university of groningen

faculty of science and engineering

Statistical learning methods for environmental DNA

Internship MSc Applied Mathematics

June 2019

Student: A.G. Wiersma

Company/Institute: Witteveen + Bos, Deventer

Internal UoG supervisor/first assessor: Prof. dr. A.J. van der Schaft

External supervisor: Dr. ir. A.C. de Niet

Internal UoG second assessor: Dr. M.A. Grzegorczyk

Abstract

In this report, the results of my internship about the data analysis of environmental DNA are presented. Environmental DNA is a new technique that gives a better insight which factors determine the water quality of lakes. This technique is being developed by Witteveen+Bos and Datura. Cluster analysis was used to check if the environmental DNA profiles of the lakes could be used to categorize the lakes. There were several methods considered. Namely, the dissimilarity coefficients: Euclidean, Correlation, Bray-Curtis and Jaccard were used in conjunction with the linkage methods: Single, Complete, Average, Weighted and Ward. In order to find the best method for this type of dataset the clustering results of the methods with several filters on the dataset were being compared. This was done to check for the robustness of the methods. Principal component analysis was used in order to visualize the clustering results and to see if there were there were factors that are characteristic for a cluster.

Contents

1 Introduction

A good method to measure the water quality of lakes, and the effects of any measures that are taken is essential for water management. It is especially important to understand which factors determine the water quality. With current measurement-techniques a part of these factors can be determined. However, there are still a lot processes that are unknown, especially for bacteria. There is a new technique called environmental DNA, which is being developed by Witteveen+Bos and Datura, that could be a better technique that is quicker, cheaper and gives a better insight which factors determine the water quality. It is a long running project that lasts several years. A first step is looking for methods that can analyze the data gathered from environmental DNA. This data is gathered from twenty lakes in the Netherlands. The locations of the lakes can be found in Figure 1. More information about the project can be found in [10] (in Dutch). In this report an overview of several unsupervised statistical learning techniques, that can be used to analyze environmental DNA, are given.

The internship was done at Witteveen+Bos which is an engineering and consultancy firm that offers solutions in the fields of water, infrastructure, environment and construction projects. With a network of 19 offices in 11 countries there are around 1100 engineers and consultants.

The idea behind environmental DNA, also called eDNA, is that each sample has it's own unique eDNA profile. Therefore, a sample from a murky lake filled with seaweed has a different profile than a sample from a clear lake with a high diversity in water plants. The objective of the project is to find a so called fingerprint for eDNA which determines the water quality of the water sample taken from the lake.

In order to analyze the data from the eDNA statistical learning is used. Statistical learning can generally be divided into two categories, supervised and unsupervised learning. In supervised learning there is an outcome measurement or categorical that we want to predict based on the variables (observations). With unsupervised learning only the variables are observed and there is no outcome. Instead the task is to describe how the data is clustered or organized.

The methods that are discussed in this report, cluster analysis and principal component analysis, are both unsupervised learning methods. We use unsupervised learning because currently there is not enough data to really predict the outcome (water quality) based on the variables (eDNA) of the samples. Later on in this project supervised learning methods would be recommended to predict the water quality of a sample based on eDNA, and to see which factors influence the water quality. For now unsupervised learning is used to see whether or not the eDNA profiles of the water samples can be used to categorize the dataset. When this is the case this gives more confidence that the eDNA profiles can be used to profile an water sample and later on used as an indication of the water quality. In addition the methods are used to analyze the data to see if there are some factors that are characteristic to some clusters which may influence the water quality.

In the next section it is explained how the eDNA samples are determined, and what the dataset that contains these samples looks like. In Section 3 several methods for cluster analysis are described. Section 4 describes how principal components analysis works. Which software has been used for the data analysis is described in Section 5. Section 6 presents the results of the cluster analysis. In Section 7 the results of the principal component analysis are presented. Finally in the last section the conclusion of this report and the recommendation is presented.

Figure 1: Locations of the water samples taken in 2018 per water management. The color of a location indicates which water management is responsible for this location.

2 Environmental DNA

In this section it is described how the eDNA samples are determined. In addition the structure of the dataset that is used for analysis is described.

As explained in the introduction the idea behind eDNA is that each water sample has a unique eDNA profile. The procedure of determining the eDNA of a water sample is as follows. From the center of the lake several water samples are taken and combined in a sample such that the sample contains around one liter water. The water sample contains both direct and indirect DNA. Direct DNA comes from species that are entirely contained in the sample like bacteria. Indirect DNA comes from species that left behind feces, urine or skin cells like fish. With eDNA we mean all the DNA that is found in a sample. The water sample is then taken to the laboratory for DNA analysis.

In order to analyze the DNA in the sample polymerase chain recation (PCR) is used to increase the amount of DNA in a sample so that it can be detected during the analysis. The process is as follows. A primer binds itself to two opposing stands of the DNA. Which part of the DNA the primer binds itself to depends on the primer. There is a primer that binds itself to DNA that is characteristic for bacteria, eukoryota and fish. During the PCR the region that lies between these DNA strands will be copied. More information about this process can be found in [11].

We call each unique DNA sequence that is detected in a sample an OTU. This stands for Operational Taxonomic Unit. Sometimes an OTU is unique for a species. However, it can also occur that an OTU is the same for a group of related species for a higher part on the taxonomic level. For instance it is found that the OTU is a particular kind of bacteria. By a taxonomic level we mean the relative level of a group of organisms in a taxonomic hierarchy of the biological classification. The main eight taxonomic levels are shown in Figure 2.

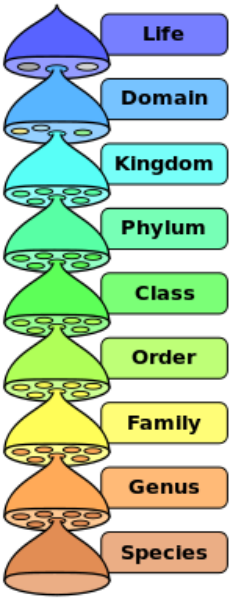


Figure 2: The main eight taxonomic levels.

In order to determine which species an OTU belongs to a reference database is used. When the reference database is used to find the matching species for an OTU it will look for the registered DNA sequence that is the best match. This match is calculated as a ratio which we call best identity. When the best identity of an OTU is 1.0 it should be a perfect match. However even for this perfect match it does not always mean that the OTU is indeed equal to the found species. Sometimes there is a genetic variation within a species, a error occurred in the lab or the OTU is not yet in the reference database. When a OTU has a best identity that is lower than 0.85 it is dropped from the dataset. This threshold has been decided by Witteveen+Bos and Datura.

2.1 Choice of primer

There are five different primers that can be used for the PCR. The difference between these primers is which part of the DNA it binds itself to. With a universal primer as much as possible DNA will be detected. There is also a bacterial primer which binds itself to DNA characteristic of bacteria so that after the PCR only DNA from bacteria is detected. Furthermore there is an eukoryota primer and two primers for fish.

Before we can begin analysis the data we first need to choose which primer is used. This can be the universal primer or the combination of the bacterial, eukoryota, the two fish primers and universal primer (also called merged primer). For the merged primer the Universal primer is used to determine the ratio's of the bacteria, eukoryota and the fish. For the merged dataset the data for fish using the two fish primers is also needed as this was not yet available for the rest of this report only Universal will be used.

2.2 Dataset

The structure of the eDNA dataset[1] is as follows. Every row in the dataset corresponds to a unique OTU, this is a DNA sequence of a species found in the sample. Sometimes the index number of an row is used to indicate that particular OTU. For this species the name of that species at the taxonomic levels is given in the associated columns. An example of a couple of taxonomic levels can be seen in Figure 3. There are a lot of names of a species per taxonomic level unknown because the reference database that is used to determine the species is not complete.

Index	best_identity	superkingdom_name	kingdom_name	phylum_name	class_name	subclass_name
0	0.902	Eukaryota	nan	nan	Spirotrichea	Choreotrichia
1	1	Eukaryota	Viridiplantae	Chlorophyta	Chlorophyceae	nan
2	0.8636	Bacteria	nan	Actinobacteria	nan	nan
3	0.9882	Bacteria	nan	Firmicutes	Negativicutes	nan
4	0.8991	Eukaryota	Metazoa	nan	nan	nan
5	0.88	nan	nan	nan	nan	nan
6	0.8506	nan	nan	nan	nan	nan

Figure 3

In addition, there are a lot of OTUs that do not appear in the eDNA database at all on any taxonomic level or match with a wrong species in the database. The percentage of unknown species per taxonomic level is given in Tabel 1.

The rest of the columns consist of the samples and per OTU the number of reeds of that OTU in that sample. The name of a column that contains the number of reeds is the identification of the sample. Which is in the format: WM.LOC.YRWK.w, where WM is the abbreviation of the water management, LOC the first three letters of the name of the lake, YR the last two numbers of

[1]The eDNA data is contained in the excel file $final\_uni.xlsx$.

the year in which the sample is taken,WK the week number in which the sample is taken and the w at the end indicates that this is a water sample.

Table 1: The percentage of unknown species per taxonomic level is given in the column Unknown. The column Dimensions contains the number of variables (species) on that taxonomic level.

Taxonomic level	Unknown (%)	Dimensions
Superkingdom	14	4
Kingdom	78	5
Phylum	42	43
Infraclass	98	7
Class	38	100
Subclass	90	30
Order	41	203
Suborder	95	25
Superfamily	98	18
Family	51	309
Genus	54	436
Species	60	488
OTU	0	2720

The number of reeds indicate the number of times that particular DNA sequence (OTU) is found in the sample. An example of a couple of samples with the number of reeds of a OTU is given in Figure 4. For the analysis in this report the data from the samples that are taken in the year 2018 are used. The water samples are taken from 20 lakes in the Netherlands and the locations can be found in Figure 1. Generally there are water samples from each of these locations at four different dates. Namely in the weeks 19, 23, 27 and 32. From four locations there is only one water sample which is taken in week 31. In total there are 69 samples.

Index	WD.BEU.1825.w	WD.BOW.1825.w	NZ.PWM.1827.w	SK.KRP.1827.w
0	482	2102	2706	2811
1	0	0	0	0
2	39519	19810	18662	30778
3	0	0	0	0
4	0	0	0	0
5	0	0	0	0
6	372303	50631	13247	48015

Figure 4

Not all samples in the eDNA dataset are used for analysis. When samples has less than 20.000 reeds there is not a lot of biodiversity so it is likely that the sample does not contain all the species that are in the water. The threshold of 20.000 is chosen by Witteveen+Bos and Datura. There are 16 samples that

do not meet this requirement and are not used for the data analysis. Therefore there are 53 samples left. The total number of OTUs for these samples are 2720. Most of the OTUs are only in a few samples. We have that there are 2081 OTUs that occur in less then 5 samples. Therefore the dataset contains a lot of zeros.

2.3 Transformations

The eDNA dataset contains the amount of species present in a sample. Because species tend to grow exponentially when conditions are favourable a transformation can be used to reduce the asymmetry of the data [5]. Several commonly used transformations are the following:

i Species profile transformation
Using this transformation we get the ratio

$$\bar{y}_{ij} = \frac{y_{ij}}{y_{j+}}, \tag{1}$$

where y_{ij} is the value of species (rows) i in sample (columns) j, y_{j+} is the sum over the samples j for species i and $\bar{y}_{ij}$ is the new value.

ii Hellinger
This is a modification of the species profile transformation and is recommended by [5].

$$\bar{y}_{ij} = \sqrt{\frac{y_{ij}}{y_{j+}}}, \tag{2}$$

where y_{ij} is the value of species (rows) i in sample (columns) j and y_{j+} is the sum over the samples j for species i.

iii Log
The log transformation reduces the asymmetry of the species distributions

$$\bar{y}_{ij} = \log\left(y_{ij} + 1\right), \tag{3}$$

where y_{ij} is the value of species (rows) i in sample (columns) j. The one was added so that the variables that are zero in the original data are again zero.

iv Chi-square distance transformation
Using the Chi-square distance transformation we reduce the value of an abundant species more than that of a rare species. Therefore this transformation should only be used when we are sure that rare species are a good indication of special ecological conditions [5].

$$\bar{y}_{ij} = \sqrt{y_{++}}\frac{y_{ij}}{y_{j+}\sqrt{y_{i+}}}, \tag{4}$$

where where y_{ij} is the value of species (rows) i in sample (columns) j, y_i+ is the sum over the species i for sample j, y_{j+} is the sum over the samples j for species i and y_{++} is the sum of values over the whole data table.

