The Activated Sludge of Danish Wastewater Treatment Plants

Exploring differences in the microbial communities using ordination

Master Thesis in Biotechnology by:

Kasper Skytte Andersen

AALBORG UNIVERSITY
DENMARK

Center for Microbial Communities

Supervisors:

Mads Albertsen

Per Halkjær Nielsen

31. May 2017

Preface

This master thesis report is written by Kasper Skytte Andersen during the period 1/9/2016 - 1/6/2017 as a requirement of the master of science degree in Medical Biotechnology at Aalborg University. The author is thankful to the supervisors Mads Albertsen and Per Halkjær Nielsen for their keen guidance during the project period, and also a big thanks to the Environmental Biotechnology group of Aalborg University for their superb laboratory preparation of samples taken during the last decade from Danish wastewater treatment plants, which this project is largely based on.

This master thesis is written so that main chapters are numbered, and figures and tables in the chapters are numbered according to the chapter and refered to hereby. References are cited with a paranthesis containing the surname of the main author and year, and multiple references are separated with a semi-colon. A full list of references with the complete citations can be found in the end of the thesis aswell as three appendices with supplementary information and plots.

Table of Contents

Chapter 1: Introduction **1**

1.1 The history of wastewater treatment and the activated sludge process 1

1.2 Identification of microorganisms in activated sludge 2

1.3 Purpose 6

Chapter 2: Ordination in Microbial Ecology **7**

2.1 Exploratory vs. Explanatory 8

2.2 Niche theory and the double-zero problem 12

2.3 Distance-based ordination 14

2.3.1 Principal Coordinates Analysis 14

2.3.2 non-Metric Multidimensional Scaling 15

2.3.3 Distance- and (dis)similarity measures 17

2.4 Eigenanalysis-based ordination 19

2.4.1 Principal Components Analysis 20

2.4.2 Redundancy Analysis 22

2.4.3 Correspondence Analysis and Canonical Correspondence Analysis 23

Aims **25**

Chapter 3: Materials and Methods **27**

3.1 Sampling 27

3.2 DNA extraction . . . 27
3.3 Library preparation, purification and pooling . . . 28
3.4 DNA sequencing and bioinformatics . . . 28
3.5 Data processing and analysis . . . 29

Chapter 4: Exploring the Microbial Communities of the WWTPs . . . 31
4.1 Overview of the differences . . . 32
4.2 How does the microbial community composition describe the WWTPs? 38
4.3 Concluding remarks . . . 44

Chapter 5: Explaining the Microbial Communities of the WWTPs . . . 45
5.1 The influence of plant design on the microbial communities . . . 45
5.2 General differences related to sampling time . . . 50

Chapter 6: General Discussion . . . 53

Conclusion . . . 57

Appendix A: Supplementary plots . . . 59

Appendix B: Scree plots . . . 69
B.1 Scree plot of Figure 4.1 . . . 69
B.2 Scree plot of Figure 4.2 . . . 70
B.3 Scree plot of Figure 4.3 . . . 70
B.4 Scree plot of Figure 5.2(A-D) . . . 71
B.5 Scree plot of Figure 5.3(A+B) . . . 73

Appendix C: Characteristics of the WWTPs . . . 75

References . . . 77

Abstract

Next-generation sequencing (NGS) technologies are today widely used to analyse the composition of microbial communities by sequencing the DNA of variable regions in the 16S rRNA gene. However, the amount of data obtained from the sequencing of numerous samples is often huge, which complicates the subsequent analysis of the microbial communities. The need for new bioinformatic tools to simplify the analysis has therefore increased proportionally to the high throughput of NGS technologies.

In this report the usefulness of ordination methods in the comparison of 622 samples from the activated sludge (AS) of 32 Danish wastewater treatment plants (WWTPs) is specifically investigated, and software is developed in the statistical programming language R to perform the analyses with ease. A detailed overview of different ordination methods and their uses in microbial ecology is provided, and the general differences between the microbial communities of the AS of the 32 WWTPs are investigated using 16S amplicon sequencing and different ordination methods.

The AS of the 32 WWTPs were found to have many abundant bacteria in common, where the 5 most abundant genus-level Operational Taxonomic Units (OTUs) made up more than 20% of the total number of OTU reads, and specifically *Tetrasphaera* were found abundant in most samples. Several groups of WWTPs appeared similar according to Principal Components Analysis (PCA) and Principal Coordinates Analysis (PCoA) with the Bray-Curtis dissimilarity measure (BCD), and Canonical Correspondence Analysis (CCA) showed similar patterns, but were able to reveal unique OTUs only present in the Ribe and Esbjerg E+W WWTPs. Furthermore,

different design characteristics of the WWTPs were found to be correlated with differences in the microbial community composition of the AS according to CCA, and minor differences were observed between which time of the year the samples were taken.

1. Introduction

1.1 The history of wastewater treatment and the activated sludge process

Cleaning wastewater is today an essential part of most modern civilised cities (Orhon, 2015). However, the practice of cleaning wastewater has only been performed within the last century or so. In fact, until the turn of the 19th century, most cities did not even have a proper disposal system for sewage, it was simply poured onto the streets (Orhon, 2015). It was first in the 1840s that the english lawyer Edwin Chadwick encouraged people to take 'sanitary responsibility' and find a solution to the many problems associated with having sewage floating on the streets (Hamlin, 1988). A few decades later, in the 1870s, the idea of leading the sewage through underground pipes called 'sewers' was widely adopted. The sewage was then primarily disposed of in one of two ways through irrigation; either it was led to the lands nearby, where farmers used it as a fertiliser, or it was distributed below the surface and penetrated into the soil in large sewage farms near the cities. Any remaining wastewater was simply led into nearby rivers, lakes or oceans, leading to eutrophication and pollution of the water bodies, which are major reasons why proper wastewater treatment is important (Orhon, 2015).

In the early 1890s, sewage disposal was revolutionised by biological treatment of the sewage using contact filters invented by William Joseph Dibdin (Hamlin, 1988). It was then possible to facilitate 'purification' of the sewage by immobilising naturally occuring bacteria in aeration tanks, which seemed to oxidise and destroy

much of the sewage. During the next few decades, remarkable research in the fundamental understanding of wastewater treatment was achieved. The study of a 'compact brown growth' suspended on filters or slates in aeration tanks was particularly investigated, as it was often associated with the removal of suspended solids and shorter aeration times (Clark & Adams, 1914). In 1914, Dr. Fowler wanted to test an idea by doing some 'experiments on the oxidation of sewage without the aid of filters', which were carried out by his graduate engineers Edward Ardern and William Lockett (Ardern & Lockett, 1914). These pioneering experiments were to set a milestone in wastewater treatment with the discovery of what they called *Activated Sludge* (AS). By reusing the biologically active 'compact brown growth', which seemed to accumulate during sewage treatment, they were able to drastically improve the rate with which the sewage was treated. They simply performed lab-scale aerated batch experiments with sewage from the Manchester sewage treatment plant, where the accumulated AS was collected after 'complete oxidation of the sewage' (the term used at the time to describe the complete removal of organic matter from the water) and added to subsequent batches of sewage. They observed that the time needed for complete oxidation decreased with every batch, from weeks in the initial batches to less than a day in the final batches. These simple experiments founded the *Activated Sludge Process* (ASP), where the accumulated sludge is continously reused, which ever since has been the backbone of most wastewater treatment plants (WWTPs) in the world (M. C. van Loosdrecht, Nielsen, Lopez-Vazquez, & Brdjanovic, 2016).

1.2 Identification of microorganisms in activated sludge

Even a century ago, when Ardern and Lockett made their groundbreaking discovery, it was well known that microorganisms played a vital role in the processes involved in the purification of wastewater (Fuller, 1915; P. H. Nielsen & McMahon, 2014). At

the time, however, little was known about which microorganisms were present in the AS or their particular role in wastewater treatment. The stability of operation and performance of early WWTPs established in the 1930-1940s were often unstable and many problems were encountered with e.g. foaming and bulking (Orhon, 2015). Research into the microbiology of AS was therefore important, but the experimental methods available at the time to identify the microorganisms were limited, and research was primarily based on biochemical characterisation by isolation and cultivation, or morphological characterisation by light microscopy (M. C. van Loosdrecht et al., 2016). Wastewater treatment was in the first half of the 20th century a highly empirical practice based on trial-and-error, and it was first in the early 1970s that significant research into the identity of problem-causing bacteria was done (Eikelboom, 1975; M. C. van Loosdrecht et al., 2016).

Until now, much research has been done into linking the identity of microorganisms present in AS to their metabolic functions relevant to wastewater treatment. The combination of the development of molecular tools, e.g. *Polymerase Chain Reaction (PCR)* (Mullis et al., 1986), various fluorescent oligonucleotide probes (Amann et al., 1990; DeLong, Wickham, & Pace, 1989), as well as DNA sequencing technologies have allowed for precise identification of the microorganisms present in the AS. This allows for a complete picture of the microbial community and has become invaluable information to the understanding of wastewater treatment. Today, the traditional *Sanger DNA sequencing* method developed in the late 1970s (Sanger & Coulson, 1975; Sanger, Nicklen, & Coulson, 1977) has been gradually replaced by high-throughput *next generation sequencing (NGS)* technologies (e.g. Roche 454 Life Sciences Pyrosequencing, Illumina, and lately Oxford Nanopore) as they are significantly cheaper and faster (Metzker, 2010). By isolating and sequencing a relatively short part of a 'fingerprint' gene like the 16S ribosomal RNA (16S rRNA) gene, it is possible to identify the microorganisms present based on variable regions in the gene (Figure 1.1) within a day.

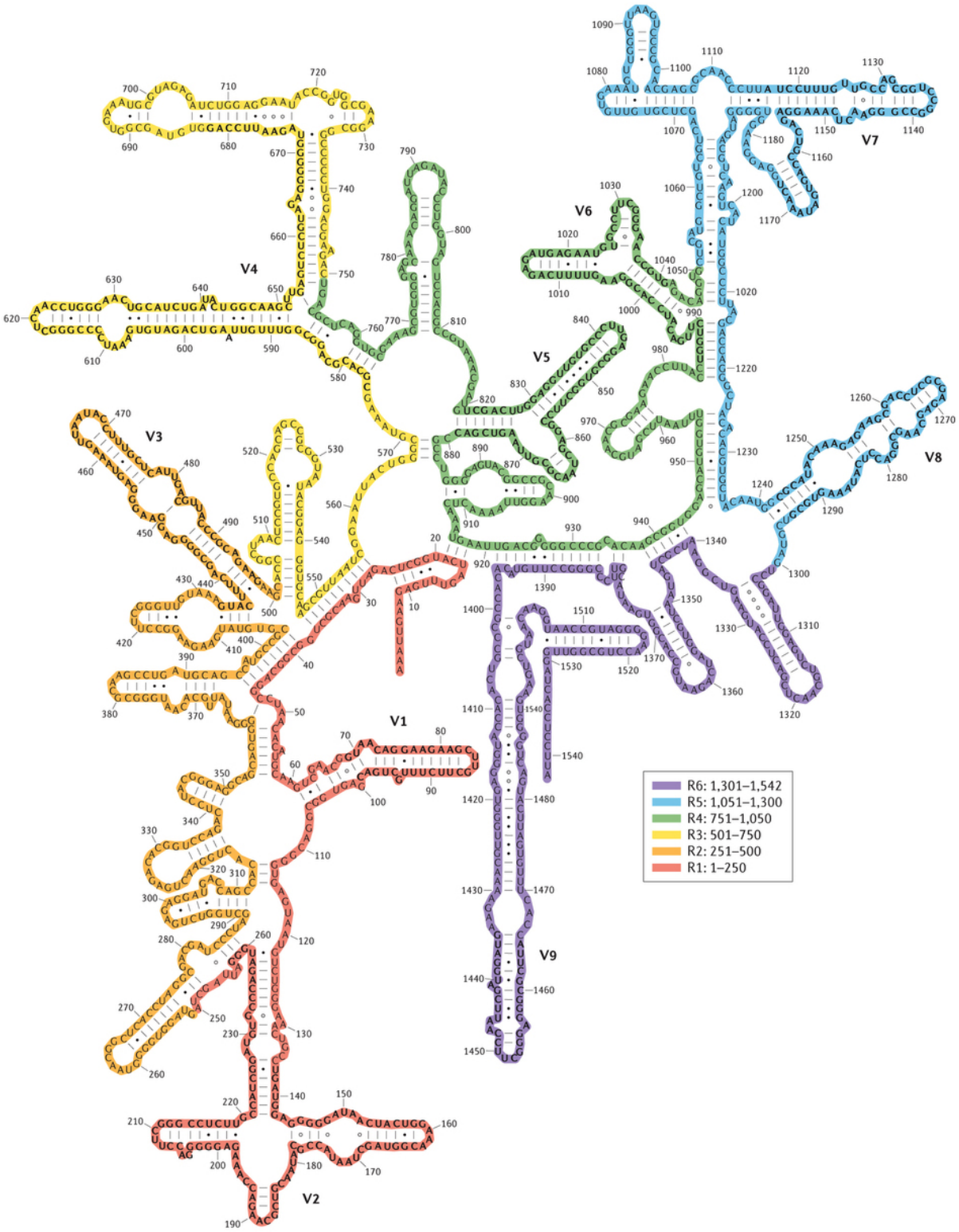


Figure 1.1: The secondary structure of the 16S rRNA gene of Escherichia coli with the variable regions V1-V9 indicated with bold text. The colors represent the separate regions of the gene that have been sequenced individually. Source: (Yarza et al, 2014)

Due to its vital function in the ribosome for protein synthesis, the 16S rRNA gene is a highly conserved gene found in almost all prokaryotes (Clarridge, 2004). Since the final gene product is RNA, mutations are fatal since the exact nucleotide sequence is essential to preserve its activity (D. Qin, Abdi, & Fredrick, 2007). Some regions of the gene are more important than others, however, which allows for evolutionary relationships to be drawn from the less preserved, variable regions (Clarridge, 2004). In the 16S rRNA gene there are 9 of these regions, V1-V9, (Figure 1.1), which can then be used as phylogenetic markers (Ashelford, Chuzhanova, Fry, Jones, & Weightman, 2005; Woese & Fox, 1977). By sequencing the DNA of one or more of these variable regions of the 16S rRNA gene using NGS, it is then possible to identify the majority of microorganisms present in a sample. This method is called 16S rRNA *Amplicon Sequencing* and is today widely used and the preferred choice to analyse microbial communities (M. C. van Loosdrecht et al., 2016). The fundamental steps of *Amplicon Sequencing* (Figure 1.2) requires both laboratory practices with DNA extraction, PCR amplification of one or more of the variable regions, and subsequent sequencing of the DNA libraries, as well as bioinformatic processing of the obtained DNA reads.

