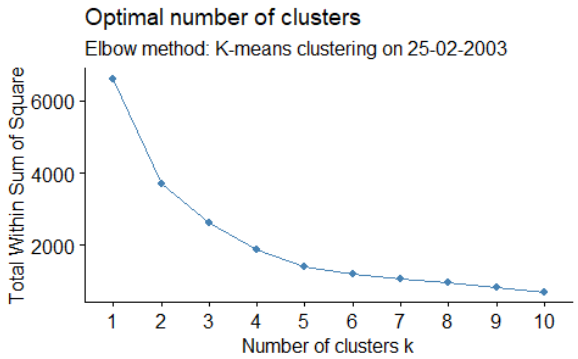


Figure A.16: The elbow method for K-means clustering on 25-02-2003

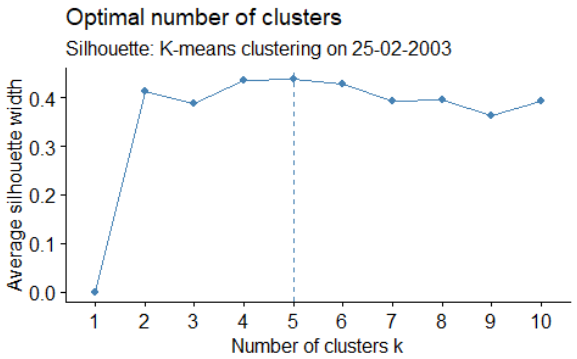


Figure A.17: The silhouette method for K-means clustering on 25-02-2003

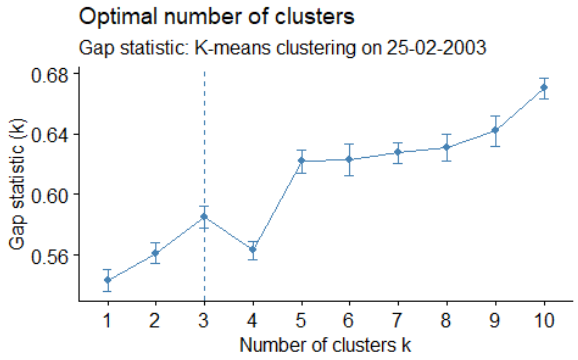


Figure A.18: The gap method for K-means clustering on 25-02-2003

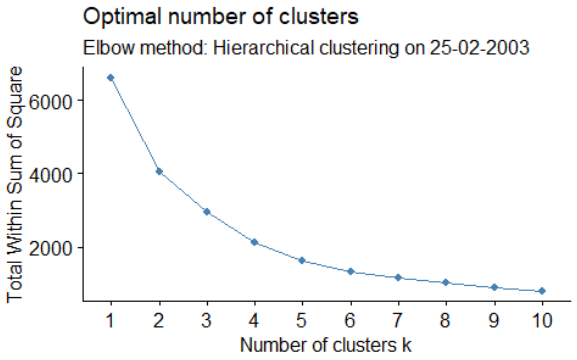


Figure A.19: The elbow method for hierarchical clustering on 25-02-2003

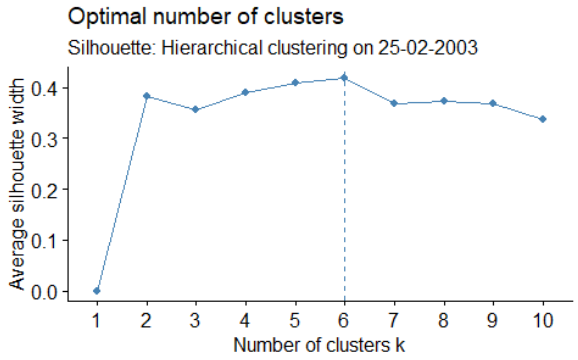


Figure A.20: The silhouette method for hierarchical clustering on 25-02-2003

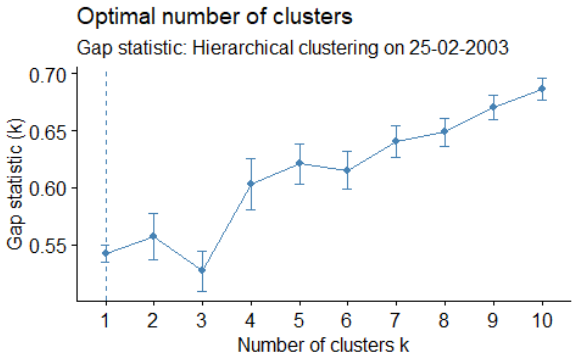


Figure A.21: The gap method for hierarchical clustering on 25-02-2003

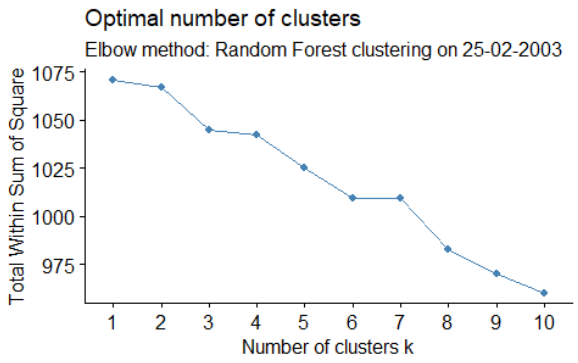


Figure A.22: The elbow method for random forest clustering on 25-02-2003

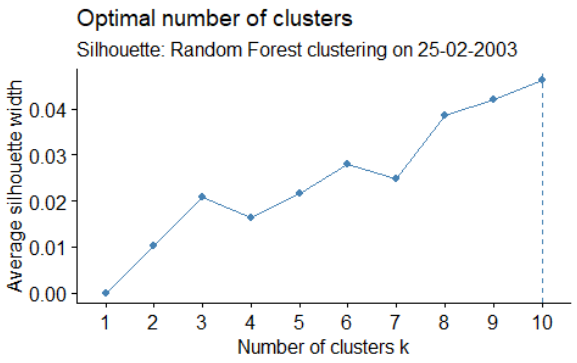


Figure A.23: The silhouette method for random forest clustering on 25-02-2003