2.4 Biomass of an OTU within a sample

Apart from the dataset containing the number of reeds of the OTUs per sample another measurement has been made. Namely, the amount of biomass in molecules/liter there is in a sample. Combining the two dataset we can get a dataset containing the biomass in molecules/liter of that OTU within a sample (called qPCR). We can do this as follows. First we perform the species profile transformation on the dataset containing the number of reeds. Now we get the ratio of a OTU within a sample. Second we multiply this value with the total biomass[2] in a sample.

[2]The total biomass per sample is in a separate file called $qPCR.xlsx$.

3 Cluster analysis

In this section several different methods for cluster analysis are described. The goal of cluster analysis is to partition the measured objects (samples) into subsets, also called clusters, such that the objects that are assigned in cluster are more similar to each other than to objects in a different cluster. We use cluster analysis for the eDNA dataset to see whether samples from similar locations cluster together, which shows that the eDNA profiles of these locations are also similar.

Cluster analysis algorithms can be divided into the following two categories.

- Partitional
  In Partitional clustering the data is partitioned in a predetermined number of non-overlapping clusters in a single step. Examples of partitional clustering methods are K-means, K-median and DBSCAN [3].
- Hierarchical
  In Hierarchical clustering the number of clusters is not predetermined and there are sub clusters allowed which are nested clusters that are organized as a tree. Such a tree is called a dendrogram and an example is seen in Figure 5. On top of the tree all samples are in one single cluster, while on bottom each sample builds its own individual cluster. There are two kinds of hierarchical cluster analysis agllomerative (bottom-up) and divisive (top-down). Bottom-up is the most commonly used and the most researched.

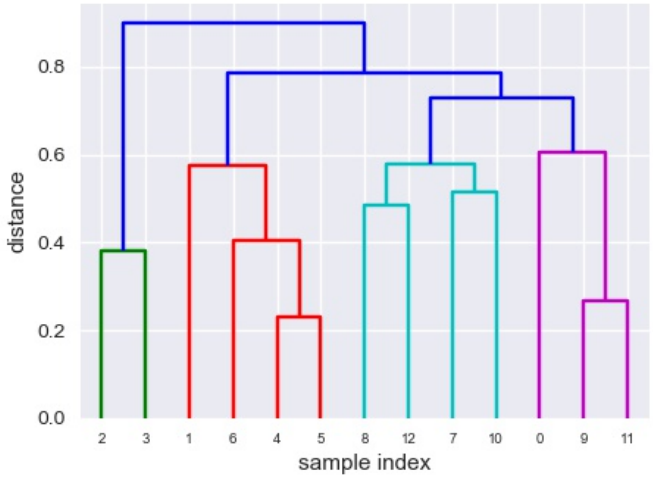


Figure 5: An example of a dendrogram, which is a result from a hierarchical cluster analysis.

For the eDNA dataset we will use agglomerative hierarchical clustering because in that case the number of cluster does not have to be predetermined. In addition

the number of clusters can be chosen depending on the level of the dendrogram cutoff. As an example consider Figure 5 at the first level there are two clusters $C_1 = \{2, 3\}$ and $C_2 = \{0, 1, 4, 5, 6, 7, 8, 9, 10, 11, 12\}$. At the second level there are three clusters $C_1 = \{2, 3\}$, $C_2 = \{1, 4, 5, 6\}$ and $C_3 = \{0, 7, 8, 9, 10, 11, 12\}$.

3.1 Agglomerative hierarchical clustering

As with most clustering methods there are two steps. The first step is to calculate a dissimilarity matrix. The dissimilarity matrix P (sometimes also called a proximity or a distance matrix) is n by n symmetric matrix, where n is the number of samples. The elements of the dissimilarity matrix contain the dissimilarities of the samples regarding each other, where P_{ij} the dissimilarity between sample i and j. There are more than a dozen methods that calculate the dissimilarity between samples several methods are described in Section 3.2.

The second step is to create the clustering based on the dissimilarity matrix and to illustrate the results in a dendrogram as seen in Figure 5. The clustering is done by a linkage method. There are several linkage methods and they are described in more detail in Section 3.3. The linkage method decides when a cluster is combined with another cluster and how the new dissimilarity between a newly formed cluster with the rest is calculated. There are several combinations posible between the method that calculates the dissimilarity matrix and the method that determines the linkage. There is however not one method that gives the best result [4]. Although there are some guidelines based on the type of data in the dataset, the final result for which method works best depends on the dataset itself. Therefore in Section 6 we test several of these methods on the eDNA dataset to determine which method gives the best result.

3.2 Dissimilarity coefficients

The dissimilarity matrix P is determined by using a dissimilarity coefficient, that calculates the dissimilarity between two samples, for every sample. Which results in a n by n matrix, where n is the number of samples. We have that the element P_{xy} is determined by the dissimilarity $d(x, y)$ between sample x and y. The resulting matrix is symmetric since we have $d(x, y) = d(y, x)$ and $d(x, x) = 1$.

There are a lot of coefficients that calculate the dissimilarity between samples. When the dissimilarity between samples is zero that means that the method cannot distinguish between the two samples so for the method these samples are equal. Some coefficients calculate the similarity instead. When the similarity between samples is zero that means that the samples do not share any characteristics according to the coefficient.

3.2.1 Symmetric vs asymmetric

Dissimilarity methods can be categorized into two categories namely symmetric and asymmetric. The difference between these two categories is how they deal with the double zero problem. The double zero problem is the problem on how to see the similarity between samples when they both contain a zero for that variable. Depending on the type of data in the dataset both samples having

a zero at a variable can mean different things. When the variables are measurements then the samples should be more similar when they are zero for that variable. However when a zero means an absence of a species it should not mean that the samples are more similar. An absence of a species can have multiple reasons, it may be because these locations corresponding to these samples have environmental conditions that are unsuitable for the species, and these conditions may be similar or very different for the two locations [5].

For symmetric coefficients the double zero is treated the same as as any other value that the samples have in common. So a double presence is treated the same way as a double presence. Asymmetric coefficients treat double presence differently then double absence. When a variable is zero for both samples the dissimilarity does not increase or decrease. When a variable is equal and nonzero for both samples the dissimilarity decreases.

3.2.2 Symmetric dissimilarity coefficients

Among symmetric methods Euclidean is the most commonly used. The dissimilarity between sample x and y is given as

$$d(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}. \tag{5}$$

When the values of both variables, i.e., x_i and y_i, are both zero then the sum in the square root does not increase. So the dissimilarity does not increase and does not decrease when the values are both are zero. When the values of the variables are equal, i.e., $x_i = y_i$. Then the sum in the square root also does not increase since $(x_i - y_i)^2 = 0$ when $x_i = y_i$. Therefore the dissimilarity does not increase and does not decrease. So the treatment is the same and therefore the coefficient is symmetric.

The correlation measures the dependence of the two samples and the function is

$$d(x, y) = 1 - \frac{(x - \bar{x})(y - \bar{y})}{\|x - \bar{x}\|_2 \|y - \bar{y}\|_2}, \tag{6}$$

where $\bar{x}$ is the mean of the elements of x.

3.2.3 Abundance

An example of asymmetric methods are methods that calculate the abundance of a variable (species). One such method is called the Bray-Curtis coefficient, which is sometimes also called the percentage difference. The Bray-Curtis coefficient is mostly used for ecological data. Instead of calculating the dissimilarity, the similarity between samples is calculated. The similarity $s(x, y)$ is the similarity between sample x and y. This similarity is given as

$$s(x, y) = \frac{2 \sum_{i=1}^{n} \min(x_i, y_i)}{\sum_{i=1}^{n} x_i + \sum_{i=1}^{n} y_i}. \tag{7}$$

The similarity can be transformed into a dissimilarity by using one of the following formula:

$$d(x, y) = 1 - s(x, y),$$

where d is the dissimilarity and s the similarity. Using this transformation for equation (7) we get

$$d(x,y) = \frac{\sum_{i=1}^{n} x_i + \sum_{i=1}^{n} y_i - 2\sum_{i=1}^{n} \min(x_i, y_i)}{\sum_{i=1}^{n} x_i + \sum_{i=1}^{n} y_i} = \frac{\sum_{i=1}^{n} |x_i - y_i|}{\sum_{i=1}^{n} |x_i + y_i|}. \tag{8}$$

When the values of both variables, i.e., x_i and y_i, are both zero then the sum on both the numerator and the denominator does not increase. So the dissimilarity does not increase and does not decrease when the value of both samples are zero. When the values of the variables are equal, i.e., $x_i = y_i$. Then the sum of the numerator does not increase, i.e., $|x_i - y_i| = 0$. However the sum of the denominator increases since $|x_i + y_i| > 0$ and the dissimilarity decreases.

3.2.4 Presence/absence

An group of methods that contain contain both symmetric and asymmetric methods are methods that calculate the presence and absence of the variables (species). There are more than a dozen similarity and dissimilarity coefficients and they all use the frequencies a, b, c and d, as seen in Figure 6.

		Object $\mathbf{x}_2$	
		1	0
Object $\mathbf{x}_1$	1	a	b
	0	c	d

Figure 6: Table of the frequencies a,b,c and d. Here a is given as the number of species the samples have in common, b the number of species that are only present in sample x, c the number of species that are only present in sample y and d the number of species that are not present in either sample.

When d is used to calculate the similarity the method is symmetric and when d is not used the method is asymmetric. The differences between these methods are mostly that they assign different weights to the species that are present in both samples and species that are present in only one of the samples being compared. Some examples of coefficients that use the presence/absence and are asymmetric are the following.

The Jaccard similarity is calculated as

$$s(x,y) = \frac{a}{a+b+c}. \tag{9}$$

For the dissimilarity we use $d(x,y) = 1 - s(x,y)$. We then get the proportion of the elements that disagree.

An example of a similarity coefficient that gives more weight to the variables that are present in both samples is the Sørensen–Dice coefficient. The similarity between sample x and y is given as

$$s(x,y) = \frac{2a}{2a+b+c}. \tag{10}$$

This similarity coefficient will not be used in the analysis as it is only given as an example.