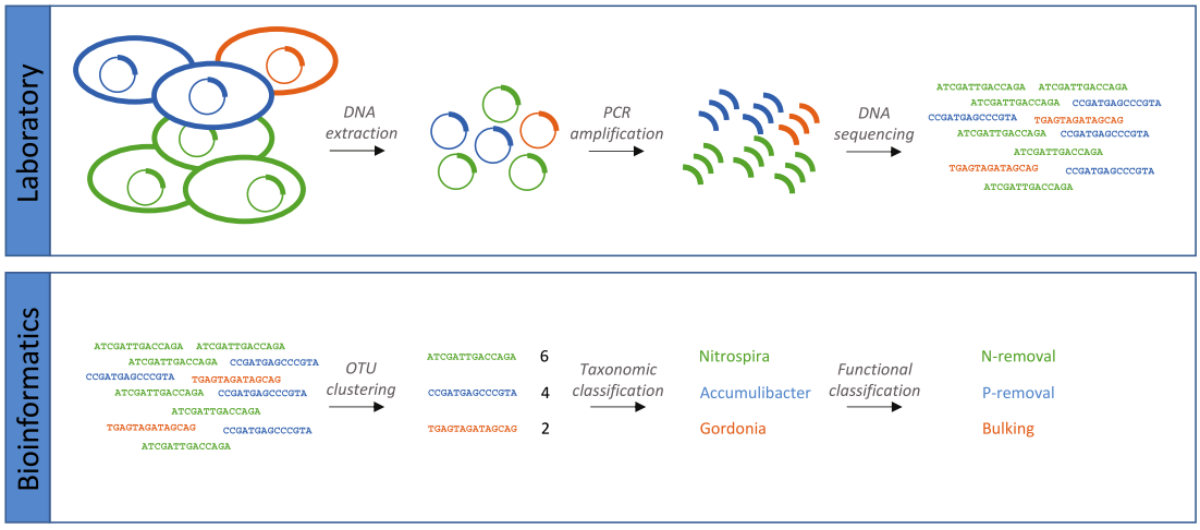


Figure 1.2: The fundamental steps in the analysis of microbial communities using 16S amplicon sequencing. Source: (Loosdrecht et al, 2016)

The latter step involves clustering of the reads by sequence identity into *Operational Taxonomic Units* (OTUs), which are then inferred taxonomic classification based on a suited reference database, of which the most commonly used today are SILVA (Quast et al., 2013), RDP (Cole et al., 2014), Greengenes (McDonald et al., 2012), or specifically for AS, the MiDAS database (S. J. McIlroy et al., 2015). The next step is data analysis, which highly depends on the purpose of the particular study.

1.3 Purpose

The rapid development of NGS technologies within the last decades has resulted in a growing need for sofisticated bioinformatic tools to analyse the large, complex and multivariate data sets obtained with these high-throughput technologies. The purpose of this study is to develop bioinformatic tools to ease the analysis of data obtained using NGS. Specifically, the usefulness of ordination for the analysis and comparison of the microbial communities of the activated sludge in Danish wastewater treatment plants will be investigated, and a handful of questions (listed in Aims) will be answered using primarily ordination.

2. Ordination in Microbial Ecology

To analyse the very large and complex multidimensional data obtained using NGS of sometimes hundreds of samples, each containing hundreds of different microorganisms, ordination is particularly suited because it can help reveal overall patterns in the data (Ramette, 2007). In essence, ordination seeks to reduce the dimensionality of a contingency table (objects in rows, descriptors in columns) into a few, usually 1-2, more important dimensions to ease interpretation, which makes it particularly suited for complex ecological data. Ordination methods are originally developed as largely heuristical approaches to solve particular ecological problems and most of the current ordination methods have been developed independently, but are similar in principle. Exactly when the first ordination method was developed is difficult to tell, but during the 1950s the term 'ordination' (from the German word 'ordnung', to order) were beginning to emerge to describe methods of classifying vegetation (Goodall, 1953), and several statistical approaches to answer specific ecological questions were proposed. For example the widely used Bray-Curtis measure was originally developed to examine the dominance behaviour of tree- and plant species in a forest in Southern Wisconsin (Bray & Curtis, 1957). Today, ordination methods have been used in countless ecological studies of the diversities of plants and animals, and have in the 21st century been readily applied in microbial ecology aswell (more than 7000 published studies) (Kuczynski et al., 2010; Ramette, 2007).

When performing ordination, $n - 1$ new, artificial dimensions are obtained through 'dimensional yoga', where n is the total number of objects, or samples in the case of ecology, each containing a part of the total inertia in the data, whether it

be (co)variance, (dis)simililarity, or any other statistical property. The first axis will then display the most inertia, the second axis the second most, the third axis the third most etc, and plotting the first, usually two, axes can then reveal interesting patterns between the samples, simply by interpreting the distances between the points. It can be difficult for the human mind to grasp more than 3 dimensions (x,y,z), because this is something that only exists in math, and the complex math behind the scenes lies beyond the scope of this report. However, there are various different types of ordination, each suited for a particular purpose and understanding the key differences between them, which to use when, and why is important. The most commonly used ordination methods in microbial ecology will be described in the following.

2.1 Exploratory vs. Explanatory

The most commonly used ordination methods can generally be divided into two groups based on their purpose. The first group is the *exploratory* analyses, also known as *unconstrained* or *indirect gradient* analyses, which are suited for identifying global patterns between the objects (samples) based on the distribution or (dis)similarity of the values of multiple variables (species abundances) associated with them. The exploratory analyses do not take environmental variables (fx sample location, pH, temperature, nutrient concentrations etc, both qualitative or quantitative) into account and thus do not explain the revealed patterns directly. It is still possible to color or shape the points by known environmental variables (see Figure 2.1), however, but the scores (coordinates) on the ordination axes remain the same. The most commonly used exploratory methods in microbial ecology are Principal Components Analysis (PCA), *non-Metric* Multidimensional Scaling (nMDS), Principal Coordinates Analysis (PCoA/*metric* MDS) and Correspondence Analysis (CA) (Ramette, 2007).

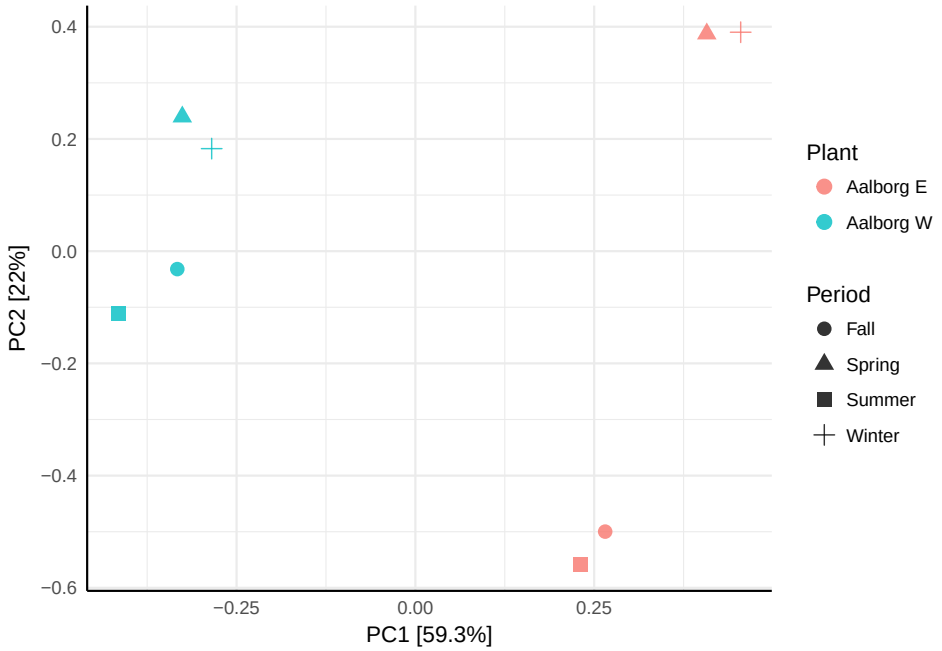


Figure 2.1: A minimal example of exploratory data analysis using Principal Components Analysis. 8 samples from two wastewater treatment plants, Aalborg East and Aalborg West have been analysed and the points have been shaped by when the samples were taken in 2012.

The second group is the *explanatory* analyses, also known as *canonical*, *constrained* or *direct gradient* analyses, which show only the variation in the data that can be explained by *known* environmental variables and not all the variation in the data as with unconstrained analysis. The response variables (species abundances) are thus considered to be the result of gradients along the environmental variables, or a combination of them. These gradients are called environmental gradients and the constrained ordination methods mainly differ in how they mathematically hypothesise the distribution of the response variables along the environmental gradient(s) to be, either linear or unimodal (P. Legendre & Legendre, 2012). Currently the two main constrained ordination methods used in microbial ecology are Redundancy Analysis (RDA) and Canonical Correspondence Analysis (CCA), which are considered extensions of Principal Component Analysis (PCA) and Correspondence Analysis (CA), respectively. RDA (and PCA for unconstrained analysis) is the optimal choice for purely linear distributions along the, preferably short, environmental

gradient(s). CCA (and CA for unconstrained analysis) is the optimal choice for unimodal distributions along longer gradients where many double-zeros occur (more about double-zeros in Chapter 2.2), but in most cases CCA also performs well with short and linear gradients, it will just show a more qualitative representation of the samples (Braak & Prentice, 1988). Both RDA and CCA are eigenanalyses and calculate their constrained/canonical axes by introducing a linear combination of the response variables and the environmental variables as an additional step in the procedure. Otherwise the procedure is identical to that of PCA (when performing RDA) or CA (when performing CCA) (P. Legendre & Legendre, 2012). An overview of commonly used ordination methods in microbial ecology can be found in Table 2.1, and the kind of plots obtained when performing ordination is illustrated in Figure 2.2.

Table 2.1: A classification of the most used ordination methods in microbial ecology. Based on (Braak & Prentice, 1988; P. Legendre & Legendre, 2012; Ramette, 2007)

Unconstrained analyses	**Constrained analyses**
Eigenanalysis-based	
Principal Components Analysis (**PCA**)	Redundancy Analysis (**RDA**)
Correspondence Analysis (**CA**)	Canonical Correspondence Analysis (**CCA**)
Distance-based	
non-metric Multidimensional Scaling (**nMDS**)	
Principal Coordinates Analysis (**mMDS/PCoA**)	

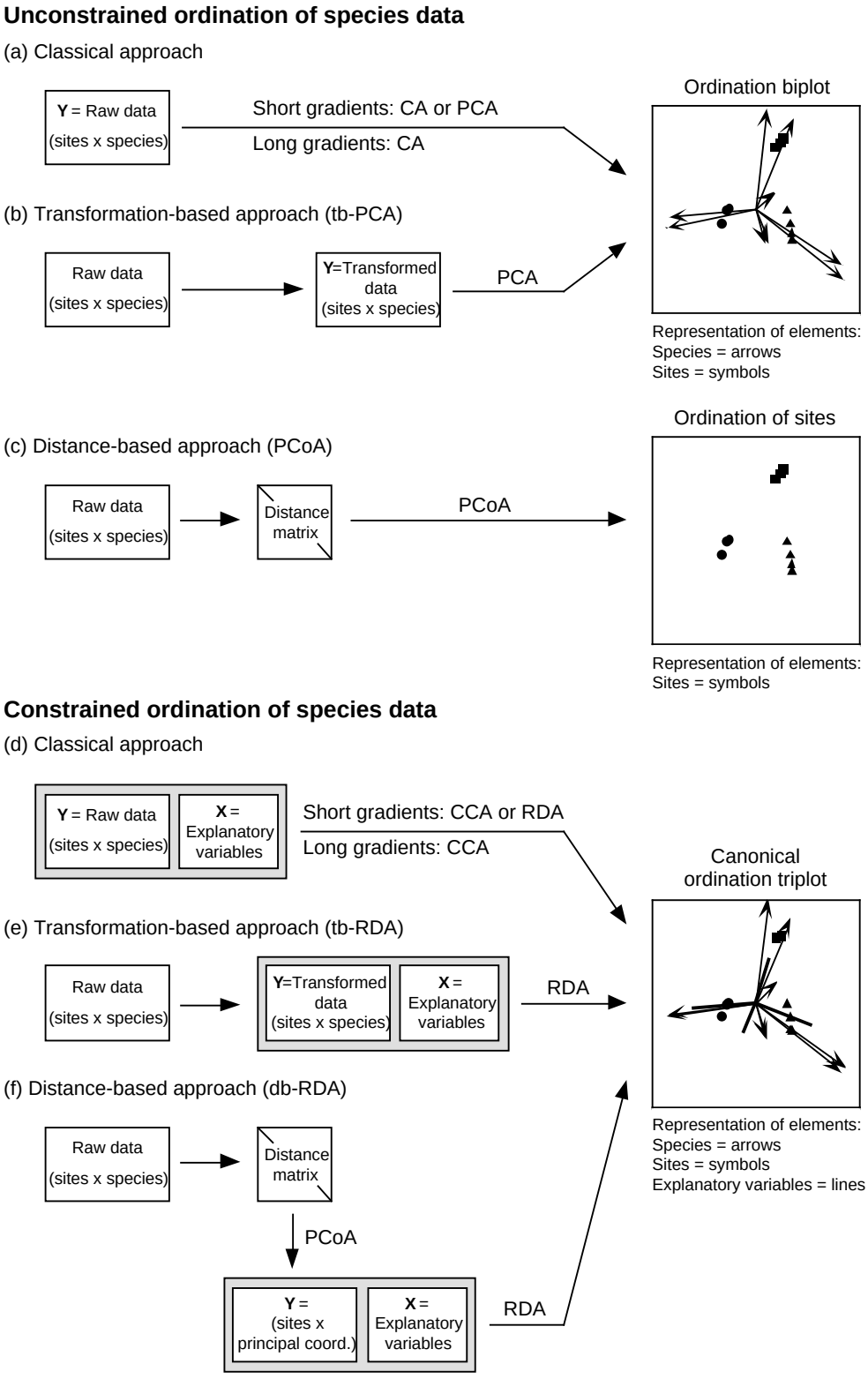


Figure 2.2: Schematic comparison of techniques that can be used to obtain unconstrained (a,b) or constrained (c-e) ordination biplots or triplots of species data tables. Source: (Legendre and Gallagher, 2001)