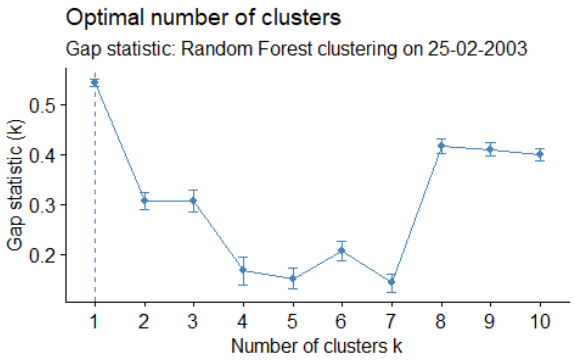


Figure A.24: The gap method for random forest clustering on 25-02-2003

B

Fitted log-normal distribution for the defined eutrophication monitoring zones

Zone	Mean	Standard Deviation
0	1.3391	0.7576
1	1.8553	0.6520
2	1.8638	0.5603
2	1.8699	0.6675
3	1.6035	0.5839

Table B.1: Parameters of the fitted Log-normal distribution for the zones in the coastal area

Zone	Mean	Standard Deviation
0	0.5812	0.5233
1	1.3059	0.7489
2	1.3161	0.7205
3	1.0802	0.7193
4	1.3443	0.6350
5	1.0619	0.7961
6	1.0961	0.4508
7	0.7327	0.5155
8	1.3975	0.6909

Table B.2: Parameters of the fitted Log-normal distribution for the zones in the offshore area

C

Fully zoomed out figures of distribution of chlorophyll-a concentration

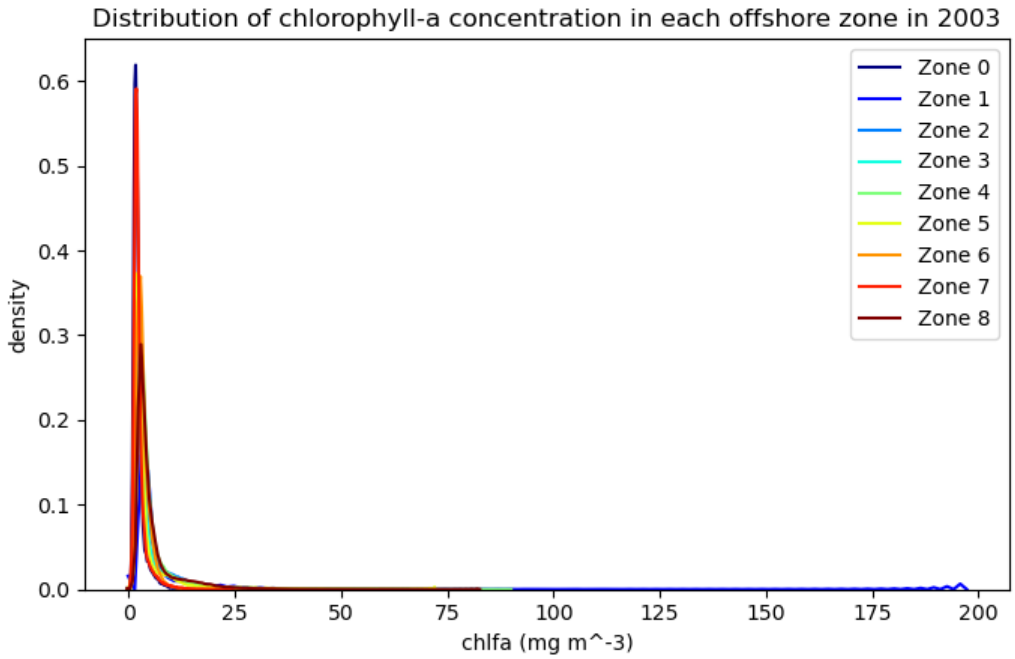


Figure C.1: Distribution of the chlorophyll-a concentrations of the zones in the offshore area zoomed out.