3.3 Linkage methods

All linkage methods start with defining every sample as a cluster of one. The clusters that are the closest together are then combined. The function of a linkage methods uses the dissimilarity matrix, that is made using one of the methods in the previous section, to calculate the new dissimilarity between the clusters. The difference between the linkage methods is that they all use a different function in order to calculate the dissimilarity of the new clusters Some linkage methods are:

i. Single

ii. Complete

iii. Average
or Unweighted Pair Group Method with Arithmetic mean (UPGMA)

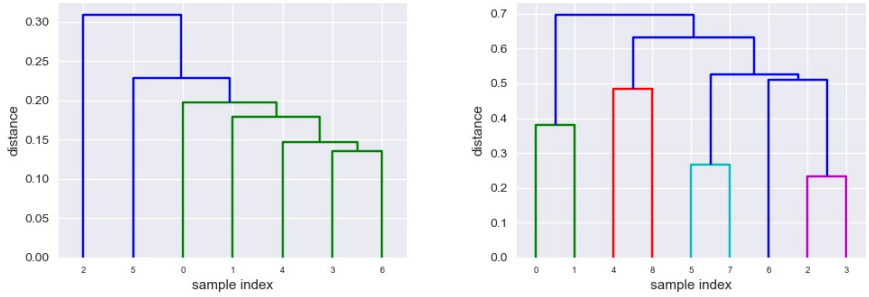
iv. Weighted
or Weighted Pair Group Method with Arithmetic Mean (WPGMA)

v. Centroid
or Unweighted Pair Group Method with Centroid Averaging (UPGMC)

vi. Median
or Weighted Pair Group Method with Centroid Averaging (WPGMC)

vii. Ward

An property of an linkage method as described by [5] is it's effect on space when clusters are formed. An linkage is called space contracting when the distance between clusters decrease as clusters are formed. The consequences of linkage methods that are space contracting is that clusters with one sample are more easily merged with clusters that have multiple samples. This is called chaining, an example can be seen in Figure 7a. The opposite of space contracting is space dilating. An linkage method is called space dilating when the distance between samples increases as clusters are formed. A consequence of space dilating methods is that sample specific clusters are more easily formed with other sample specific clusters. An example can be seen in Figure 7b.



(a) Chaining (b) Contracting

Figure 7: The effects of two linkage methods on space.

Let $D(C_i, C_j)$ be the dissimilarity between cluster C_i and C_j. When the dissimilarity (or distance) between the clusters is preserved after clusters are formed it is called space conserving. An linkage method is space conserving if

$$D(C_l, C_k) < D(C_l, C_j) < D(C_k, C_j), \tag{11}$$

holds for all clusters C_l, C_k and C_j, then the following also holds for all clusters C_l, C_k and C_j

$$D(C_l, z) < D(\{C_l, C_k\}, C_j) < D(C_k, C_j).$$

a linkage method is called space contracting. The procedure of a linkage method is as follows. All linkage methods start with n clusters, assigning each of the n samples to its own sample-specific cluster. Then the two cluster that have the lowest value in the dissimilarity matrix are combined in on cluster. The second step is generally the same for all linkage methods. The third step is to calculate the new reduced dissimilarity matrix between the clusters. This process is repeated until one cluster remains. The linkage method all use a different function in order to calculate the new reduced dissimilarity matrix. As an example consider the following sample specific clusters: $C_1 = \{x\}, C_2 = \{y\}, C_3 = \{z\}, C_4 = \{u\}$, where x, y, z and u are the samples. In the second step samples x and y are combined in one cluster, i.e., $C_5 = C_1 \cup C_2 = \{x, y\}$. Then a new reduced dissimilarity matrix is calculated that contains the dissimilarity between the new clusters, i.e., the dissimilarity $D(C_5, C_3)$ and $D(C_5, C_4)$. Using this new dissimilarity matrix the two clusters with the lowest dissimilarity are combined in a new cluster.

3.3.1 Single linkage

Single linkage (nearest neighbor) defines the distance between two clusters as the minimum distance between the elements in that cluster and can seen in Figure 8. The distance between two clusters C_i and C_j is defined as

$$D(C_i, C_j) = \min(D(C_k, C_j), D(C_l, C_j)), \tag{12}$$

where C_i was formed with cluster C_k and C_l, i.e., $C_i = C_k \cup C_l$. For single linkage we can also calculate the dissimilarity between every cluster based on

original dissimilarity matrix calculated in Section 3.2.

$$D(C_i, C_j) = \min_{x \in C_i, y \in C_j} (d(x, y)). \tag{13}$$

Single linkage is space contracting since the distance between clusters decreases. Therefore a disadvantage of single linkage is that chaining can occur. It can however be useful to use Single linkage to check if chaining occurs as it gives some information about the structure of the data, namely that intermediates are present. Single linkage can also be used to detect outliers in the data, which are the samples that are added to a cluster at a large distance. For example in Figure 7a sample 2 is an outlier.

3.3.2 Complete linkage

Complete linkage (furthest neighbor) defines the distance between two clusters as the maximum distance between the elements and can be seen in Figure 8. The distance between two clusters C_i and C_j is defined as

$$D(C_i, C_j) = \max(D(C_k, C_j), D(C_l, C_j)), \tag{14}$$

where C_i was formed with cluster C_k and C_l, i.e., $C_i = C_k \cup C_l$. For complete linkage we can calculate the dissimilarity between every cluster based on dissimilarity matrix calculated in Section 3.2.

$$D(C_i, C_j) = \max_{x \in C_i, y \in C_j} (D(x, y)). \tag{15}$$

As the algorithm progresses the clusters are moving further away from each other and the distance between clusters increases. It is therefore a space dilating method. This is the opposite of single linkage where the distance decreases. An disadvantage of complete linkage is that samples within a cluster can be more similar to samples in another clusters than to the samples in its own cluster [7].

3.3.3 Average linkage

Average linkage is a compromise of single and complete linkage. The distance between clusters is defined as the average distance between all samples from one cluster to the other cluster and can be seen in Figure 8. The distance between two clusters C_i and C_j is defined as

$$D(C_i, C_j) = \frac{n(C_k)D(C_k, C_j) + n(C_l)D(C_l, C_j)}{n(C_i)n(C_j)} \tag{16}$$

where C_i was formed with cluster C_k and C_l, i.e., $C_i = C_k \cup C_l$. In addition $n(C_i)$ is the cardinality of cluster C_i. The cardinality of a set is the number of elements in that set. So $n(C_i)$ is the number of samples in cluster C_i. For average linkage we can also calculate the dissimilarity between every cluster based on original dissimilarity matrix calculated in Section 3.2.

$$D(C_i, C_j) = \frac{1}{n(C_i)n(C_j)} \sum_{x \in C_i} \sum_{y \in C_j} D(x, y), \tag{17}$$

Average linkage gives equal weight to every member of a cluster. So a larger cluster has more influence in the distance after a merging when merged with a

smaller cluster. An consequence of this is that when a smaller cluster is merged with a larger cluster, where the samples of the larger cluster are similar to each other because of their common origin, the results may be distorted [3]. Therefore average linkage should only be used when the data is obtained by simple random or systematic sampling, i.e., the number of samples obtained per location should be equal.

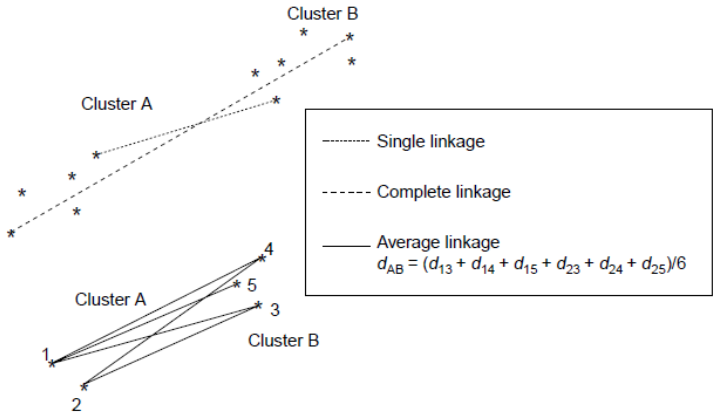


Figure 8: Examples of linkage methods: single, complete and average from [4].

3.3.4 Weighted linkage

Weighted linkage gives the clusters that are being merged an equal weight when determining the new distance. So a smaller cluster is weighted equally to a larger cluster. The consequence is that the samples in the smaller clusters are weighed more (which is where the term weighted comes from). The distance between two clusters C_i and C_j is defined as

$$D(C_i, C_j) = \frac{D(C_k, C_j) + D(C_l, C_j)}{2}, \tag{18}$$

where cluster C_i was formed with cluster C_k and C_l and C_j is a remaining cluster in the forest.

3.3.5 Ward linkage

Ward's method attempts to minimize the sum of the squared distances of points from their cluster centroids. The distance $D(C_i, C_j)$ is determined by how much the sum of squares increases when the cluster C_i and C_j are merged, i.e.,

$$D(C_i, C_j) = \sum_{x \in C_i \cup C_j} D(x, m_{C_i \cup C_j})^2 - \sum_{x \in C_i} D(x, m_{C_i})^2 - \sum_{x \in C_j} D(x, m_{C_j})^2, \tag{19}$$

where m_{C_i} is the centroid of cluster C_i. We can write the Ward linkage in the same form as the previous equations by writing the sum of squares in terms of pairwise distances [9], i.e.,

$$\sum_{x \in C_i} D(x, m_{C_i})^2 = \frac{1}{n(C_i)} \sum_{x,y \in C_i \times C_i} D(x, y)^2. \tag{20}$$

Using equation (20) we can rewrite equation (19), and we get

$$D(C_i, C_j) = \left\{ \frac{n(C_j) + n(C_k)}{n(C_i) + n(C_j)} D(C_j, C_k)^2 + \frac{n(C_j) + n(C_l)}{n(C_i) + n(C_j)} D(C_j, C_l)^2 - \right.$$

$$\left. \frac{n(C_j)}{n(C_i) + n(C_j)} D(C_k, C_l)^2 \right\}^{1/2}$$

where cluster C_i was formed with cluster C_k and C_l and C_j is a remaining cluster in the forest. In addition, $n(C_j)$ is the cardinality of cluster C_j.

3.3.6 Centroid

Centroid defines the distance between two clusters as the distance between the centroids, i.e., the mean point of the samples, of the clusters. As average linkage centroid gives equal weight to the samples in a cluster so that a larger cluster has more influence in the final distance then a smaller cluster when merged.

3.3.7 Median

As Centroid, Median defines the distance between two clusters as the distance between the centroids of the clusters. The difference with the centroid method is that the clusters are given equal weight when merged as in Weighted linkage.

Table 2: Overview linkage methods

Linkage Method	Effect on space	Additional properties
Single	contracting	Is susceptible to noise and outliers (can be used to identify outliers). Can lead to chaining.
Complete	dilating	Less susceptible to noise and outliers, but it can break large clusters [3].
Average	conserving	Should only be used when the data is obtained by simple random or systematic sampling [5]. Relatively robust [4].
Weighted	conserving	Preferred when the data is not obtained by systematic sampling [5].
Centroid	conserving	Works best with Euclidean distance. Preferred when the data is obtained by systematic sampling [5].
Median	conserving	Works best with Euclidean distance. Preferred when the data is not obtained by systematic sampling [5].
Ward	conserving	Works best with Euclidean distance. Tends to produce clusters of the same size [4].

4 Principal component analysis

In this section we describe how principal component analysis (PCA) works and how the result of the analysis can be interpreted when applying PCA to the eDNA dataset. We want to use principal component analysis to visualize the results of the clustering analysis (Bray-curtis average linkage). When the clusters are still visible in the biplots of the PCA this gives some more confidence that the clustering is correct. In addition we want to see if there are species or OTUs that are characteristic for some clusters.

Principal component analysis is a linear technique that finds new variables (principal components) which are linear combinations of the original variables. These principal components have the properties that they are orthogonal and capture the maximum amount of variation in the data. The first principal component captures as much variation as possible. The second principal components captures the remaining variation in that direction with the constraint of being orthogonal to the first. The third principal component is orthogonal to the second and captures the remaining variation in that direction and the first and so on.

PCA was created as a dimensional reduction technique as the variation can be captured by a small fraction of the original dimensions, but it has several other applications. Namely finding patterns and outliers in the data in addition to reducing the noise of the data. In Figure 9 it is shown how the 3 dimensional data is reduced to two dimensional data.