2.2 Niche theory and the double-zero problem

In reality there are often many, known or unknown, environmental variables affecting the presence of species and the gradient is then considered a complex environmental gradient. As niche theory states, species have ecological preferences and are present under a set of optimal environmental conditions, including the presence of other species (Hutchinson, 1957). The theory also predicts that species have unimodal distributions along environmental gradients, illustrated in Figure 2.3, so that they are found in greater abundances at some intervals along the major environmental gradients and gradually less present away from that optimal set of thriving conditions, ultimately absent (Whittaker, 1967). This has the consequence that community composition data typically contain many zeros, which can pose a problem for some ordination methods, specifically those using Euclidean distances like PCA and RDA (P. Legendre & Legendre, 2012).

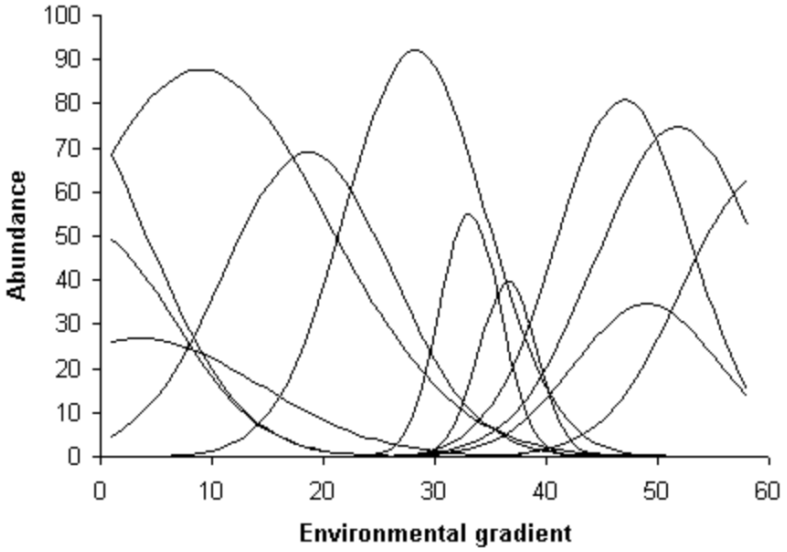


Figure 2.3: Species response (or abundance-) curves in most cases show a unimodal distribution along an environmental gradient. Adopted from (Whittaker, 1967)

The goal of using ordination is to represent patterns or differences between samples based on species abundances to draw ecologically meaningful conclusions

about the sampling site(s) and their corresponding β- or γ-diversity (the diversity of local sites or whole regions of sites, respectively). The species abundances are considered to be response variables in the sense that they are indicators of their nearby environment. Variation in the environment is expected to be reflected in the relative productivities or abundances of the species (Whittaker, 1972). If a species is present at two sites, this means that the sites share conditions that are favorable for the species, indicating a similarity between the sites. On the contrary, if a species is present at one site and not the other, this indicates that the ecological conditions at the sites most likely are dissimilar. However, if a species is absent from both sites, it provides no valuable ecological information since both sites either have ecological conditions outside the niche of the species, and these conditions may be very similar or very different, or the niche has been occupied by another species prefering the same conditions. Including double zeros in the comparison of sites thus results in a higher similarity between the sites than there is ecologically, and this phenomena is called the *double-zero problem* (P. Legendre & Legendre, 2012). It is therefore important to consider whether there is relevant ecological information in double-zeros in the particular study, but in most cases it is preferable not to conclude anything ecologically about them and avoid including them in the analysis. This is called asymmetric analysis as double presences are treated differently than double absences and is done in practice when the distance coefficients between sites are computed, either implicitly or explicitly during ordination (P. Legendre & Legendre, 2012). Choosing the correct ordination method and, for some, also a distance measure therefore depends on whether this information is meaningful for the study. If linear ordination methods like Principal Components Analysis or Redundancy Analysis is used, then they should first be subjected to appropriate data transformation to correct for the problem and reflect the ecological differences more correctly (P. Legendre & Legendre, 2012).

2.3 Distance-based ordination

Fundamental to any ordination method is not only dimensional reduction, but more importantly also how it does so and how it displays and calculates distances between objects (sites) or variables (species). As stated in Table 2.1, ordination methods can be classified by whether they are distance-based or not, however they could just as well have been classified by whether they use a distance measure explicitly or implicitly to calculate distances, respectively (P. Legendre & Legendre, 2012). As such, eigenanalysis-based ordination methods use a distance measure of their own as part of their algorithm, while distance-based ordination can only be performed on a distance- or (dis)similarity matrix, which has to be calculated first using one of many distance- or (dis)similarity measures. When calculated, a symmetrical matrix is then obtained containing *association coefficients* between all pairs of sites. This gives the ecologist flexibility as to how it would be ecologically meaningful to represent the differences between sites or species, however it also implies the importance of knowing how to interpret the results based on the particular measure and choosing it wisely (more on distance measures in Chapter 2.3.3) (P. Legendre & Legendre, 2012). One of the disadvantages of using distance-based ordination, however, is that it is not possible to plot both sites and species together in a biplot, only the sites are plotted.

Currently the most used distance-based ordination methods are *non-Metric* Multidimensional Scaling (nMDS) and Principal Coordinates Analysis (PCoA), also called *metric* Multidimensional Scaling (mMDS) (Ramette, 2007). There is also Polar Ordination, which is rarely used and will not be detailed in the following.

2.3.1 Principal Coordinates Analysis

Principal Coordinates Analysis (PCoA) is very useful at exploring microbial ecology data because it can represent relationships between samples measured by any distance coefficient in Euclidean space (also called metric space, fx a Cartesian

coordinate system). Therefore PCoA is also called *metric* Multidimensional Scaling, because it can represent the *metric* properties of the distance- or (dis)similarity matrix. When the distance measure is Euclidean, the results obtained by PCoA is identical to that of PCA (P. Legendre & Legendre, 2012).

Since the choice of distance measure directly influences the result, it has to be done with care. However, this is also one of the advantages of using PCoA because it is then possible to better deal with ecological problems like the *double-zero problem* while still using Euclidean mapping. The choice of distance measure also influences how the results are to be interpreted, as the original data is then a function of the chosen measure, which may be non-Euclidean, and does not always allow for a true representation in Euclidean space. Furthermore, it is possible to obtain negative eigenvalues of the resulting axes, especially when semi-metric or non-metric measures are used, and this should be corrected for if they occur on the main axes (ie the axes being plotted). The eigenvalue of an axis is an indication of its contribution of inertia to the total inertia in the data, which is therefore also obscured by the choice of distance measure and its value should not be directly refered to when performing PCoA (Ramette, 2007). PCoA is partly based on eigenanalysis, however it is more appropriate to classify it as a distance-based analysis, since it is highly dependent on the chosen distance measure.

2.3.2 non-Metric Multidimensional Scaling

Non-Metric Multidimensional Scaling (nMDS) is similar to PCoA in that it is also performed on a distance- or (dis)similarity matrix calculated using a suited measure. However, nMDS is different from PCoA on numerous aspects. PCA aswell as PCoA both try to maximise a linear correlation (of course dependent on the metric used in the case of PCoA) of species abundances along the environmental gradient, which will often result in an artifact called *the horseshoe effect* when the response variables are the result of a non-linear or long gradient (Podani & Miklós, 2002). This results in an arch shaped and incorrect pattern of the points leading to false conclusions. nMDS eliminates this by only preserving the ranked order of the distance coefficients

between samples. As the name suggests, this makes the procedure *non-metric*, and the distances between points is therefore not to be interpreted numerically.

nMDS does not try to explain as much variation in the data as possible as with PCA or PCoA, but more the discontinuities in the data. It is a very robust technique that can handle missing values aswell as multiple data types at once. It has no distributional assumptions about the data compared to all other ordination methods, where the data is assumed to have for example linear or unimodal distributions as with PCA/RDA and CA/CCA, respectively. nMDS is therefore the ordination method of choice when the nature of the data is unknown (Buttigieg & Ramette, 2014; P. Legendre & Legendre, 2012).

nMDS is furthermore very different from other ordination methods by how it is computed. nMDS is an iterative procedure where the number of dimensions is chosen *a priori* (before analysis) and the algorithm tries to find a solution from either random starting points or from the results of a PCoA on the same data provided by the user. The solution is not unique as with all other ordination methods mentioned in this chapter and it is therefore recommended to run it multiple times to validate the result, preferably with different numbers of dimensions. During the algorithm, a stress value is calculated to express the goodness-of-fit of the solution and the procedure is repeated many times (20+ is not unusual) using the solution of the previous cycle as the starting point until the stress value does not change significantly and reach an acceptable, low value. A stress value is considered good when below 0.05, while below 0.1 is acceptable. A stress value above 0.2 is suspect and the results should not be trusted (Ramette, 2007). Generally choosing more dimensions will lower the stress value.

One of the major drawbacks of nMDS is that it is very computationally demanding. However, modern computers are getting increasingly more powerful, so this is less of a concern compared to the time during which the method was developed by the psychometricians Kruskal and Shepard at the Bell Telephone Labs in the 1960's (Kruskal, 1964; Shepard, 1966).

2.3.3 Distance- and (dis)similarity measures

As mentioned, choosing a distance- or (dis)similarity measure that makes ecological sense is crucial for the analysis and subsequent interpretation of the ordination. A distance measure is basically a mathematical function with which to calculate distances between objects or variables in the data. There are many, many ways (60+) of calculating distance/association coefficients between objects and/or variables, however only the general concepts and most important differences will be covered in the following. For in-depth knowledge of how to calculate association coefficients using all metrics, semi-metrics, non-metrics and their exact formulae, refer to Chapter 7 in *Numerical Ecology* (P. Legendre & Legendre, 2012).

For association coefficients to be considered metric, the four properties listed below have to be satisfied. When this is true, the coefficient is called a distance coefficient or a metric coefficient, since it can be fully represented in Euclidean (metric) space (P. Legendre & Legendre, 2012):

The four metric properties

1. The distance between identical objects is 0, which is the lowest possible value:
   if $a = b$, then $D(a, b) = 0$

2. When the compared objects are not identical, the coefficient, and the distance, has a positive value:
   if $a \neq b$, then $D(a, b) > 0$

3. Symmetry: the distance from A to B is the same as the distance from B to A:
   $D(a, b) = D(b, a)$

4. Triangle inequality: The sum of two sides of a triangle of points in Euclidean space is equal to or longer than the third side. In other words, the shortest distance between two points in Euclidean space is a straight line:
   $D(a, b) + D(b, c) \geq D(a, c)$

When one or more of the four properties are not satisfied, in most cases the fourth property (triangle inequality), the coefficients calculated are not considered to be metric coefficients, they are then termed *semi-metric* coefficients or (dis)similarity coefficients. When this is the case, distances cannot be ordinated (at least reliably) in Euclidean space and nMDS is the optimal ordination method. PCoA can also be used, but the eigenvalues of the axes plotted has to be first checked for negative eigenvalues. This is normally not the case for the main axes being plotted, however. Therefore considering whether the chosen measure is suitable for the ordination method used is important and the interpretation of the result should be done according to the logic of the chosen distance measure (P. Legendre & Legendre, 2012). Dissimilarity and similarity coefficients are often just termed similarity coefficients because it is straight forward to convert dissimilarity coefficients (D) to similarity coefficients (S) and vice versa: $S = 1 - D$ (P. Legendre & Legendre, 2012). An overview of some of the most commonly used metrics in ecology is summed in Table 2.2

Table 2.2: A classification of some of the most common distance- and (dis)similarity measures. The distance metrics are often just referred to as metrics or distances and the semi-metrics as (dis)similarity measures. Non-metrics are often problematic and counter-intuitive. (Buttigieg & Ramette, 2014)

Distance metrics	**semi-Metrics**	**non-Metrics**
Euclidean	Bray-Curtis (dissimilarity)	Kulczynski
Chord	Sørensen (dissimilarity)	
Mahalanobis		
χ^2 (Pearson chi-square)		
Manhattan		
Canberra		
Jaccard		

2.4 Eigenanalysis-based ordination

The eigenanalysis-based ordination methods have more specific purposes than the distance-based methods, because they are limited to represent the distances only by the capabilities of their implicit distance measure, where distance coefficients are not first calculated manually by the user. They all have a few things in common, however (P. Legendre & Legendre, 2012):

- There is always one unique solution with the particular data
- Each axis is an eigenvector associated with an eigenvalue expressing the axis' contribution to the inertia in the data
- The axes are ranked by (and plotted by) the eigenvalues, highest to lowest
- The axes are orthogonal to eachother, thus uncorrelated and express their own 'unique' inertia

Normally, the two axes with the highest eigenvalues are plotted in a Cartesian coordinate system, where the highest eigenvalue axis is represented by the first axis and the second highest on the second axis. The most inertia in the data is therefore always represented by the first (x-)axis, which can, for example in the case of two distinct groups of sample points as seen with the example in Figure 2.1, often be interpreted as 'between-group variation'. The secondmost inertia is expressed by the second (y-)axis and can then be interpreted as 'within-group variation' (P. Legendre & Legendre, 2012).