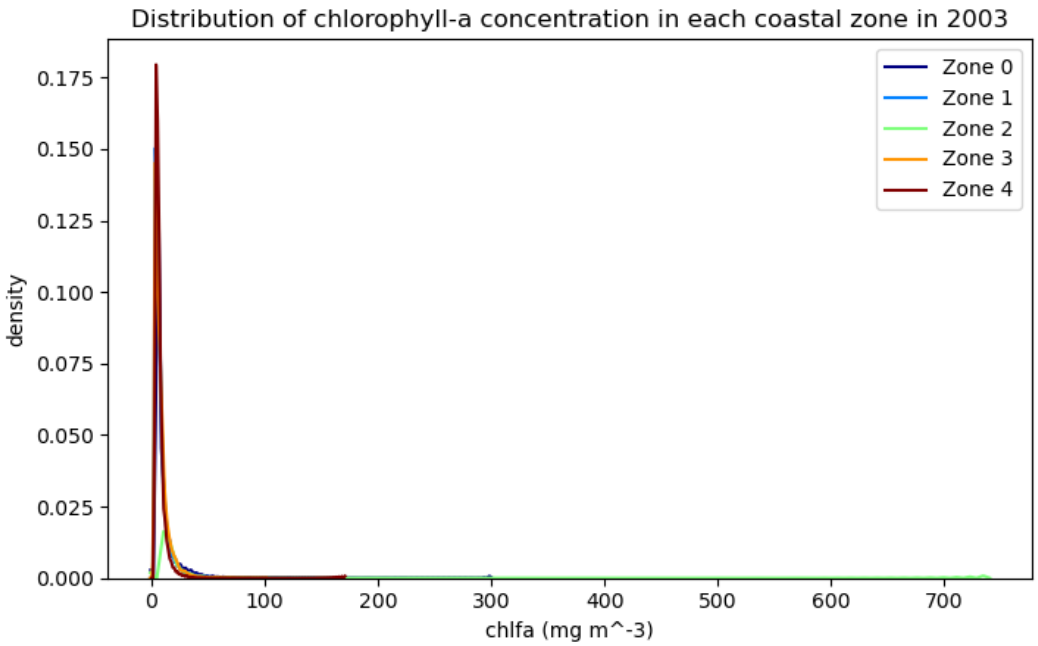


Figure C.2: Distribution of the chlorophyll-a concentrations of the zones in the coastal area zoomed out.

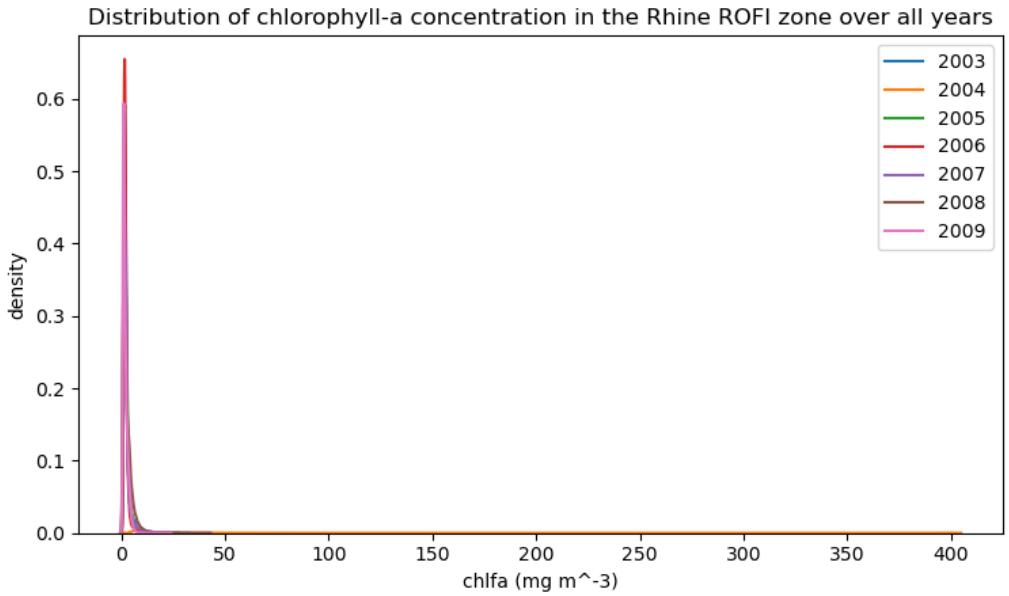


Figure C.3: Distribution of the chlorophyll-a concentrations of the Rhine ROFI zones over the years zoomed out.

D

Code

The implementation of the clustering algorithms together with the visualisations found in this report can be accessed through the following link:
`https://github.com/LauraVeerhoek/BEPeutrophication/`