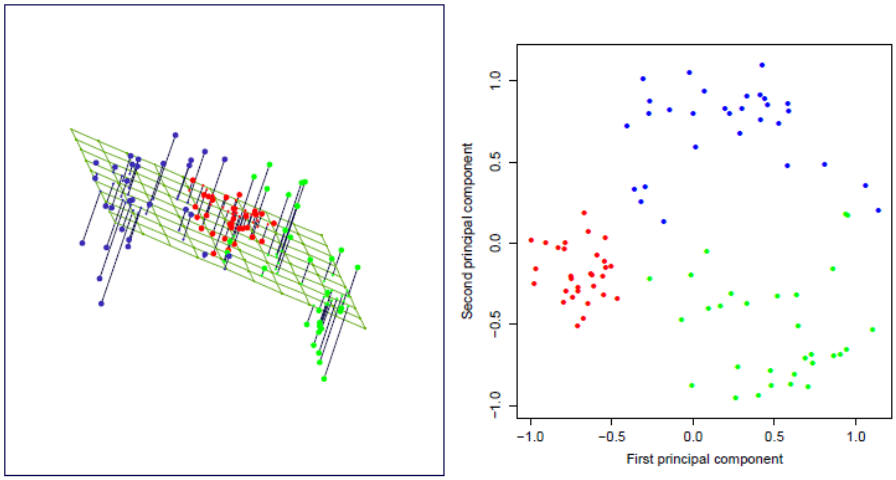


Figure 9: Going from 3D data to 2D using PCA from [7].

4.1 How PCA works

Principal component analysis works as follows. Let the data consists of n samples with m variables (the species). First the dataset needs to be standardized with respect to the species. By standardization we meant that the mean of the variables needs to be equal to zero and the variance equal to one. This is done

because the units or scale of the variables may be different and if the data is not standardized then the PCA will give emphasis on the variables that have the biggest range (and therefore variation).

For every variable in the dataset the covariance is calculated. The covariance is the measure of the joint variability between two variables. Let x_i be a variable with $0 < i \leq m$, then the covariance of every x_i with every x_j is calculated and given in the covariance matrix $C \in \mathbb{R}^{n \times n}$ as follows.

$$\{c_{ij}\} = cov(x_i, x_j) = \sum_{k=1}^{m} \frac{(x_{i_k} - \bar{x}_i)(x_{j_k} - \bar{x}_j)}{n-1},$$

where $\bar{x}_i$ is the mean of the elements of x_i. Then, the eigenvectors are calculated of the covariance matrix C and are sorted by the eigenvalues from highest to lowest eigenvalue. The eigenvectors with the highest eigenvalue explain the most of the variance in the data. The principal component y_i is the linear combination of the original variables and determined as follows.

$$y_i = a_{i1}x_1 + ... + a_{in}x_n, \tag{21}$$

where a_{ij} is the j component in the eigenvector a_i and $x_i, 0 < i \leq n$ are the original variables (after standardization). So the eigenvectors of the convariance matrix determine the weights in the linear transformation. The eigenvectors can therefore be used to interpret the principal components. If some elements a_{ij} are large in an eigenvector a_i for a principal component y_i then those elements are important in defining that principal component. In addition if the correlation is positive and of the same magnitude then the elements in the eigenvector are positive and of the same magnitude.

There are now $\min(n, m) - 1$ principal components and the number of dimension in the dataset can now be reduced to p dimensions by choosing the first p principal components. How much variation in the data is kept is given by the explained variance. The explained variance is calculated using the eigenvalues. Let λ_+ be the sum over all of the eigenvalues. Then the explained variance V_i of the principal component i is given by

$$V_i = \frac{\lambda_i}{\lambda_+}.$$

If the sum of the explained variance of the first p principal components is for example 0.95 that means that 95% of the variance in the dataset can be explained by these principal components. When only the first two or three principal components are chosen then the data can be plotted in a two or three dimensional plot so that the data can be more easily visualized. More information about how PCA works and its uses can be found in [1] and [2].

4.2 Rotation

When the scores of the principal components have intermediate as well as large and small ones. Or when the scores of the principal components are all small it can be can be difficult to interpret. If this happens for most of the principal

components rotation may be a solution. One of the most used rotations is the VARIMAX rotation which discussed and used in this report. The VARIMAX rotation maximizes the sum of the variance of the squared loadings. The loadings are the eigenvectors multiplied with the squared root of the eigenvectors, i.e.,

$$loadings = eigenvectors \times \sqrt{eigenvalues}.$$

The goal of rotation is that each variable (species) is associated to one principal component, i.e., the species have a high score for only one principal component and low scores at the other principal components. An property of the VARIMAX rotation is that the rotation is orthogonal and therefore preserves an essential property of the PCA. In addition the explained variance of the principal components also remains equal.

5 Software

The programming language Python was used for the cluster analysis and the principal component analysis. Python was chosen because it has several packages which make data analysis easier. In addition Witteveen + Bos has written some functions in Python scripts which were used as a basis. The data analysis can also be done in other software languages. Some software languages have libraries that contain functions for cluster analysis and PCA which makes implementation easier. Examples are Julia, R and Matlab.

In general the following packages were used:

(i) Pandas
The package pandas was used to read and manipulate the eDNA dataset. The eDNA data that was used was contained in excel files.

(ii) NumPy
NumPy was used for several mathematical functions.

(iii) Matplotlib.pyplot
All of the plots for the data analysis were made using the package Matplotlib.pyplot.

In addition some functions from Witteveen + Bos were used and adapted. Namely, several function to prepare the data for analysis and delete samples that contained less than 20.000 reeds. There were also some new functions written. Among these function was a function that transformed the data according the one of the transformations given in Section 2.3.

5.1 Cluster analysis

For the cluster analysis some additional packages were used. Namely, scipy.spatial.distance in order to calculate the dissimilarity matrix. For the linkage scipy.cluster.hierarchy was used. In addition some functions were used and adapted from Witteveen + Bos to preform the final cluster analysis. In order to compare the cluster results of the methods a couple of new functions were written. This includes the function that compares two dendrograms with a plot as given in Figure 10, and some functions that compare the results of the methods as given in Section 6.2.

5.2 Principal component analysis

For the principal component analysis the package sklearn was used. Which contained functions that were used to preform the PCA. Namely, sklearn.decomposition.pca and sklearn.preprocessing.StandardScaler. From Witteveen + Bos functions were adapted that perform the PCA for the eDNA dataset with and without the Varimax rotation.

6 Results cluster analysis

In this section the results of the cluster analysis for the eDNA dataset are given. Their is no best recommended method for hierarchical clustering, and which method preforms better than another method often depends on the data. Therefore, we will test several methods to see which method provides the most robust clustering. The dissimilarity coefficients Euclidean, Correlation, Bray-Curtis and Jaccard will be used in conuction with the linkage methods Single, Complete, Average, Weighted and Ward.

The eDNA dataset contains a lot species that are rare (around 2081 OTUs that occur only in 5 samples of the total of 53 samples). Therefore the dataset contains a lot of zeros and in this case when two variables (species) are zero for both samples it does not mean that the samples are more similar. Bray-Curtis and Jaccard both do not look at the OTUs that don't occur in both samples and are therefore useful for this kind of data set [5]. Correlation was not recommended in [5] for comparison between samples, we do use it as as a comparison because this method was used in a previous cluster analysis for the eDNA dataset and therefore is used as a comparison to see if the other method preform better or worse.

6.1 Cluster validation

A cluster algorithm will provide clusters of the data even when there are no clusters in the data [4]. One method to validate a clustering result of a Hierarchical clustering is the cophenetic correlation. The cophenetic correlation measures the correlation between the original dissimilarity matrix, as calculated in step one, with the cophenetic distance matrix. The cophenetic distance between two samples is the distance in the dendrogram when they are clustered together. Consider the example of an dendrogram in Figure 5, the cophenetic distance between sample 6 and 7 is around 0.8, since that is the distance when they are first together in a cluster. When the cophenetic correlation is close to one this means that the dissimilarity between samples is preserved after clustering. When this is not the case it means samples in the same cluster are not necessarily more dissimilar than samples in a different cluster.

For the eDNA dataset another two tests are performed. When more samples are added to the eDNA dataset it will result in more OTUs and most of the increase in OTUs will be OTUs that occur in only a few samples. When these samples are added we do not want an entirely different result from the cluster analysis. Therefore it will be checked whether or not the cluster analysis is the same or similar when is is performed again using only OTUs that occur in at least x_s samples. Another test is to perform the cluster analysis using a subset of the samples of the eDNA dataset and check whether or not the samples still cluster together the same.

6.2 eDNA clustering results

The dissimilarity methods Bray-Curtis, Jaccard and Correlation were used with the linkage methods: Single, Complete, Average, Weighted and Ward. For Cor-

relation and Jaccard the qPCR data will have the same result as the ratio. Bray-Curtis and Euclidean will have a different results, will be tested with both. For the qPCR the data was first transformed using the function $log(x+1)$, as recommended by [3]. The one was added so that the variables that are zero in the original data are again zero.

Without the log transform Bray-Curtis will partially cluster samples together that have a similar amount of total biomass (per sample over all variables). For example the sample WF.SNM.1823 with RW.VOL.1823 that have a $1.848 * 10^8$ and $1.969 * 10^8$ respectively. The total biomass of all the samples is between $8.235 * 10^7$ and $1.099 * 10^9$. On the other hand the cluster DL.OMP.1823 and SK.KRP.1823 will not form an immediate cluster as the total biomass differs a lot ($1.099 * 10^9$ and $1.111 * 10^8$ respectively). However these samples will form a cluster when the log transformation is used.

Single linkage resulted in chaining for every dissimilarity coefficient. We can therefore conclude that the data contains intermediate points and single linkage is not a suitable method for finding clusters. There were however two outlier detected using single linkage. The sample WD.PWT.1823 was considered an outlier for all distance methods. In addition the sample RW.VOL.1819 was considered an outlier for both Bray-Curtis and Jaccard.

6.2.1 Test: Cophenetic correlation

The cophenetic correlation for the dissimilarity methods can be seen in Table 3. For all dissimilarity methods the Average linkage method gives the high cophenetic correlation. The highest cophenetic correlation is when using the Euclidean distance with average linkage and using the ratio data.

Table 3: Cophenetic correlation for all tested methods.

	Single	Complete	Average	Weighted	Ward
Bray-Curtis (qPCR)	0.7	0.6	0.76	0.63	0.55
Bary-Curtis (ratio)	0.69	0.64	0.78	0.77	0.44
Jaccard	0.73	0.71	0.84	0.67	0.5
Correlation	0.65	0.73	0.79	0.78	0.49
Euclidean (qPCR)	0.85	0.73	0.87	0.83	0.53
Euclidean (ratio)	0.93	0.93	0.96	0.94	0.71

6.2.2 OTU occurs in at least x_s samples

For this test the cluster result which was made using all the OTUs was compared with the cluster result which was made using OTUs that occur in at least x_s samples (with $1 \leq x_s \leq 15$). The amount of OTUs deleted by this action for some values of x_s can be found in Table 4.

Table 4: Number of OTUs in dataset after deleting species that occur in less then x_s samples.

	Number of OTUs kept	% OTU kept	Number of OTUs deleted
x_s = 0	2720	100	0
x_s = 5	639	24	2081
x_s = 10	253	13	2367
x_s = 15	238	9	2482

The comparison of the dendrograms of Bray-Curtis using average linkage with and without the filter x_s = 5 can be seen in Figure 10.