It is important to examine all eigenvalues obtained for each axis to confirm that the axes being plotted are significant and represent a large portion of the inertia in the data. To do this, a simple plot called a *scree plot* can be made where all axes are plotted on the first axis ordered by eigenvalue in decreasing order and their corresponding eigenvalue on the second axis, as illustrated in Figure 2.4. Optimally, the first two axes represent more than half of the inertia in the data and have high values compared to the rest of the axes, where the latter should make up a slightly decreasing, straight line. The worst case scenario is when all the values make up

a nearly horizontal, straight line. In this case, the ordination is either failing at representing the inertia in the data and a different ordination method may be better suited for the data or there is simply no inertia to represent at all. The sum of all eigenvalues is always equal to the total inertia in the data and the corresponding eigenvalues of the axes plotted are often shown as a percentage of the total inertia on the axis labels. In the following, the remaining four ordination methods listed in Table 2.1 will be explained briefly (Ramette, 2007).

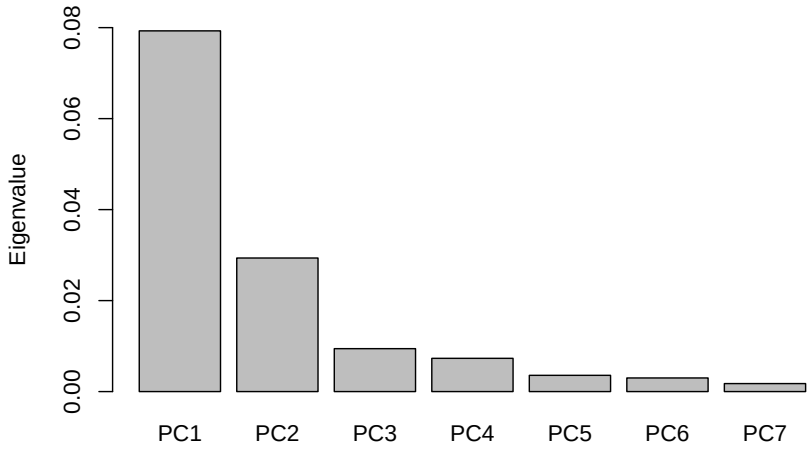


Figure 2.4: A scree plot of the eigenvalues of the 7 axes obtained from the PCA ordination seen in Figure 2.1.

2.4.1 Principal Components Analysis

Principal Component Analysis (PCA) is the oldest ordination method and still also the most used, perhaps due to its simplicity (Ramette, 2007). It has its roots all the way back to 1901 when Karl Pearson explained how to represent objects by the 'best-fitting' line or plane (Pearson, 1901). The simplest example of dimensional reduction is the representation of 2-dimensional data in 1 dimension, which is normally called linear regression. It is simply a straight line, drawn through the *centroid* of the data, see Figure 2.5.

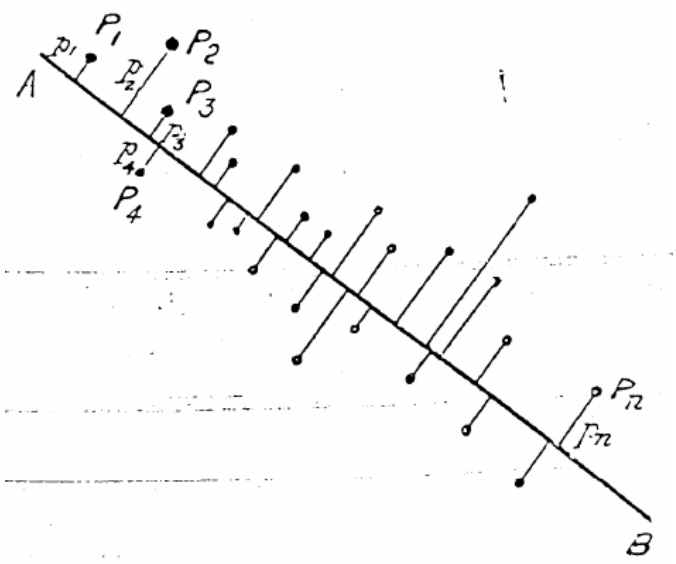


Figure 2.5: The main structure of two-dimensional data can often be described in one dimension by positions on a straight line. Adopted from (Pearson, 1901)

Pearson described how this is done mathematically, which formed the concepts of how to describe the overall structure of complex data. In 1933, Harold Hotelling (Hotelling, 1933) applied the concepts on multidimensional data and explained how to represent the data by its *principal components*, which in turn formed the fundamentals of PCA.

The goal of PCA is simple: express the maximum amount of variation in the data. This is done by generating new, synthetic axes, which are synonymous to eigenvectors, where the first axis is alligned so that it represents the most inertia in the data, where inertia in the case of PCA is specifically: *variance*. PCA is therefore considered more of a quantitative ordination method, as it excels at highlighting differences between samples based on numerical differences in species abundances and is most reliable when the same species are present in most or all of the samples (P. Legendre & Legendre, 2012). As mentioned in Chapter 2.2, this is rarely the case with ecological data, and the double-zero problem thus has a major impact on the results obtained by PCA, questioning its usefulnes without appropriate data transformation. Of course, this is not a problem when analysing α-diversities (the diversity of a single sample) since there will be no zero-abundances (P. Legendre &

Legendre, 2012).

PCA is a linear method because it represents the linear correlation or covariance of species abundances between samples using Euclidean distances. Both the sample scores (points) and species scores (arrows) are usually plotted together to form a biplot (refer to Figure 2.2), where arrows indicate the linear gradients of species abundances and the relative right-angle projections of the samples onto the arrows then approximate their corresponding abundance of the particular species. This makes PCA suited to answer questions like for example: "How are samples different in terms of the abundances of species X and Y?", where X and Y are also among the most abundant species (Ramette, 2007).

Fundamental to PCA (and Redundancy Analysis) is the Euclidean distance, which is calculated by the following formula, equivalent to the pythagorean theorem:

The euclidean distance

$$D_{Euclidean}(x_1, x_2) = \sqrt{\sum_{j=1}^{p} \left(y_{1j} - y_{2j}\right)^2} \tag{2.1}$$

where $Y = [y_{ij}]$ is a species abundance table of the size $(n \times p)$ with sites (rows) $i = [1...n]$ and species (columns) $j = [1...p]$. (P. Legendre & Gallagher, 2001)

2.4.2 Redundancy Analysis

To answer a question like: "How do species abundances correlate with a measured carbon source concentration?", PCA would be insufficient. To answer that question, Redundancy Analysis (RDA) is the perfect choice. RDA is considered an extension of PCA, or the constrained version of PCA, which can directly explain the observed patterns based on known environmental variables. This is done by making a linear combination of the species abundance matrix and a matrix containing information about one or several known environmental variables. These variables are then hypothesised to be able to explain a portion of the observed variance in the data and

the resulting constrained axes are plotted. Just as with unconstrained ordination, the two axes with the highest eigenvalues are plotted and their contribution to the total variance in the data is usually indicated on the axis labels to give an indication of the significance of the constrained variable(s) (Buttigieg & Ramette, 2014). When doing constrained analysis, both with RDA and CCA (Chapter 2.4.3), a plot with not only sites and species are made, but also environmental vectors are plotted together in one plot called a triplot (refer to Figure 2.2). This is a very convenient way to visualise all three types of information at once to easily draw ecological conclusions about the sites sampled.

Because RDA is based on Euclidean distances just like PCA, it is only suited for the analysis of short, linear environmental gradients with few or no species absences, which can limit its use in ecology (Minchin, 1987; Ramette, 2007).

2.4.3 Correspondence Analysis and Canonical Correspondence Analysis

Perhaps the most appropriate ordination method for ecological data is Correspondence Analysis (CA) because it represents the differences between sites hypothesising a unimodal distribution, which fits the species niche theory well, as discussed in Chapter 2.2. As the name suggests, CA tries to represent the *correspondence* between samples and species by showing how "this species correspond to that site". The relative positions between species points and samples points are then to be interpreted as *probabilities* of the samples to contain the particular species nearby (P. Legendre & Legendre, 2012). CA is thus considered more of a qualitative representation of the data and is based on the Pearson chi-squared statistic of the form $\chi^2 = \sum \frac{(observed-expected)^2}{expected}$.

CA is developed by several authors independently (though mainly the work of Braak & Prentice (1988) and Hill (1973) is credited) and has several different names, for example *contingency table analysis*, *reciprocal averaging*, *weighted averaging*, *dual scaling* or *homogeneity analysis*. Fundamentally, CA calculates the weighted averages

of the species as defined by the χ^2-distance described below, where the weights are, with community composition data, specifically species abundances (Braak & Prentice, 1988; P. Legendre & Legendre, 2012). This has the advantage that absent species in the samples have exactly zero weight and therefore do not result in higher similarities between the samples. It also means that species abundances contribute to the calculated distances, but only to a lesser degree because they are calculated relative to the average abundances and therefore do not influence the results as much as with the Euclidean-based ordination methods (PCA/RDA). This has the consequence that low abundant species may have an unduly high influence on the results, because the abundances of common species contribute less to the calculated distance compared to low abundant species, which are often numerous (P. Legendre & Gallagher, 2001). With CA, and CCA, it can therefore be relevant to give less weight to the low abundant species if needed.

The eigenvalues of the axes in CA (and CCA) are not equivalent to those in PCA/RDA and should not be interpreted as *variance* but as correlation coefficients, which reflects the degree of correspondence between the samples and species (Buttigieg & Ramette, 2014). Canonical Correspondence Analysis (CCA) will not be described further as it is simply the constrained version of CA in the exact same way as RDA is the constrained version of PCA, where a linear combination of the species abundance matrix and the explanatory variables are calculated as an additional step in the procedure, otherwise CA and CCA are the same. Fundamental to CA and CCA is the χ^2-distance, which is calculated by the following formula:

The χ^2-distance

$$D_{\chi^2}(x_1, x_2) = \sqrt{y_{++}} \sqrt{\sum_{j=1}^{p} \frac{1}{y_{+j}} \left(\frac{y_{1j}}{y_{1+}} - \frac{y_{2j}}{y_{2+}} \right)^2} \tag{2.2}$$

where $Y = [y_{ij}]$ is a species abundance table of the size $(n \times p)$ with sites (rows) $i = [1...n]$ and species (columns) $j = [1...p]$, with row sums y_{i+}, column sums y_{+j} and total sum y_{++}. (P. Legendre & Gallagher, 2001)

Aims

By using primarily the ordination methods described in Chapter 2, the following key questions about the AS of 32 Danish wastewater treatment plants (WWTPs) will be investigated. Different types of ordination methods and distance measures will be used to describe the general differences in the microbial community composition of the AS of the WWTPs:

- Are there significant differences between Danish wastewater treatment plants with respect to the microbial community composition of the activated sludge?
- Are there microorganisms that are generally abundant in most or all treatment plants or microorganisms that are characteristic of individual treatment plants?
- Does the way in which the treatment plants are designed seem to influence the structure of the microbial communities?
- Are the microbial community composition stable over time?

3. Materials and Methods

The experimental procedures performed will only be covered briefly here. For detailed protocols refer to the 3 protocols "DNA extraction from activated sludge", "MiSeq Sequencing of Amplicons" and "16S rRNA V1-3 Amplicon Preparation v1.2" at `http://midasfieldguide.org/en/protocols/` or the supplementary data from (S. J. McIlroy et al., 2015).

3.1 Sampling

The activated sludge (AS) from 32 Danish wastewater treatment plants (WWTPs) were sampled at a 1 meter depth below the surface. Some WWTPs were sampled 2 times a year (summer, winter), while others 4 times a year (summer, fall, winter, and spring), in the period 2006 to 2015. The samples were kept on ice as much as possible and stored at -80ºC.

3.2 DNA extraction

The DNA was extracted based on the manufacturer's instructions of the FastDNA™ 2 mL SPIN Kit for Soil (MP Biomedicals, USA), optimised for DNA extraction from activated sludge by using increased bead beating with 4x40s at 6 m/s.

3.3 Library preparation, purification and pooling

Polymerase Chain Reaction (PCR) was used to amplify the V1-3 variable regions of the 16S rRNA gene using the barcoded forward (F) and reverse (R) primers listed below:

- 27F: AGAGTTTGATCCTGGCTCAG
- 534R: ATTACCGCGGCTGCTGG

PCR was performed with the thermo cycler settings: Initial denaturation at 95 °C for 2 min, 30 cycles of 95°C for 20s, 56°C for 30s, 72°C for 60s and final elongation at 72°C for 5 min. All PCR reactions were run in duplicates and pooled afterwards. The amplicon libraries were purified using the Agencourt® AMpure XP bead kit (Beckmann Coulter, USA) and the DNA concentration and quality was assured afterwards using a Quant-iT™ HS DNA Assay (Thermo Fisher Scientific) and Tapestation 2200 with D1K ScreenTapes (Agilent), respectively. Based on the measured library DNA concentrations the samples were pooled in equimolar concentrations before sequencing.

3.4 DNA sequencing and bioinformatics

The libraries were then sequenced using a MiSeq (Illumina, USA) with a PhiX control library. The libraries were subsampled to 50000 raw reads, and low quality reads were removed using Trimmomatic v0.32. Using USEARCH9 (Edgar, 2013), the sequence pairs were merged, PhiX sequences were filtered, unique sequences were identified, classified and clustered into Operational Taxonomic Units (OTUs, $\geq$ 97% sequence identity). The OTUs were then infered taxonomic identification based on the MiDAS taxonomic database (S. J. McIlroy et al., 2015), which is a manually curated version of the SILVA database (Quast et al., 2013) suited for activated sludge microorganisms. The result is a sample-by-OTU table which contains the approximate read abundances of each OTU in each sample alongside their taxonomy.

3.5 Data processing and analysis

The raw sample-by-OTU table was analysed using the R statistical language (Ihaka & Gentleman, 1996) and the RStudio IDE (`https://www.rstudio.org`). I wrote an R-function (500+ lines of code) based on the vegan and ggplot2 R-packages for ordination of the data called ord_mep() (ordination of microbial ecology profiles) specifically for the purpose of this report, available at `https://github.com/KasperSkytte/ord_mep`. The function is inspired by the function amp_ordinate() from the R-package "ampvis" by Mads Albertsen (Albertsen, Karst, Ziegler, Kirkegaard, & Nielsen, 2015), extended with support for all ordination methods mentioned in Chapter 2, data transformations, plotting options, and more.

The OTUs have been filtered so that only those with a read abundance larger than 0.1% (across *all* samples) in at least one sample are being analysed. If not otherwise noted, ordination of the data has been done by first applying the Hellinger transformation to account for incorrect similarities due to double-zeros and to also downweight low abundant species (Buttigieg & Ramette, 2014; P. Legendre & Gallagher, 2001):

The Hellinger transformation

$$D_{Hellinger}(x_1, x_2) = \sqrt{\sum_{j=1}^{p} \left(\sqrt{\frac{y_{1j}}{y_{1+}}} - \sqrt{\frac{y_{2j}}{y_{2+}}} \right)^2} \tag{3.1}$$

where $Y = [y_{ij}]$ is a species abundance table of the size $(n \times p)$ with sites (rows) $i = [1...n]$, species (columns) $j = [1...p]$ and row sums y_{i+}. (P. Legendre & Gallagher, 2001)

4. Exploring the Microbial Communities of the WWTPs

In this chapter the WWTPs will be described using mainly exploratory/unconstrained ordination methods to get an overview of the differences between the WWTPs with respect to their microbial communities. In Chapter 5 these differences will be explained using also the explanatory/constrained ordination methods based on information about the WWTPs and how they are designed. The differences are represented by the distances between sample points if not noted otherwise. The points are colored by a unique color for each WWTP, but because there are 32 different WWTPs it can be difficult to distinguish between the colors and the corresponding name of the WWTPs are therefore written at the approximate center of all the sample points from the particular WWTP. The same colors listed in the legend in Figure 4.1 will be used in subsequent ordination plots in the chapter. Furthermore, the eigenvalues of the axes plotted are indicated by the axis titles as a percentage of the total sum of eigenvalues, and scree plots can be found in Appendix B.

Before filtering, the 622 samples from 32 Wastewater Treatment Plants (WWTPs) contained a total of 21728 different OTUs. After filtering the low abundant OTUs the size of the data reduced remarkably with a total of 2366 different OTUs in all the samples (mean per sample: 1078, SD: 291.7). Because this is still a very large amount of data to visualise (even when using ordination), the following plots may be better viewed in the online *bookdown* version of this report, where it is possible to zoom in the plots and hover the points. It is available at `https:`

//github.com/KasperSkytte/MasterThesis. If you encounter any problems or need help, please email me at ksan12@student.aau.dk

4.1 Overview of the differences

The 622 samples were initially analysed using Principal Components Analysis to provide a brief overview of the data (Figure 4.1). At a first glance, there seem to be many similarities between the WWTPs, as the groups of samples tend to overlap. However, it is possible to identify global clusters of WWTPs, which seem to be similar at least partly. The largest dissimilarities observed in an ordination plot are points positioned diagonally of one another (when there is large variation on both axes), which is the case with for example the Ribe and Bjergmarken WWTPs. Now, imagine a line from the Ribe label towards the Bjergmarken label. Two clusters of similar WWTPs can then be identified by separating the line with a perpendicular line roughly in the direction of Fornaes-Haderslev, where the WWTPs on either side of this perpendicular line can be considered two different clusters. Of course, this is a rough clustering of the WWTPs, but along the Ribe-Bjergmarken diagonal seems to be the greatest variation between the WWTPs. Large variation between the samples within individual WWTPs can also be observed, for example within Bjergmarken, whose samples cover a broad area of the top group of WWTPs. This is the case with many of the WWTPs and these differences are mostly evident on the first axis. It is worth noting that the eigenvalues of the two axes plotted are nearly identical (10.1% vs 9.2%), which further highlights that the differences between the samples are large within the individual WWTPs. If there would have been a clear difference between the WWTPs, they would have been positioned horisontally with few overlaps, the first axis would have had a much greater eigenvalue than the second axis, and the within-group variation would have been evident mostly on the second axis.

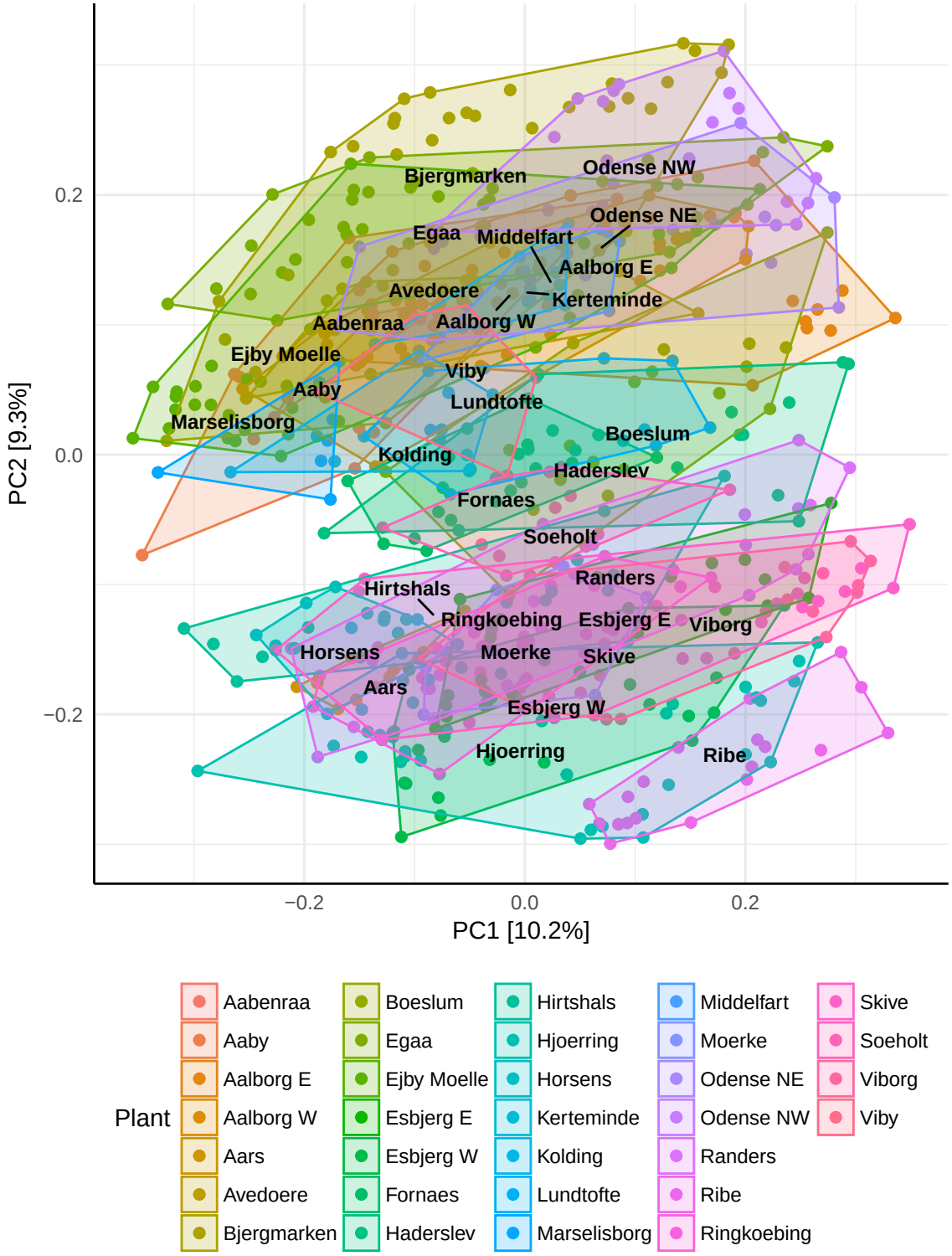


Figure 4.1: Principal Components Analysis of samples from the 32 WWTPs. Each WWTP has been assigned a unique color as indicated by the legend and labels have been positioned approximately at the center of the points.

With PCA the read abundances of the OTUs (aka the weights) contribute considerably to the distances between the samples (Buttigieg & Ramette, 2014). To represent the differences between the WWTPs where the OTU abundances have less of an impact on the distances, the widely used Bray-Curtis Dissimilarity index (BCD) is appropriate. With this measure the abundances have less of an impact because the abundance of an OTU is relativised to the total abundance of the OTU in the two samples being compared. As BCD is a semi-metric (does not satisfy the triangle inequality property), Principal Coordinates Analysis (PCoA) has to be used and the result can be seen in Figure 4.2. The relative positions of the sample points on the axes are not always in the same orientation between different ordination methods and reversing the first axis (mirroring the plot vertically) reveals relative positions similar to those in the PCA (Figure 4.1, see Figure A.4 in Appendix A for a procrustes comparison). Again, the groups are overlapping and are not well separated on the first axis, but there is a slightly clearer separation of the two (top and bottom) groups observed with PCA (Figure 4.1). The fact that the read abundances contribute less to the distances when using the BCD index and the result is similar to that of PCA indicates that abundances are not the primary cause of the differences and that the WWTPs may have many OTUs in common. The eigenvalues of the axes are not ideal, however, and using non-Metric Multidimensional Scaling with the BCD index had a bad stress value of 0.247 (see Appendix A, Figure A.2), which confirms that all the variation in the data cannot be fully represented in two dimensions.

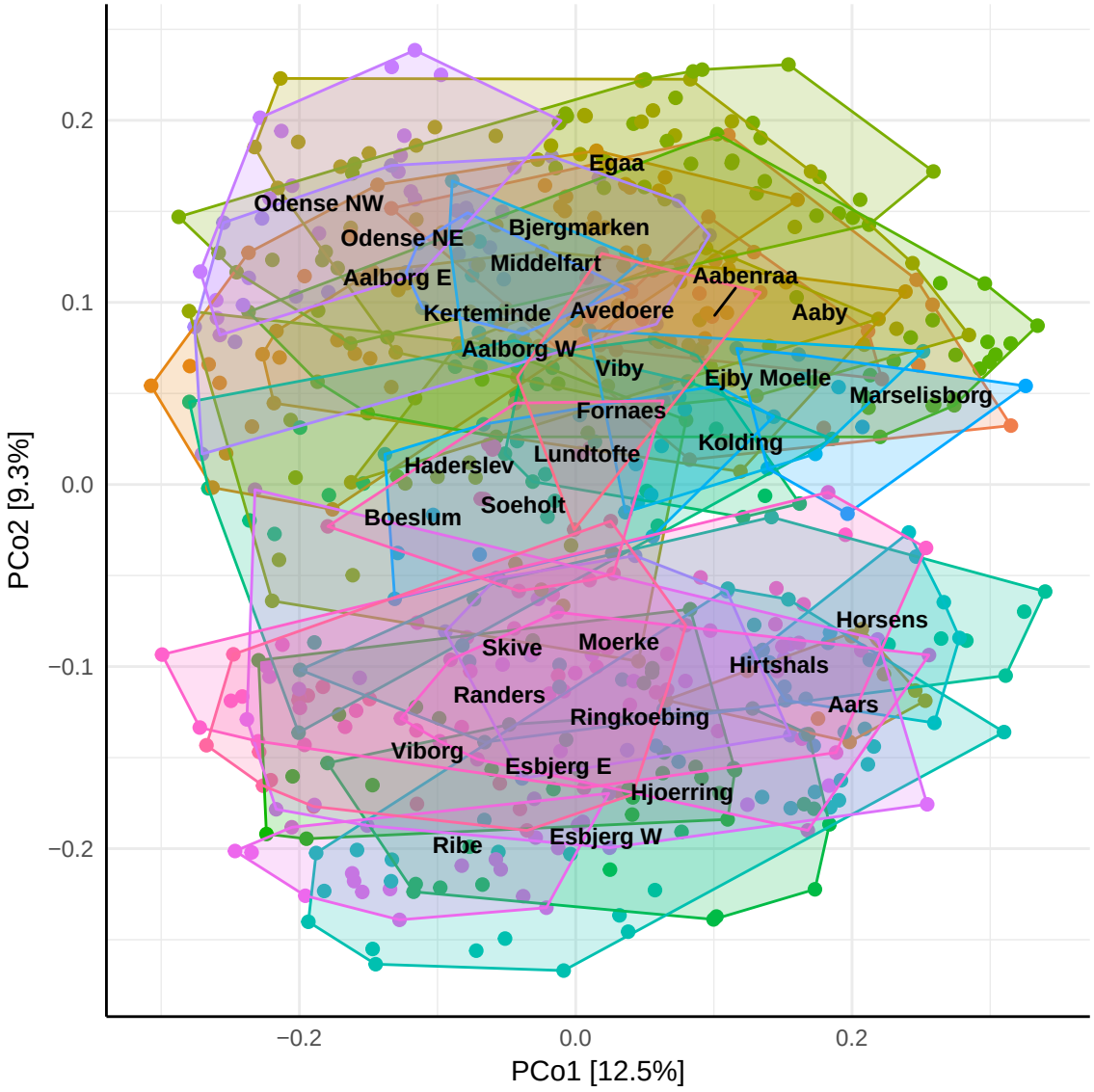


Figure 4.2: Principal Coordinates Analysis of samples from the 32 WWTPs using the Bray-Curtis Dissimilarity (BCD) index, with no Hellinger transformation. Each WWTP has been assigned a unique color as indicated by the legend in Figure 4.1 and labels have been positioned approximately at the center of the points.