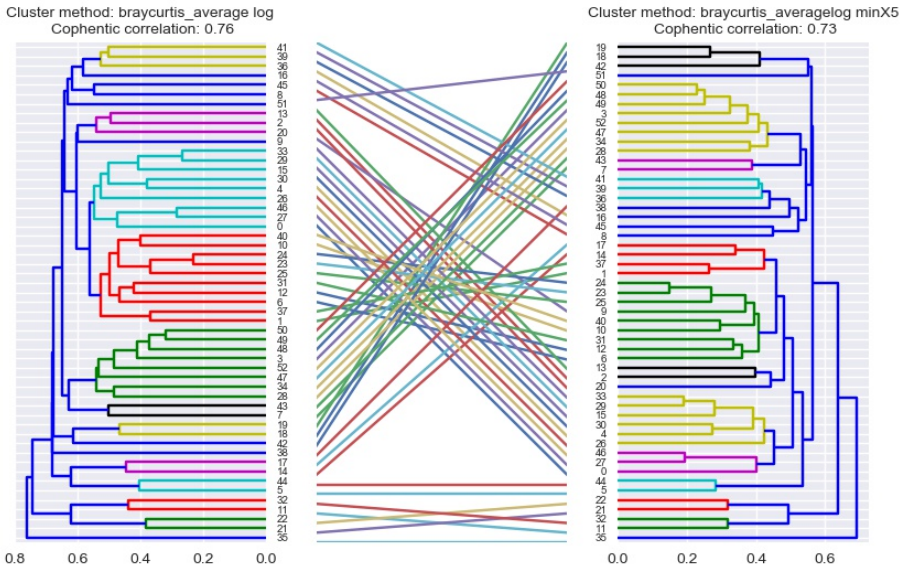


Figure 10: Cluster results of Bray-Curtis with average linkage compared with the results using the filter x_s=5.

In addition the correlation coefficients between the cophenetic distance matrix of the original dendrogram (without filter x_s) and the cophenetic distance matrix of the new dendrogram (with filter x_s) were calculated. This was done for values of x_s between 1 and 16. The results for all the dissimilarity coefficients with average linkage for the ratio data is given in Figure 11 and for the qPCR data in Figure 12. The plots for all of the linkage methods are given in Appendix A. Jaccard and Bray-Curtis using qPCR with the log transformation in conjunction with average linkage provides the most consistent results. For the Euclidean and Correlation dissimilarity the correlation coefficient decreases rapidly when x_s increases.

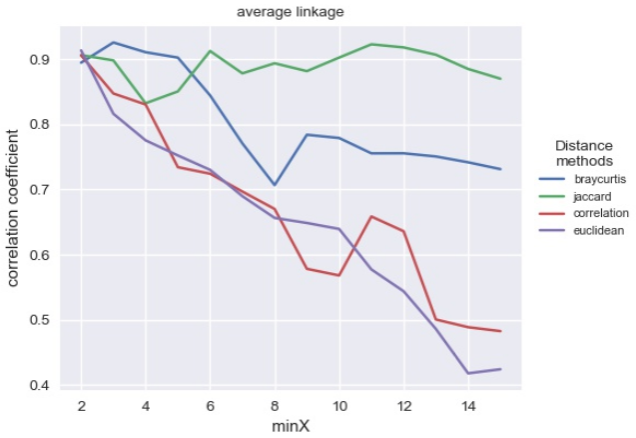


Figure 11: The correlations coefficients of the cophenetic distance matrix of the original clustering with the clustering result using the filter x_s. The data from the ratio was used.

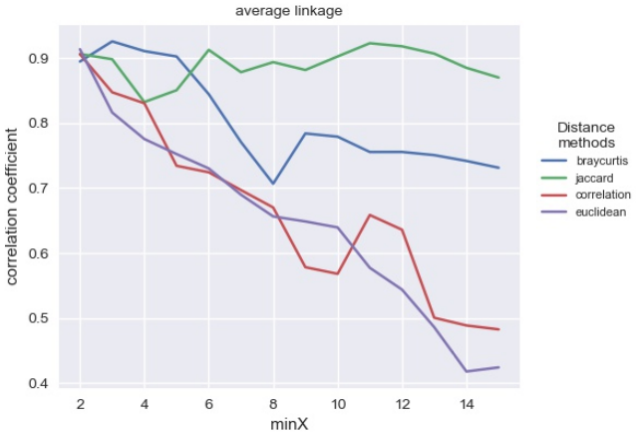


Figure 12: The correlations coefficients of the cophenetic distance matrix of the original clustering with the clustering result using the filter x_s. The data from the qPCR was used together with a Log transformation.

6.2.3 Test: Subset of samples

In this case the results of the cluster analysis where all 53 samples are used are compared with the cluster analysis where a subset of the samples are used (samples measured in week 23 and week 25). All distance coefficients in conjunction with average linkage provided good results in the sense that the distance between samples using all samples was similar to the distance between samples using only a subset of the samples. The result of the clustering of Bray-Curtis using average linkage and with qPCR data and Log transformation is given in Figure 13.

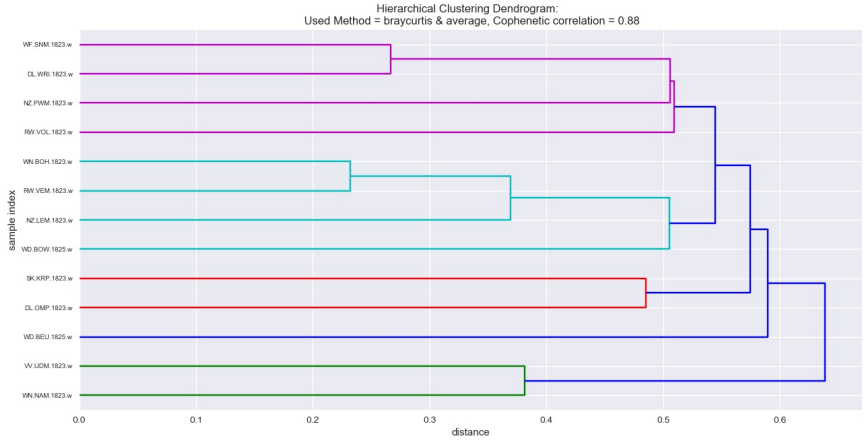


Figure 13: Bray-Curtis using average linkage and with qPCR data and Log transform. Only samples taken in week 23 and 25 of the year 2018 were used.

6.2.4 Comparison cluster results

For a final test of the clustering results of the methods we look at which samples form a clusters. From a ecological point of view we expect that some samples will form a clusters. For example samples that are taken from the same location but at a different data. The method Bray-Curtis with average linkage using qPCR with the Log transform provided the best clustering, the resulting dendrogram can be seen in Figure 14. The results of the other clustering method can be found in Appendix B.

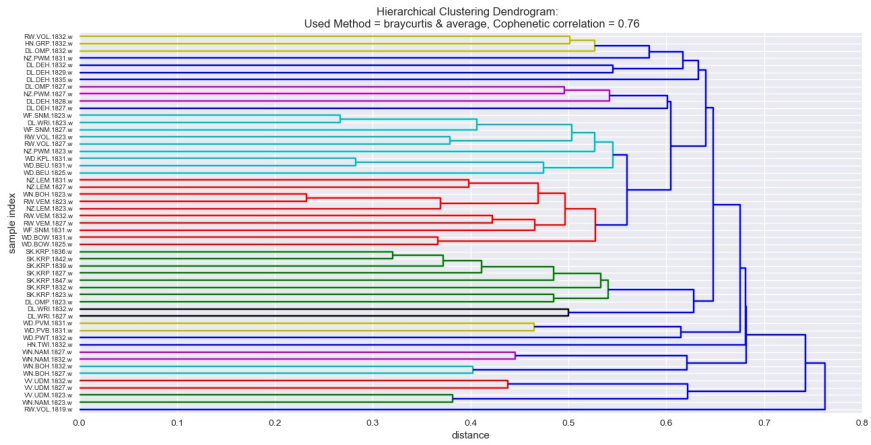


Figure 14: Bray-Curtis using qPCR data and average linkage, with all samples.

6.3 Discussion results

Using average linkage all dissimilarity coefficient had a high cophenetic correlation. The highest cophenetic correlation was with the Euclidean distance using the ratio data. With the filter of only using OTUs that occur in at least x_s samples the dissimilarity coefficients Jaccard and Bray-Curtis using qPCR and Log transform with average linkage provided the best results. For the test using only a subset of the samples all dissimilarity methods scored well with average linkage. Overall Bray-Curtis with average linkage using qPCR with the log transformation provided the most consistent results.

7 Results PCA

Before we can use the data from the Universal primer of the eDNA dataset we use a transformation as described in Section 2.3. This is because PCA detects linear relationships and the species tend to have exponential growth [3]. The Hellinger and the Log transformation was used. The Hellinger transformation was used on the dataset containing the number of reeds of an OTU per sample. The Log transformation was used on the qPCR data as described in Section 2.4. The Hellinger transformation resulted in a slightly higher explained variance over the first two principal components. Additionally the Hellinger transformation was recommended by [8]. Therefore for further analysis the Hellinger transformation will be used.

7.1 Options data filtering

As mentioned before the eDNA dataset contains a lot of rare species. Namely, 2081 of 2720 species only occur in a few samples. When a PCA is performed a lot of the variation will be explained using the rare species. There are several options for this problem. The first is to group the data by taxonomic level. By this we mean that the data is grouped by the species name of that level. As an example consider the taxonomic level of superkingdom. This level consists of Archaea, Eukaryota, Bacteria and unknown OTUs in this dataset. The OTUs that fall under Bacteria are grouped together in a single row and the ratio or biomass within a sample is added together. When a species is unknown at that particular taxonomic level it will be grouped together with the species that are also unknown. This unknown group will be deleted as it consists of species that may have nothing in common. The amount of species that are unknown at a taxonomic level and the dimensions the data is reduced to for every taxonomic level is given in Table 1.

Another reason to group by taxonomic level is that we are also interested in how the groups of species at a taxonomic level interact with each other. The disadvantage of this grouping is that it relies a lot on the reference database. Some species are unknown in this dataset or listed wrong which influences the result. For the analysis of the PCA we will only consider the following taxonomic levels: Phylum, Class , Order and Genus. As for the other taxonomic levels to much of the species are unknown at this level. Therefore a lot of the data is lost. When this changes in the future the other taxonomic levels can also be considered.

As we also want to look at the level of OTU we need an additional filter for the data. One of these filters is that species that occur in less than x_s samples will be dropped. This filter was also used for the cluster analysis in Section 3. The results of this filter for several values of x_s will be shown in Section 7.3.

Another option is to use only the top x_t species that have the highest sum over the samples (after the transformation). This filter will also be used for the grouping by taxonomic level as these also still contain a large number of dimension. The results of this filter for several values of x_t will be shown in Section 7.3.

7.2 Analysis PCA figure

The results of the PCA can be shown in a plot. The first two principal components are combined in a 2D plot, the third and fourth are also in a 2D plot and so on. However as mentioned before for the analysis we will only use the results of the taxonomic levels: Phylum, Class , Order and Genus. In the plot the samples are plotted as points and the ten species with the highest principal components are plotted as vectors of the two principal components. The number ten is chosen arbitrarily. For more than ten species the plots will be less clear when looking at the vectors of the species. In this report the color of a sample depends on which cluster it is in based on the cluster analysis using Bray-Curtis with average linkage. However the colors can also be determined by the results of the other method of the cluster analysis, by the date of when the sample was taken or the location of the sample. In Figure 15 an example of a plot of the PCA is given. In this case only the species that are known on Phylum level is used.