As mentioned in Chapter 2, microorganisms are most likely present when a set of optimal environmental conditions are met at the sampling site resulting in a unimodal abundance distribution across samples. The ecological differences between the WWTPs are therefore expected to be reflected by both unique bacteria and to a lesser extend abundances of shared bacteria. To compare the WWTPs more in terms of their distribution of OTUs, the Pearson χ^2-statistic used in Canonical Correspondence Analysis (CCA) is more appropriate than the measures of PCA and PCoA (with BCD), as it better reveals the unique OTUs that would correspond to each WWTP. As seen in Figure 4.3, CCA shows that Esbjerg E, Esbjerg W and Ribe seem to be significantly different from the rest of the WWTPs. In general the sizes of the groups are smaller, more distinct and the overlaps are less prevailing. Furthermore, the axes plotted span a much larger range (roughly from -4 to 2) than those in PCA and PCoA (roughly from -0.3 to 0.3). When interpreting a CCA plot it is important to note that the sample points positioned closest to the center of the plot (0,0) have the highest *probabilty* of containing the most common OTUs across all samples and the samples closer to the edges of the plot have a higher probability of containing unique OTUs. This means that the Esbjerg E+W and Ribe WWTPs must either contain several unique OTUs which the rest of the WWTPs are highly unlikely to contain, or the opposite, common OTUs in the other WWTPs are of low abundance in these 3 WWTPs (this will be investigated in the following Chapter 4.2). Except for these 3 WWTPs, the overall groupings of WWTPs observed with PCA and PCoA are also evident with CCA. Considering only the relative positions of the text labels, the differences between the WWTPs seem to be relatively similar to that observed with PCA and PCoA, but now on the primary axis. It can be difficult to see in the figure, but other than the Esbjerg E+W and Ribe WWTPs, there are a few additional WWTPs that are (almost) not overlapping with other WWTPs, namely Lundtofte, Marselisborg and Moerke, indicating that their distribution of OTUs are slightly different from the rest of the WWTPs. They are closer to the center, however, indicating that the differences are most likely due to differences in abundances of common OTUs and not due to unique OTUs.

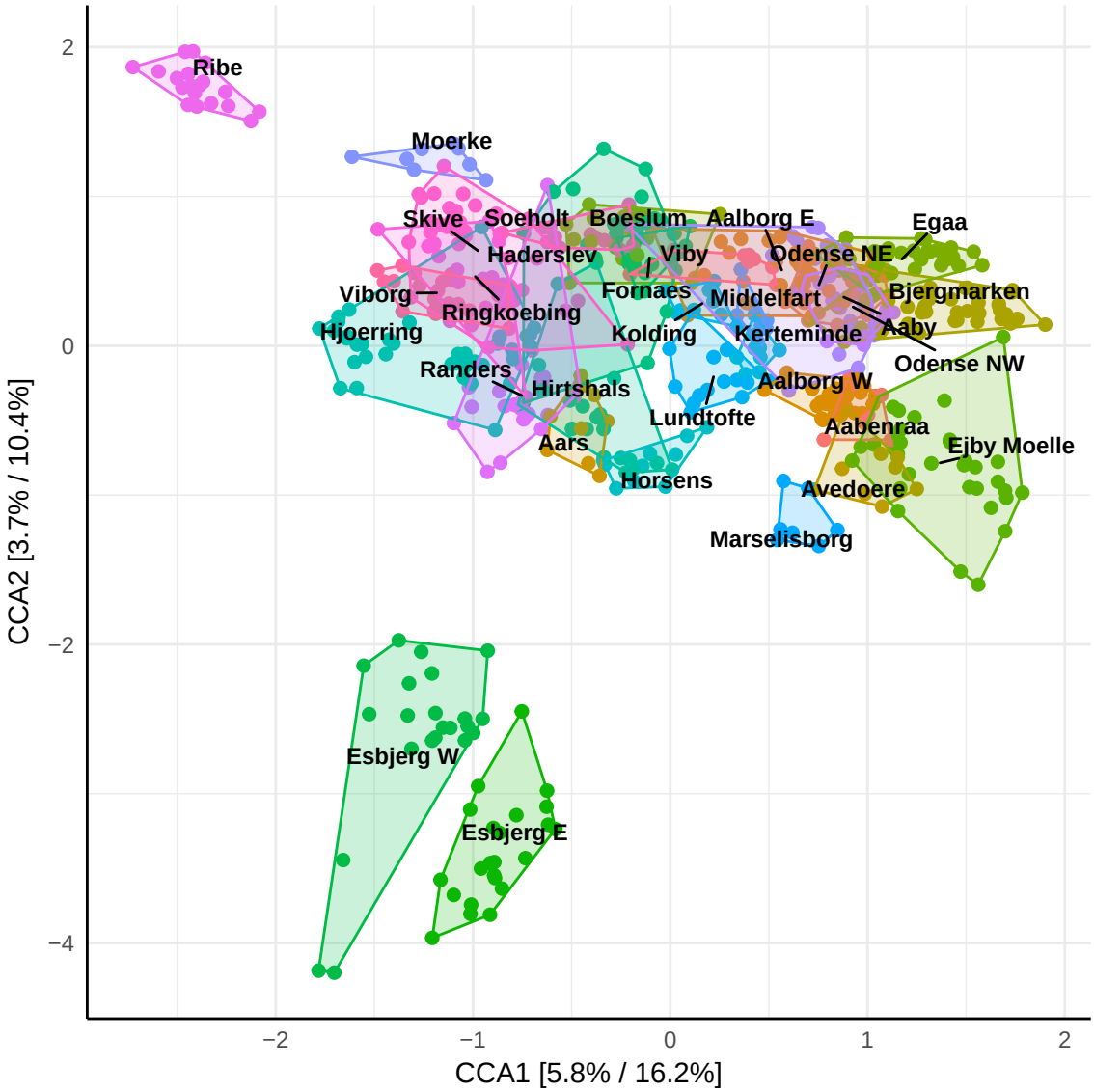


Figure 4.3: Canonical Correspondence Analysis of samples from the 32 WWTPs constrained to the WWTP where samples were taken. Each WWTP has been assigned a unique color as indicated by the legend in Figure 4.1 and labels have been positioned approximately at the center of the points. The percentages indicated on the axis titles are (left): the eigenvalue of the axis relative to the total sum of eigenvalues and (right): the eigenvalue relative to the total sum of only the constrained eigenvalues.

Again, the eigenvalues of the axes are low, but this is expected since the WWTPs must have many OTUs in common considering the fact that there are an average of 1078 different OTUs in each sample, and only 2366 different OTUs in all the 622 samples.

4.2 How does the microbial community composition describe the WWTPs?

Describing the differences between the WWTPs with respect to their microbial communities is not an easy task. As the differences are the result of variation in the presences and/or abundances of 2366 different OTUs, it is impossible to provide an extensive overview while covering all aspects of the differences. The heatmap shown in Figure 4.4 is a good example of the challenge of visualising the complex microbial communities characteristic of the individual WWTPs. The most abundant OTUs are often of most interest, however. The heatmap shows an overview of the 40 most abundant genera in *all* samples. Noticably, there seem to be only a few genera in high abundance in almost all the WWTPs, namely *Tetrasphaera*, *Candidatus Microthrix*, *Trichococcus*, *Rhodobacter*, *Rhodoferax* (the top 5). These 5 genera (65 OTUs) together made up roughly 20% of the total number of reads. Specifically *Tetrasphaera* is the only genus abundant in all WWTPs while other genera are (at least nearly) absent in at least one of the WWTPs. It is also clear that there are several genera which are only abundant (>5%) in one or only a few WWTPs, for example *Gordonia* at the very bottom of the figure, which is mostly only abundant in Moerke. These somewhat 'unique' genera abundant in only one WWTP are numerous and are often only abundant in one or a few samples, usually from the same WWTP. Generally, the remaining 324 of the 364 identified genera which are *not* shown in Figure 4.4 are dominant in only one and occasionally a few samples, which could be simply due to differences in the influent. Depending on the ordination method used to represent the differences between the WWTPs, these unique OTUs will either have

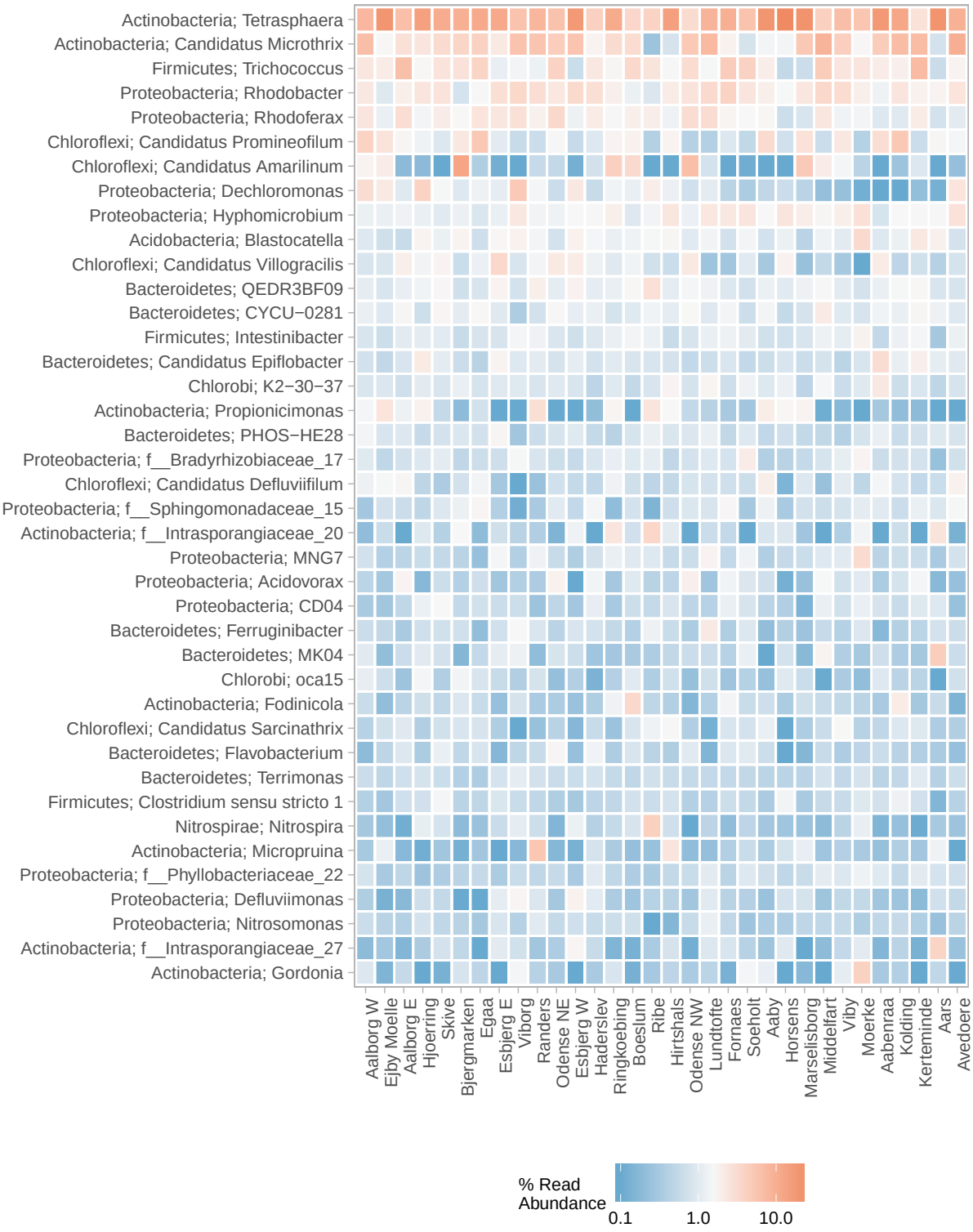


Figure 4.4: A heatmap of the 40 most abundant genera in all the samples grouped by the average in each WWTP. The WWTPs are ordered by similarity. The relative abundances are indicated by a color gradient according to the legend. The corresponding phylum of each genus is written to the left of the semi-colon: 'phylum; genus'. OTUs which have not been classified to a genus are indicated with the family name instead (the prefix 'f__')

a large impact on the distances calculated, as with Correspondence Analysis, or almost no impact at all, as with PCA, where variation in the abundances of the most abundant OTUs across *all* samples will have the largest influence on the distances. This is evident in the fact that the most abundant OTUs across all samples also seem to be among the 20 most extreme OTUs in a PCA biplot (Figure 4.5(A)). This is one of the characteristics of PCA - it shows a more quantitative-like representation of the data, because the distances are weighted directly by read abundances (Buttigieg & Ramette, 2014). To interpret the differences between samples with respect to their species in a PCA biplot, the species points are normally plotted as an arrow outwards from the center (0,0). This, as well as the plotting of sample points, has been omitted for clarity.

Because the groups of samples from the individual WWTPs do not form distinct groups in PCA (Figure 4.1), it is also expected to be reflected in the positions of the OTUs. It is clear that *Tetrasphaera* is the predominant genus and seems to be the most abundant genus across all the samples, because it is positioned in the far left, lower corner (coordinates: (-0.53,-0.17)), but it does not explain the clusters of WWTPs alone. The remaining OTUs points are positioned closer to the WWTPs, which confirms that the differences between the WWTPs can be explained partly by differences in the read abundances of common OTUs. The WWTPs near the top of the plot can then be characterised by having a higher abundance of *Trichococcus*, *Candidatus Defluviifilum*, and more, as opposed to the WWTPs near the bottom, which have a higher abundance of for example *Rhodobacter* and *Defluviimonas.* There seem to be no unique *and* highly abundant OTUs, which could have been characteristic of individual WWTPs or groups of WWTPs. This is expected, since the samples seem to share many OTUs that are generally abundant and since there are no clearly separate groups of samples observed with PCA. Noticably, the positions of the OTUs observed with PCA do not appear to be completely representative when compared to the actual read abundances of the OTUs in each WWTP (Figure 4.5(B)).

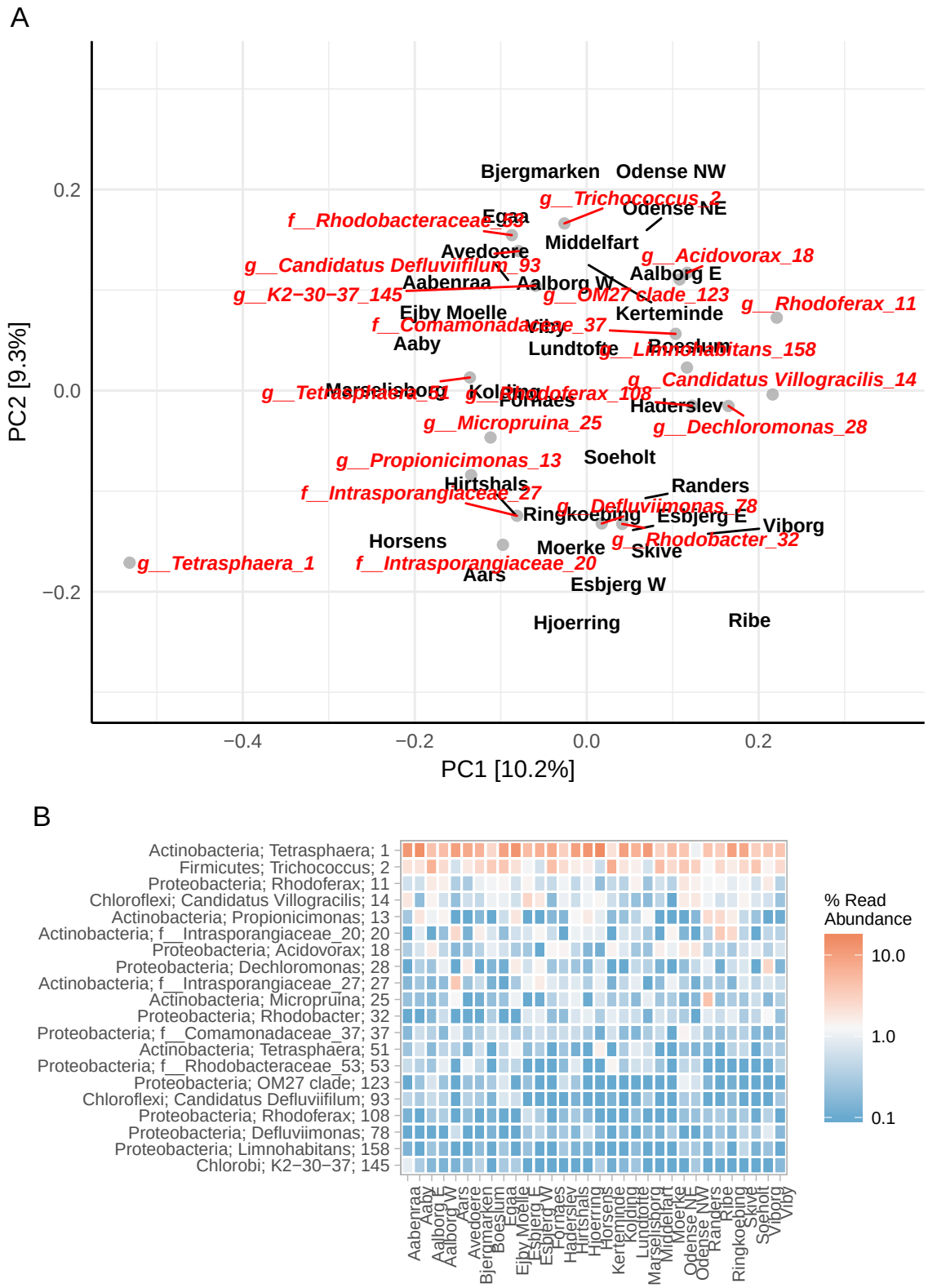


Figure 4.5: (A): Principal Components Analysis of samples from the 32 WWTPs. The names of the lowest known taxonomic rank of the 20 most extreme OTUs are shown in red text, where g__ is genus name, f__ is Family name, etc. This figure is identical to Figure 4.1, except that OTUs are plotted instead of samples. (B): Heatmap of the mean read abundances of the same 20 OTUs in each WWTP. The corresponding phylum and genus names of the OTUs (the numbers) are also indicated in the form: phylum; genus; OTU.

For example, *Rhodobacter* seems to have a larger influence on Randers than *Micropruina*, even though *Micropruina* is considerably more abundant in Randers compared to *Rhodobacter.* This phenomena is believed to be caused by the fact that only the centroids (labels) of the WWTPs are being interpreted and not all samples at once, however.

It is also clear that low abundant and unique OTUs have almost no contribution to the distances in PCA, because they are positioned very close to the center (this is only visible in the *bookdown* version of the report, there are too many labels to plot). This somewhat qualitative information can be valuable, but is lost with PCA. As mentioned, CCA reveals these OTUs clearly, even when given lower weights. The relative positions of OTU points in CCA are not to be interpreted as linear gradients of abundances as with PCA, but as *probabilities* of the OTUs to be present in the samples positioned nearby. It is impossible to show text labels of the OTUs with CCA, so again, please refer to the *bookdown* version of the report for their identity.

A CCA biplot (Figure 4.6) shows that there are numerous OTUs which seem unique to Esbjerg E+W and Ribe. It is clear that there are OTUs very close to the Ribe and Esbjerg E+W samples, but there are also OTUs positioned roughly on a gradient from the 3 individual WWTPs towards the center (0,0) of the plot. Because these OTUs are not exactly positioned among the samples from the 3 WWTPs, they may be present in low abundances in other samples from other WWTPs. Among the unique OTUs that are present in Ribe samples are for example from the genera *Aquicella* and *Haliangium*. Both of these genera are known halophilic bacteria which have previously been isolated from coastal saline waters (Fudou, Jojima, Iizuka, & Yamanaka, 2002; Valenzuela-Encinas et al., 2009). This could explain why they are present at the Ribe WWTP, because it is located near the west coast of Denmark and the influent water may have a higher degree of salinity. In the case of Esbjerg E+W, there is a much more diverse set of unique OTUs. These OTUs are not well defined as most of them have only been assigned to a family or higher taxonomic rank, which makes it difficult to explain their presence based on physiological properties. Some of the identified genera positioned the closest to Esbjerg E+W are for example

Brooklawnia[1], *Herpetosiphon*[2], *Thermovirga*[3], and *Denitratisoma*[4]. These will not be characterised further here, refer to the articles noted.

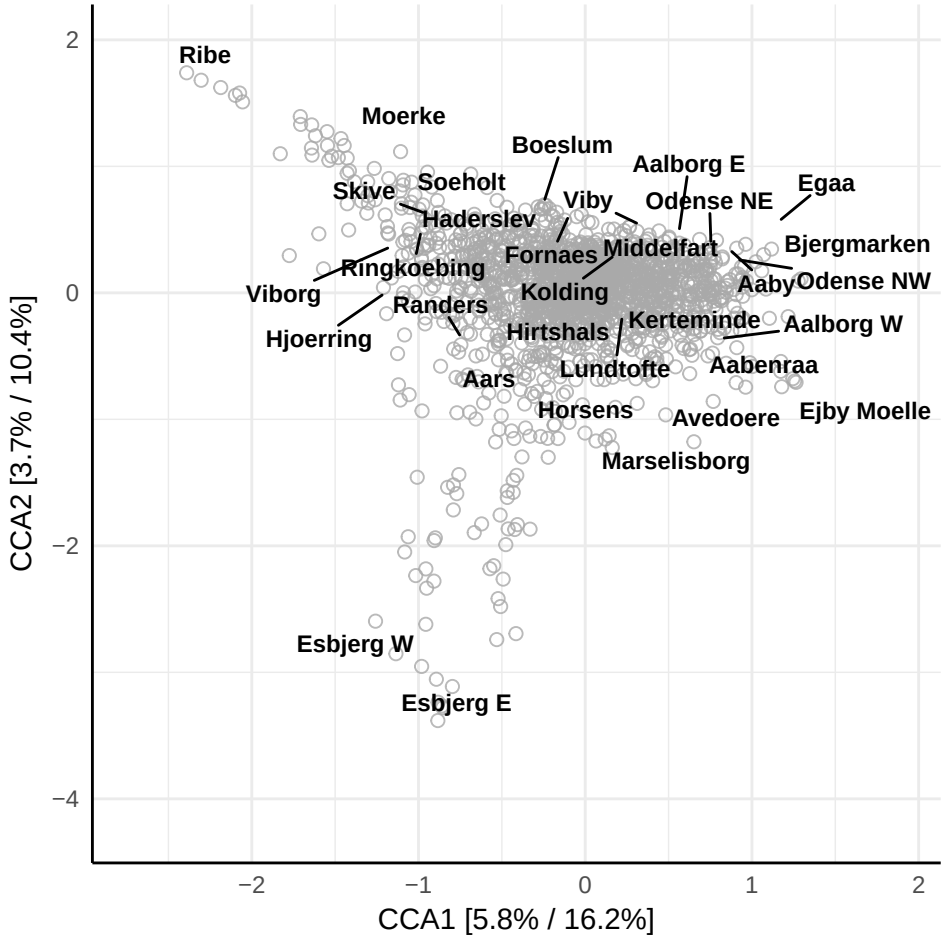


Figure 4.6: Canonical Correspondence Analysis of samples from the 32 WWTPs. This figure is identical to Figure 4.3, except that only the text labels of the WWTPs have been plotted and the OTUs are shown as grey circles (to better reveal overlaps). The percentages indicated on the axis titles are (left): the eigenvalue of the axis relative to the total sum of eigenvalues and (right): the eigenvalue relative to the total sum of only the constrained eigenvalues.

[1]Bae et al. (2006)
[2]Quinn & Skerman (1980)
[3]Dahle & Birkeland (2006)
[4]Fahrbach, Kuever, Meinke, Kämpfer, & Hollender (2006)

4.3 Concluding remarks

The 32 WWTPs generally seem to be similar with many OTUs in common, and the differences between the samples often seem to be large within individual WWTPs when using PCA and PCoA. The differences between the samples can either be thought of as qualitative differences, where a handful of unique OTUs are only present at one particular WWTP, or as quantitative differences, where the differences are primarily the result of variation in the read abundances of shared (usually also abundant) OTUs among the WWTPs. As such, there are differences in terms of both aspects and it is important to consider the ecological importance of both abundant OTUs and also the presence of unique OTUs and their abundances. Therefore it is reasonable to not only use one type of ordination and/or distance measure to represent the differences, but a combination of a few methods to reveal as many important aspects of the data as possible.

5. Explaining the Microbial Communities of the WWTPs

5.1 The influence of plant design on the microbial communities

To investigate possible causes of the differences between the WWTPs other than the influent wastewater, the influence of plant design of the WWTPs is interesting to investigate. Whether differences in plant design have an influence on the microbial communites of the activated sludge of a full scale WWTP is nearly impossible to test experimentally. This would require two identical WWTPs with the exact same influent wastewater, weather conditions, dimensions etc, where one WWTP is designed with one feature which the other WWTP does not have, and the microbial communities of each WWTP then compared. By using ordination it is possible to examine the correlation between the microbial community and environmental variables like plant design as an alternative approach. This can be done by either plotting the environmental variables onto an ordination plot creating a triplot (Figure 5.1), where the relative positions of the labels represent the probabilities of the variables to correspond to samples nearby, or it can be done by constrained/canonical ordination of the variables individually (Figure 5.2), as already described. A table of the design characteristics of each WWTP which will be investigated can be found in Appendix C.