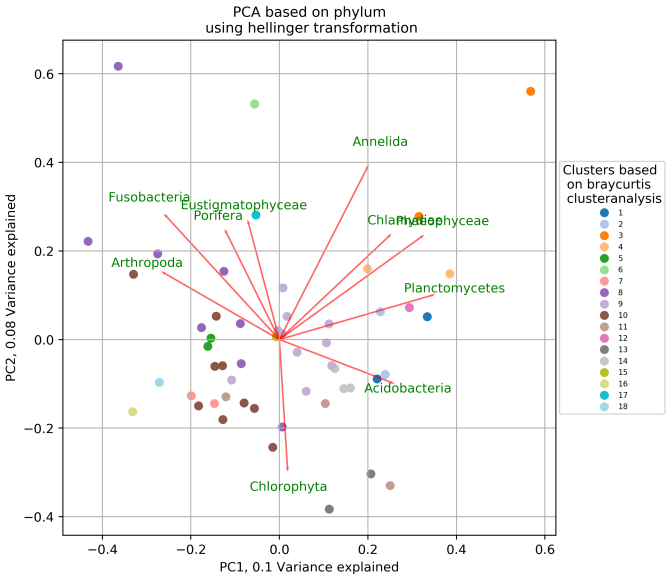


Figure 15: Plot of the samples using the first two principal components at Phylum level. For this plot no filter was used.

The vectors of the species that are plotted can also be seen as an axis in the sense that when a sample is close to that vector the sample contains a lot of that species in comparison to other samples.

7.3 Application to dataset

First a PCA was done for the taxonomic levels: Phylum, Class, Order, genus and additionally at OTU level. With the exception of Phylum and Class the samples are mostly clustered together especially at OTU level as seen in Figure 16. The

reason for this is that the first two principal components only explain around 11% of the variance. Additionally most of this variance is explained by a species that only occurs in a few samples (in this case the blue and green samples). An overview of the explained variance for the taxonomic levels without using an additional filter is given in Table 5.

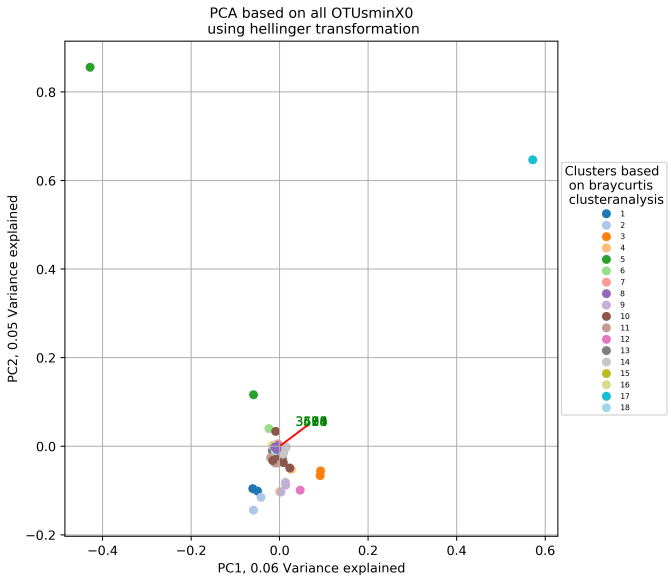


Figure 16: Plot of the samples using the first two principal components using data at OTU level. The Hellinger transformation was used.

Table 5: Explained variance per taxonomic level when using no additional filters.

Taxonomic level	Number of principal components needed to explain around 75% of the variance	sum of the explained variance for the first 10 principal components
Phylum	14 for 76%	63%
Class	21 for 76%	53%
Order	22 for 76%	49%
Genus	24 for 76%	45%
OTU	24 for 75%	40%

7.3.1 OTU occur in at least x_s samples

At OTU level using the filter of using only OTUs that occur in $x_s = 5$ samples increases the explained variance of the first two principal components. Additionally the samples are more spread out as can be seen in Figure 17.

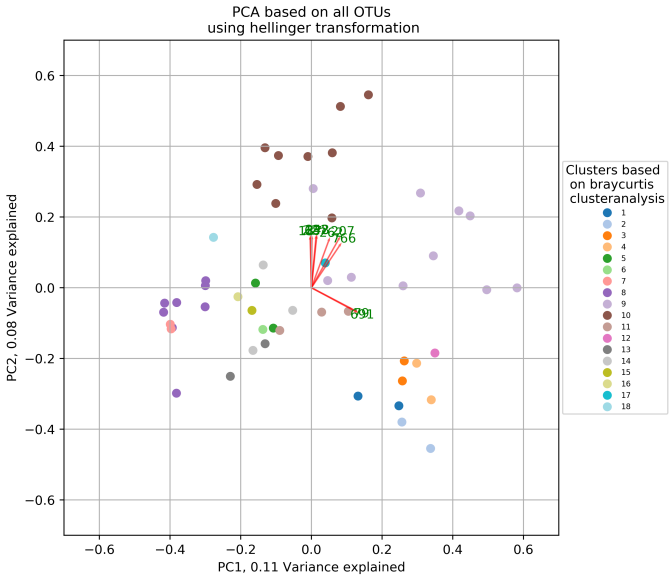


Figure 17: Plot of the samples using the first two principal components at OTU level. The Hellinger transformation was used. With the filter of using only species that occur in at least $x_s = 5$ samples.

7.3.2 Filter using top x_t species

The other filter over the data was using only the top x_t species over all samples. The explained variance using only the top x_t species for $x_t = 10, 20, 40$ is given in Table 6.

Table 6: Sum of the explained variance for the first 10 principal components with several values of x_t.

Taxonomic level	s $x_t = 10$	$x_t = 20$	$x_t = 40$
Phylum	100%	88%	78%
Class	100%	83%	64%
Order	100%	85%	71%
Genus	100%	84%	74%
OTU	100%	85%	73%

When performing a PCA with only the top x_t species the explained variance is much higher. This can also be seen in Figures 18,30 and 31 as the samples are more spread out. Do note however that is done on a subset of the data. Using filter for the top x_t species for $x_t = 20$, $x_t = 40$ clusters stay mostly together on the first 6 principal components, for the principal components 7,8,9 and 10 the clusters are more mixed, although these principal components do not explain a lot of the variance so that is to be expected.

In addition we wanted to see if there are species which are characteristic for a cluster. This is the case for some species namely for the taxonomic level class we have that cluster 8 stays around insecta for x_t = 10,x_t = 20 and x_t = 40. In addition cluster 10 is split around spirotrichea and cyptophagia in the plot of the first two principal components. This can also be seen in Figure 18. For the taxonomic level genus we see that cluster 9 is around thermostilla for x_t = 10,x_t = 20 and x_t = 40. This can also be seen in Figure 30. For phylum we have that cluster 11 and 14 are around cyanobacteria for x_t = 10 and x_t = 20 which can be seen in Figure 31.

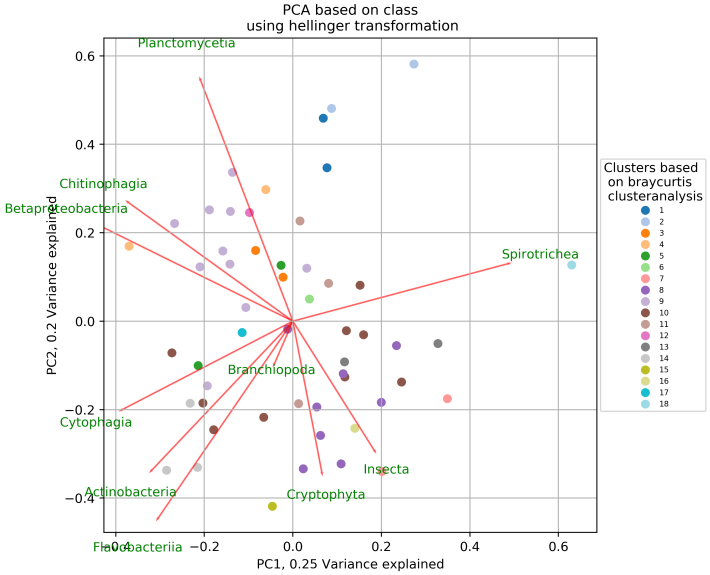


Figure 18: Plot of the samples using the first two principal components using data for Class level with x_t=40.

7.4 PCA scores

Another thing that can be analyzed are the scores of the principal components. Which can be used to interpret a principal component. When the coefficients for a principal component are all the same size that principal component can be seen as the average of all the variables (species). It can also occur that only a few are large and the rest are small, the principal component can be easily interpreted in that case, namely that the variation in that component is mostly determined by those variables (species). When one species scores high on a principal component for example 0.9 and another species around -0.8 that means that these two species are negatively correlated, i.e., samples that one species do not contain the other species. The scores of the species with the highest values for the principal components for Phylum level without any filters is given in Table 8.

Table 7: PCA scores of three species and the sum of the explained variance for the first 10 principal components for Phylum level.

principal component	explained variance	Streptophyta	Bacillariophyta	Colponemidia
1	0.19	-0.03	-0.03	0.02
2	0.11	0	0.01	-0.04
3	0.09	-0.02	-0.05	-0.01
4	0.09	-0.01	0.09	0.03
5	0.07	-0.02	0.26	-0.07
6	0.06	-0.02	0.01	0.13
7	0.05	0	0.08	-0.36
8	0.03	-0.08	-0.05	0.02
9	0.03	-0.09	-0.23	0.09
10	0.03	-0.07	-0.01	-0.54

The PCA scores for the other taxonomic levels, with and without filters, are also difficult to interpret as the scores also contain only low values that are between -0.3 and 0.3.

7.5 Results Rotation

We perform a rotation on the data with the filter that only the top x_t species are used. Since in that case the explained variance of the principal components is high and the goal of the rotation is to find important species that explain a lot of the variation. For x_t = 40 we have that 70% of the variance is explained by the first 10 principal components so a good option would be to rotate to 10. For x_t = 20 we have that 70% of the variance of the data is explained by the first 7 principal components so a good option would be to rotate to 7.

The plots of the data using the VARIMAX rotation are not as conclusive as the previous plots, i.e., there are no particular species that are characteristic for a cluster. The plots for taxonomic level Class and for OTU are given in Figure 19 and Figure 20 respectively.

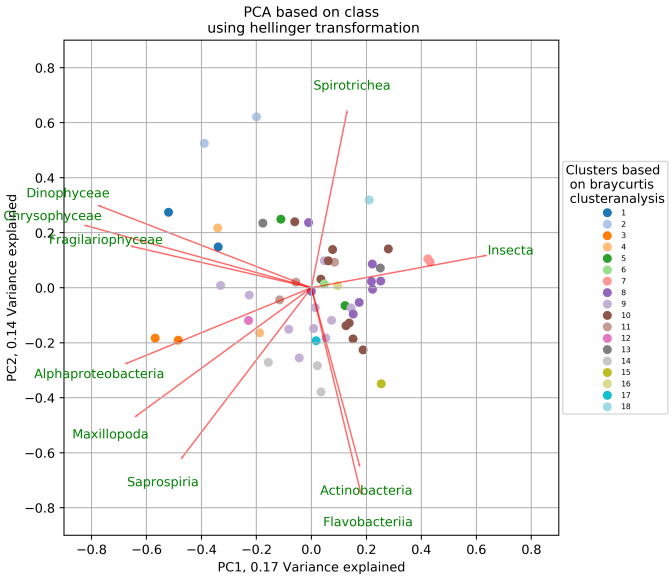


Figure 19: Plot of the samples using the first two rotated principal components. The Hellinger transformation was used.

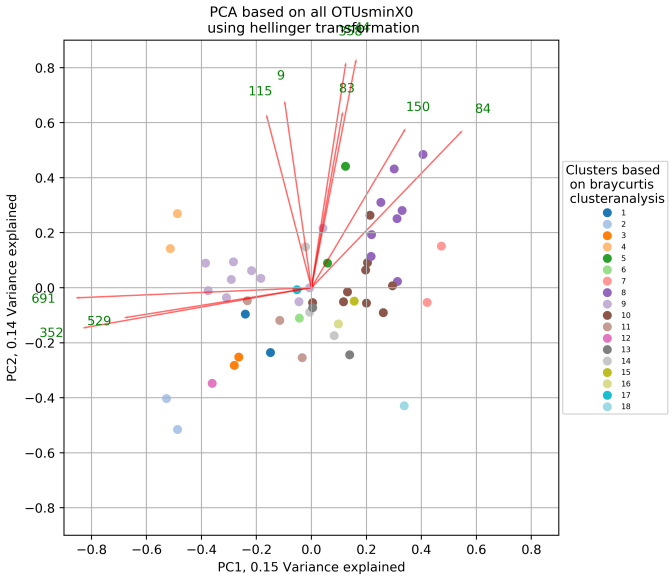


Figure 20: Plot of the samples using the first two rotated principal components.The Hellinger transformation was used.

The scores of the principal components are easier to interpret as there are

some high values and some low values for a species on a principal component. Since most species have a high value on one principal component and low values on the rest of the principal components. The first principal component is mostly determined by Arthropoda, Planctomycetes, Acidobacteria and Phaeophyceae. In addition the second principal component is determined by the Proteobacteria and the Cyanobacteria.

Table 8: PCA scores of three species and the sum of the explained variance for the 7 rotated principal components for Phylum level.

Principal component	explained variance	Proteobacteria	Bacteroidetes	Actinobacteria	Arthropoda
1	0.2	-0.25	0.02	0.25	0.75
2	0.15	0.76	0.45	0.37	0.13
3	0.12	0.08	0.21	-0.15	-0.04
4	0.13	0.08	-0.49	-0.01	0.18
5	0.14	-0.08	-0.28	0.12	0.15
6	0.12	-0.18	-0.02	-0.68	0.1
7	0.13	-0.01	0.25	0.09	0.13

7.6 Conclusion

For the PCA we can conclude that cluster stay mostly together. The filters of using the top 20 and top 40 species (x_t = 20 and x_t = 40) gives results that are the most similar to using no filter. In addition there are some species which are characteristic for a cluster it is not yet known however if these species have any ecological importance. However it is difficult to interpret the principal components as even after using the filter of x_t the PCA scores are between -0.3 and 0.3. therefore it is difficult to interpret the principal components.

The VARIMAX rotation was used in order to better interpret the principal components. For the rotation we found that rotating does help to interpret the principal components. Most species have a high value on one principal component and low values on the rest of the principal components. However it is not yet known if these species have any ecological importance. In addition after rotation the clusters are less clear in the plots for most of the taxonomic levels.

8 Conclusion/Discussion

We used the unsupervised learning methods: cluster analysis and principal component analysis, to analyze the eDNA dataset. This was done to check whether or not the eDNA profiles of the water samples could be used to categorize the dataset. Namely, when this is the case this gives more confidence that the eDNA profiles can be used to profile an water sample.

For cluster analysis the result was the samples that were taken from the same location but at a different data would be in a cluster together. This gives some confidence that the clustering result is not a random clustering. Several methods for cluster analysis were described and tested to see if they gave consistent results when using a subset of the data. Overall Bray-Curtis using qPCR with the log transformation and average linkage provided the most consistent results. In the plots from the principal component analysis the biggest three clusters were also clustered together for most of the principal components. As this was done on several taxonomic levels it gives shows that the clustering is not random.

In addition we wanted to use principal component analysis to analyze the data to see if there are species (on any taxonomic level) that are characteristic to some clusters. There was an additional filter on the data so that the dimensions of the dataset were reduced. The clusters were still visible for these subsets. In addition it was found that there are some species which are characteristic for a cluster it is not yet known however if these species have any ecological importance. It was difficult to interpret the principal components of the PCA because the PCA scores were low for all the species. As an option we used rotation to better interpret the principal components. We found that did help to interpret the principal components as most species have a high value on one principal component and low values on the rest of the principal components. However the clusters were less clear in the biplots for most of the taxonomic levels for the rotation.

For future analysis it would be recommended to also look into supervised learning methods when there is more data available.

References

[1] Jollife, I.T., *Principal Component Analysis.* Springer Verslag, 2nd edition, (2002).

[2] Dunteman,G.H., *Principal Component Analysis.* Sage Publications (1989).

[3] Tan, P., Steinbach, M., Kumar, V. *Introduction to Data Mining.* Addison-Wesley: Boston, Mass, 2nd edition, (2013).

[4] Everitt, B. *Cluster Analysis.* Wiley, 5.th ed., (2011).

[5] Legendre, L., Legendre, P., Rohlf, F. J., Belanger Robert. *Numerical Ecology.* Elsevier Scientific, 3.rd edition, (2012).

[6] Borcard, D.; Gillet Francoi,; Legendre, P. *Numerical Ecology with R.* Springer, 2nd edition, (2018).

[7] Hastie, T., Tibshirani, R., Friedman, J. H. . *The Elements of Statistical Learning : Data Mining, Inference, and Prediction.* Springer Series in Statistics, 2nd edition (2009).

[8] Legendre, P.,Gallagher. E. D. *Ecologically meaningful transformations for ordination of species data* Oecologia, Volume 129 (2001), no. 2, pp. 271–280.

[9] Murtagh, F., Legendre, P. *Ward's Hierarchical Agglomerative Clustering Method: Which Algorithms Implement Ward's Criterion* Journal of Classification, Volume 31 (2014), no. 3, pp. 274-295.

[10] Stowa. (2017). *Toepassing eDNA voedselwebanalyses voor toestandsbepaling en systeembegrip.* Retrieved from `https://www.stowa.nl/onderwerpen/waterkwaliteit/realiseren-van-ecologische-waterkwaliteitsdoelen-krw/toepassing-edna`

[11] Khan Academy. (2016). *Polymerase chain reaction (PCR).* Retrieved from `https://www.khanacademy.org/science/biology/biotech-dna-technology/dna-sequencing-pcr-electrophoresis/a/polymerase-chain-reaction-pcr`

Appendix A: Figures correlation coefficients

The correlation coefficients between the cophentic distance matrices of the original dendrogram (without filter x_s) and of the new dendrogram (with filter x_s). The results for all the dissimilarity coefficients with average linkage for the ratio data is given in Figure 21 and for the qPCR data in Figure 22.

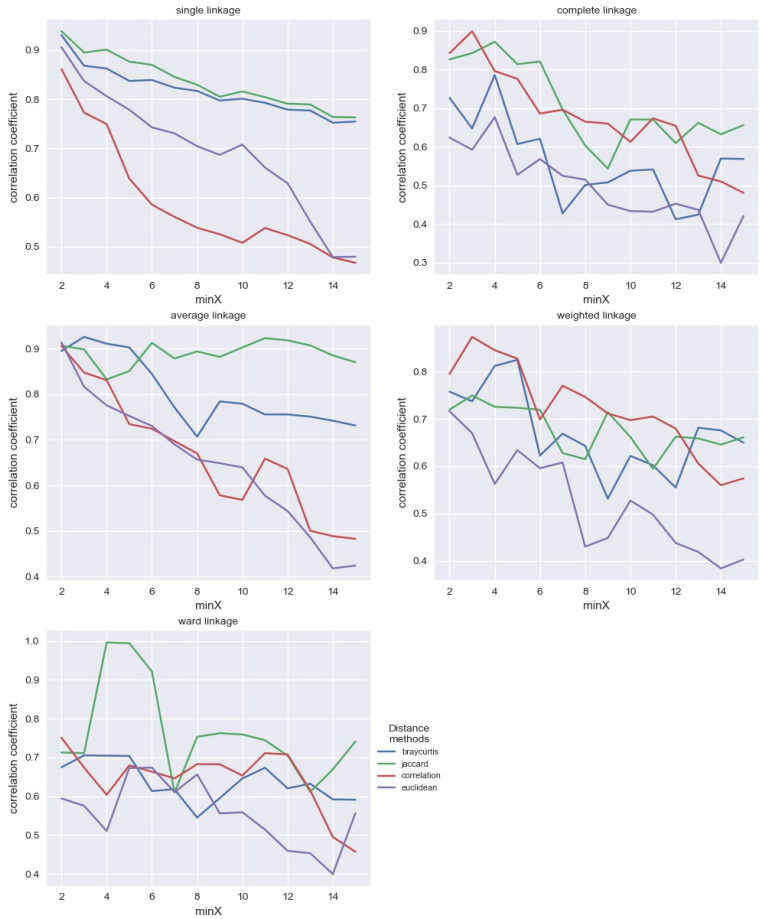


Figure 21: The correlations coefficients of the cophentic distance matrix of the original clustering with the clustering result using the filter x_s. The data from the ratio was used.

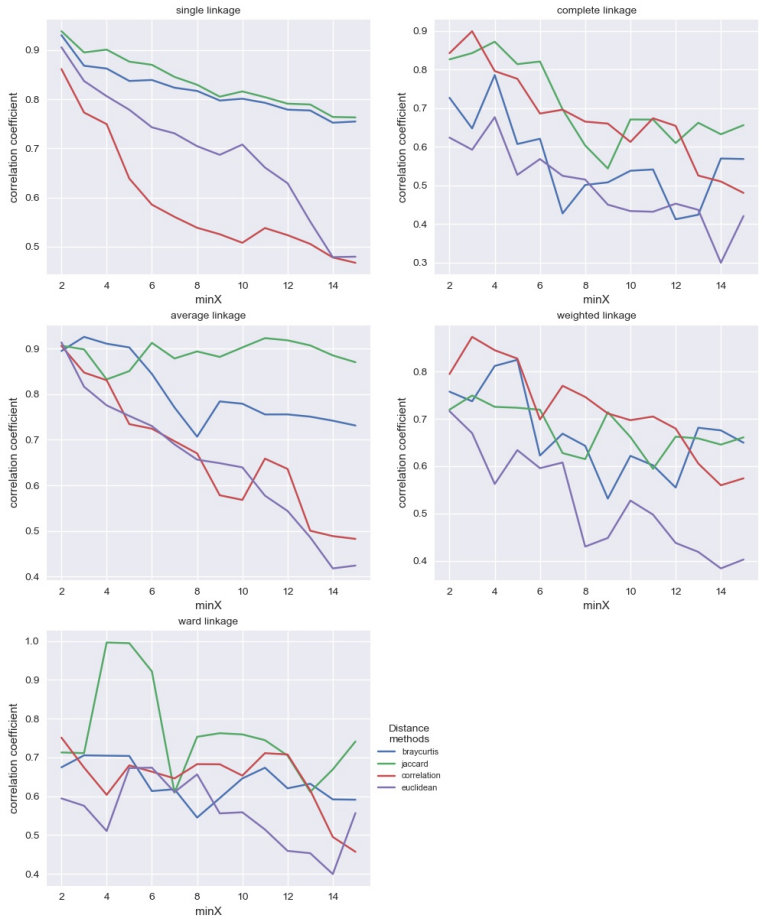


Figure 22: The correlations coefficients of the cophentic distance matrix of the original clustering with the clustering result using the filter x_s. The data from the qPCR was used together with a Log transformation.

Appendix B: Clustering results

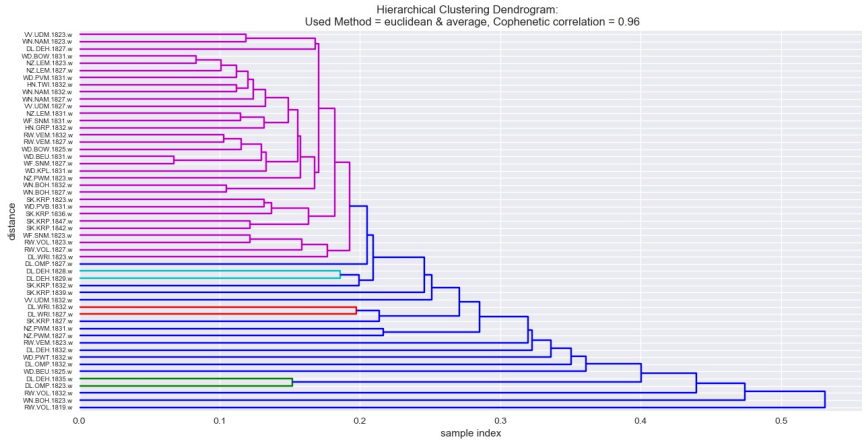


Figure 23: The clustering result using Euclidean distance with average linkage. The data from the ratio was used.

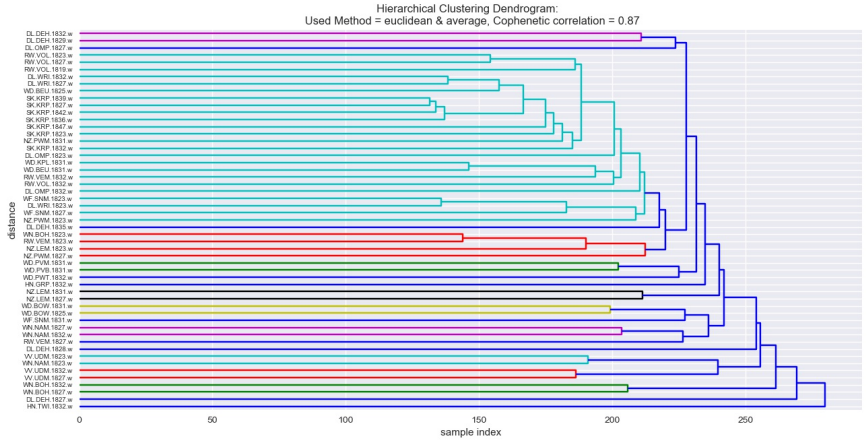


Figure 24: The clustering result using Euclidean distance with average linkage. The data from qPCR was used together with the Log transformation.

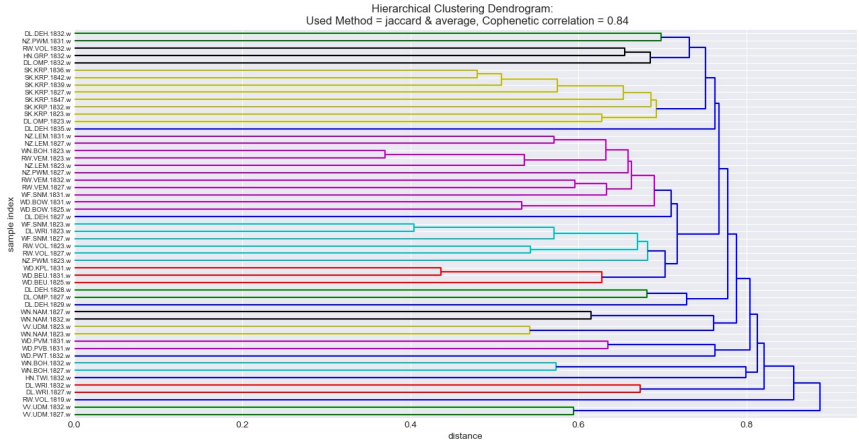


Figure 25: he clustering result using Jaccard dissimilarity with average linkage. The data from the ratio was used.

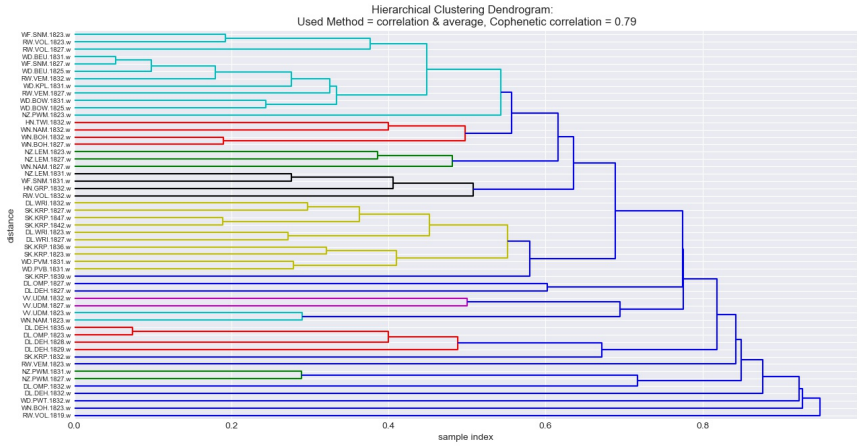


Figure 26: The clustering result using correlation dissimilarity with average linkage. The data from the ratio was used.

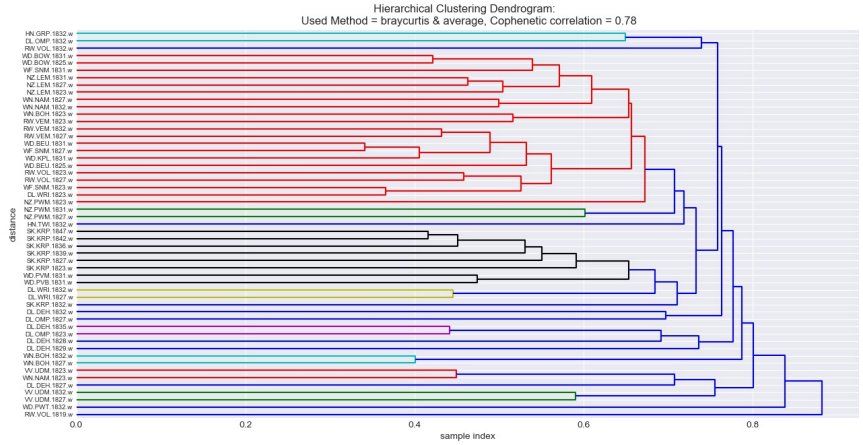


Figure 27: The clustering result using Bray-Curtis dissimilarity with average linkage. The data from the ratio was used.

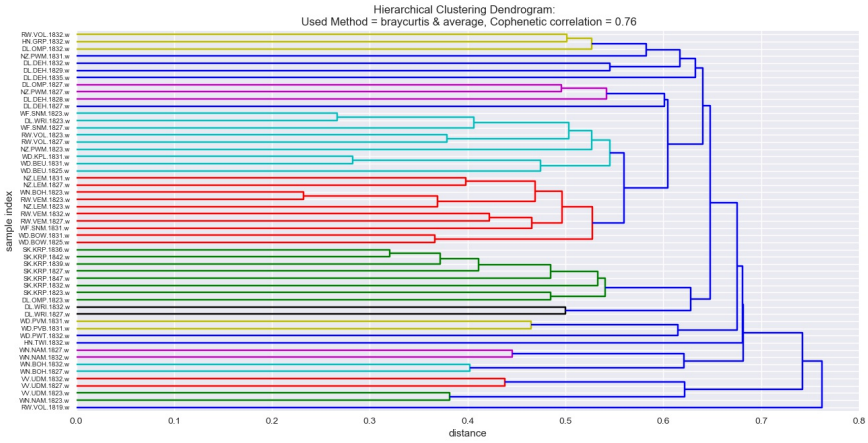


Figure 28: The clustering result using correlation dissimilarity with average linkage.The data from qPCR was used together with the Log transformation.

Appendix C: Additional PCA figures

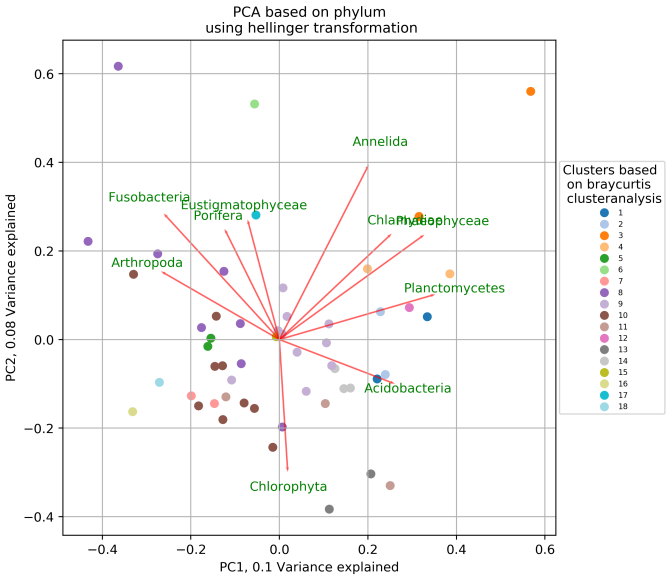


Figure 29: Plot of the samples using the first two principal components using the Hellinger transformation for Phylum level.

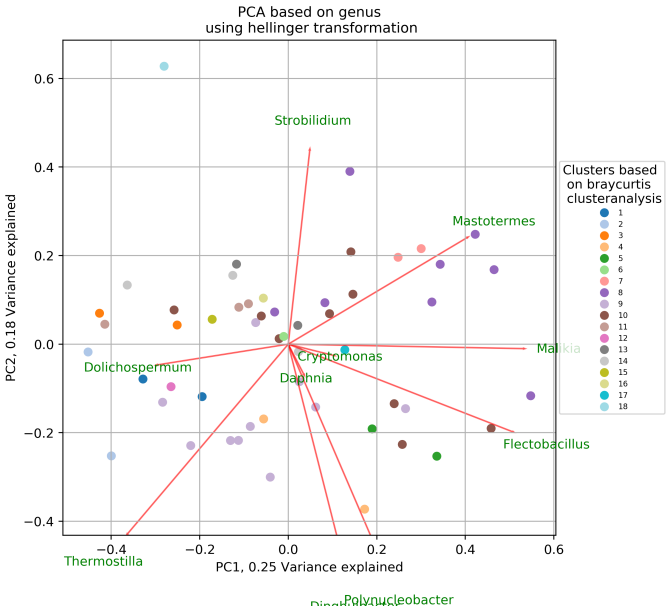


Figure 30: Plot of the samples using the first two principal components using data for Phylum level with $x_t = 20$.

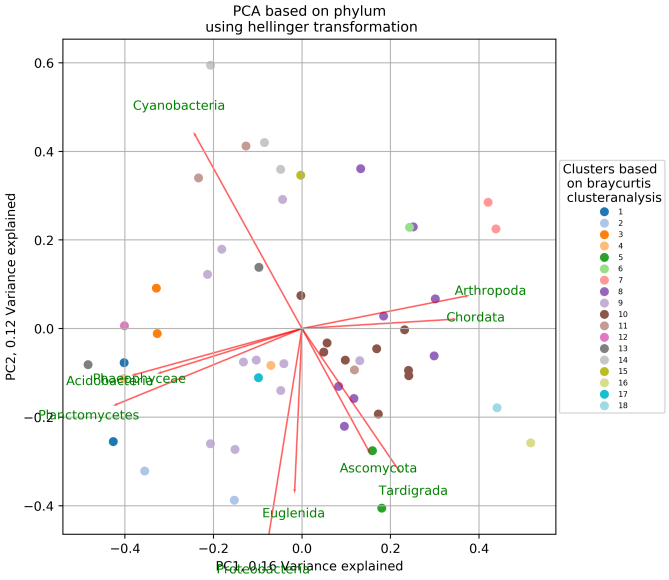


Figure 31: Plot of the samples using the first two principal components using data for Phylum level with $x_t = 20$.