A CCA triplot (Figure 5.1) shows that the AS samples seem to be correlated with most of the design characteristics. The four characteristics plotted are whether the WWTPs have a digester or not ("DigesterYes" and "DigesterNo"), whether they have primary setling or not ("PrimarySetlingYes", and "PrimarySetlingNo"), whether they utilise Enhanced Biological Phosphorous Removal ("DesignEBPR") or Biological Nutrient Removal ("DesignBNR"), and lastly whether there are alternating aerobic/anaerobic conditions or not ("ConfigurationAlternating" or "ConfigurationRecirculation"). The correlation of the characteristics (or factors) are significant as all four have a P-value below 0.001 with 999 random permutations, so the design of the WWTPs seems to be correlated with the composition of the microbial communities. The fact that the positions of the labels are not all exactly at the center (0,0) supports this. However, by interpreting the positions of the labels, the magnitude of the correlation seems small and sometimes misleading when interpreting the individual WWTPs. Take for example Aars (-0.5,-0.8), which is positioned very close to PrimarySetlingYes, but it does not feature it (see Appendix C). Similarly Kerteminde (0.5,0.0) is positioned very close to ConfigurationAlternating, but is expected to be positioned near ConfigurationRecirculation. There are several more examples of this kind, which suggests that there must be additional, unknown factors having a larger direct influence on the microbial communities than the ones investigated. When doing analyses involving correlation, it is thus important to remember that correlation does not imply causation, and the observed correlations can only provide a hint of what could be interesting to investigate and confirm experimentally.

Supporting the CCA triplot (Figure 5.1) with constrained ordination of some of the individual characteristics (Figure 5.2), it is again clear that the differences between the WWTPs can be partly explained by their design characteristics. It is important to note that both axes can be interpreted in Figure 5.2(D), but only the first axis in Figure 5.2(A+B+C) (due to the number of axes obtained with constrained ordination are always one less than the number of different possibilities of the particular variable).

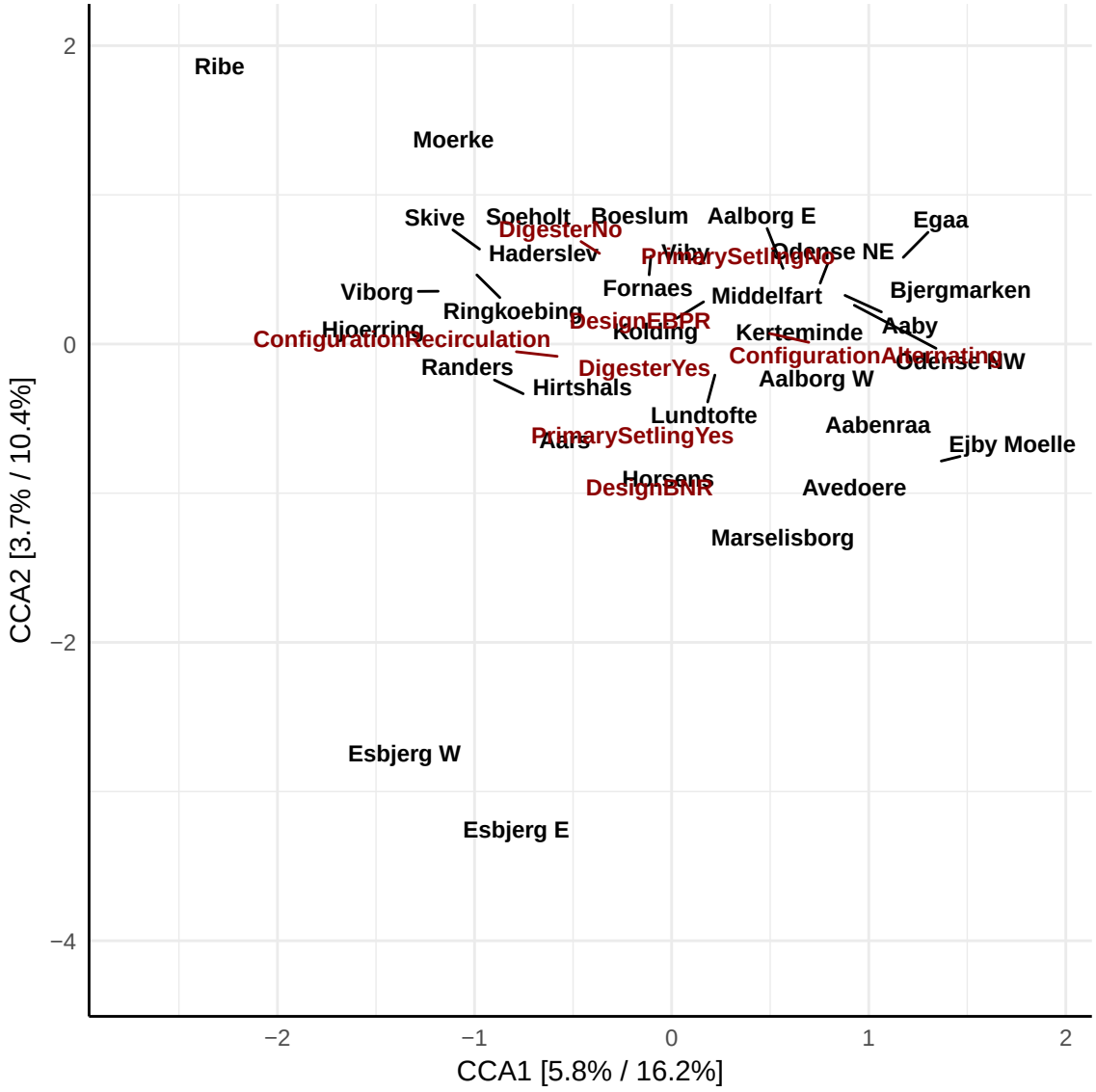


Figure 5.1: Canonical Correspondence Analysis of samples from the 32 WWTPs. This figure is identical to Figure 4.3, except that only the labels of the WWTPs have been plotted for clarity. Labels of the correlation between samples and different plant designs are plotted at their center of gravity, where their relative position represents their significance.

In general, the percentages of the first axes are low in all four plots (between 1.8% and 2.4%), which indicates that the differences are small.

Alternation vs recirculation (Figure 5.2(A)) seems to be the characteristic with the least influence on the OTUs as they are not clearly separated on the first axis and the two groups both overlap the OTUs on the first axis. In Figure 5.2(B+C) some of the OTUs are positioned on a vertical line directly through the center of the two groups, which is a clear indication that these OTUs are representative of the particular group. Any OTUs between these two vertical lines are shared among both groups. This is most evident in Figure 5.2(B), where there are many OTUs positioned on a wide line at x=-0.5 corresponding to EBPR, and similarly also OTUs on a line at x=2 corresponding to BNR. It is worth noting that only 9 WWTPs utilise BNR, while the remaining 23 WWTPs utilise EBPR. This explains why most of the OTU points near the center (0,0) are closer to the EBPR group and not BNR. This is also evident in Figure 5.1, where the DesignEBPR label is close to (0,0) and DesignBNR at (0,-1). The microbial communities of WWTPs that are utilising Enhanced Biological Phosphorous Removal (EBPR) have been studied extensively in recent years and several key bacteria are confirmed to play an important role in the process, for example the Phosphate Accumulating Organisms (PAO) *Tetrasphaera* and *Accumulibacter* (Kristiansen et al., 2013; Mino, Loosdrecht, & Heijnen, 1998; P. H. Nielsen et al., 2010; R. Seviour & Nielsen, 2010). In Figure 5.2(B) these bacteria are positioned to the left near the EBPR label (unfortunately it is not possible to search in the plots), which confirms that the WWTPs utilising EBPR may indeed have a different microbial community composition. With primary setling (Figure 5.2(C)) the positions of the OTUs are not as clearly separated on the first axis, and there do not seem to be particularly unique OTUs corresponding to each group, a large majority of the OTUs seem to be shared. Lastly, the amount of industrial wastewater in the influent wastewater (Figure 5.2(D)) also seems to be influencing the microbial communities of the WWTPs.

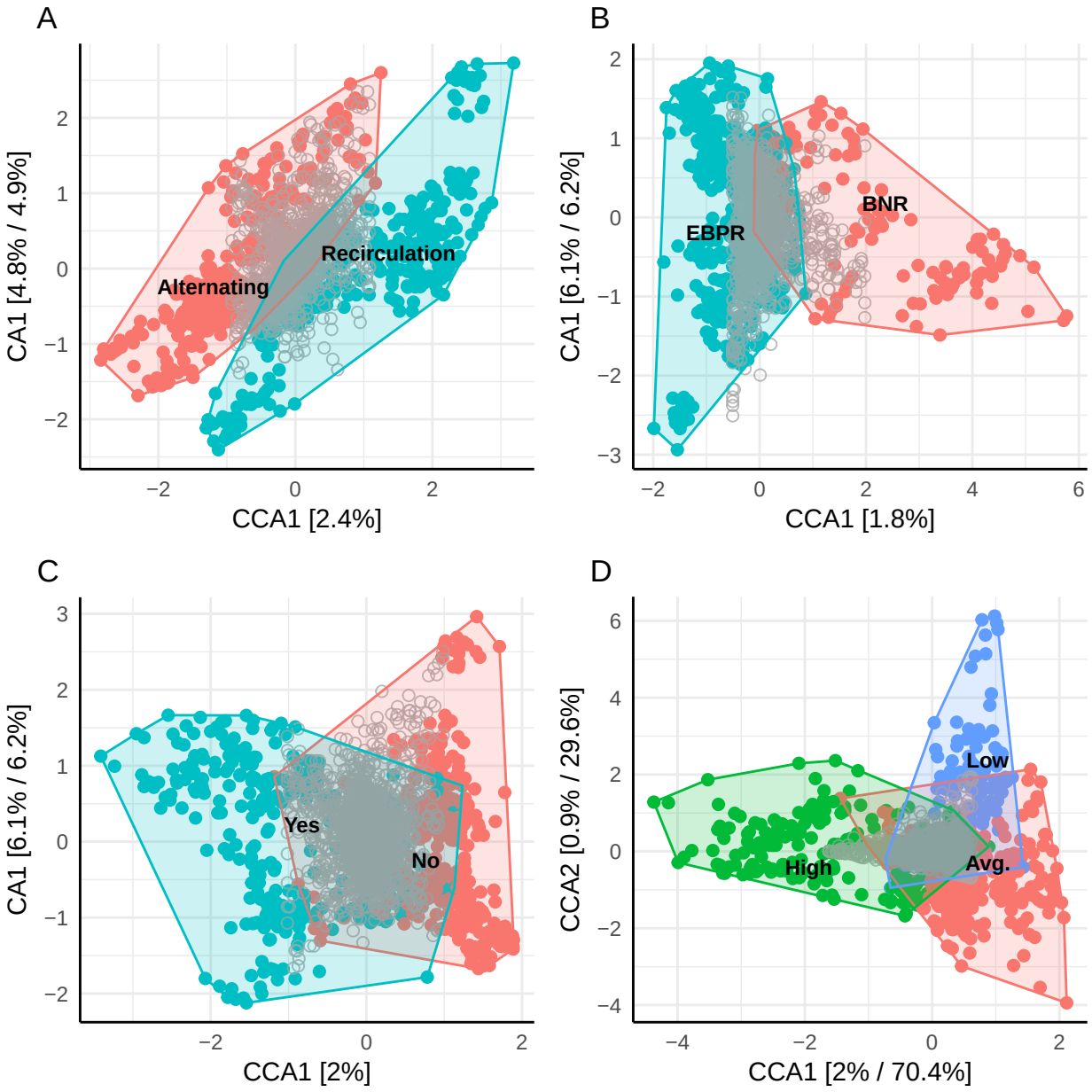


Figure 5.2: Canonical Correspondence Analysis with the constraints (A): Alternation vs Recirculation, (B): Enhanced Biological Phosphorous Removal (EBPR) vs Biological Nutrient Removal (BNR), (C): Primary Setling, and (D): the amount of industrial wastewater content, where Low is 5%<10%, Avg. is 10%<35% and High is 35%<100%. The points represent samples colored by the particular constraint, OTUs are marked as grey circles.

As expected, all three groups (Low, Avg., and High) have many OTUs in common, but there are many OTUs which seem to correspond to a high industrial wastewater content, and likewise several OTUs that correspond only to the WWTPs having a low amount of industrial wastewater content, however not as many. The influence of the amount of industrial wastewater in the influent have not been extensively studied previously, however, but one study by Ibarbalz, Figuerola, & Erijman (2013) used constrained ordination to analyse the differences between industrial WWTPs, and concluded that their microbial communities were clearly distinct from those of municipal WWTPs.

5.2 General differences related to sampling time

The microbial communities of the 32 WWTPs appear to be changing slightly over time as revealed by CCA (Figure 5.3(A)). It is clear that the year 2014 seems strikingly different from the rest of the years, but this is believed to be the result of a different way of preparing the samples from this particular year. When taking a closer look into which OTUs seemed unique to the year 2014, many of the OTUs have only been classified to a kingdom (bacteria) or a class level, which indicates that something went wrong during one of the experimental steps, either during DNA extraction, PCR, or DNA sequencing. It would simply seem impossible that all 32 WWTPs in one year differed from other years in the exact same way, when the WWTPs have nothing in common except that they are all located in Denmark. It is important to mention that the samples from 2014 have not been filtered in the analyses presented in this report, however doing so did not result in any noticable differences. A CCA plot without the 2014 samples can be found in Figure A.10 in Appendix A. Besides 2014, there seem to be a small difference between each year. It is not easily distinguishable exactly which OTUs could be representative of each year, it seems more like the differences are generally caused by differences in the OTU read abundances.

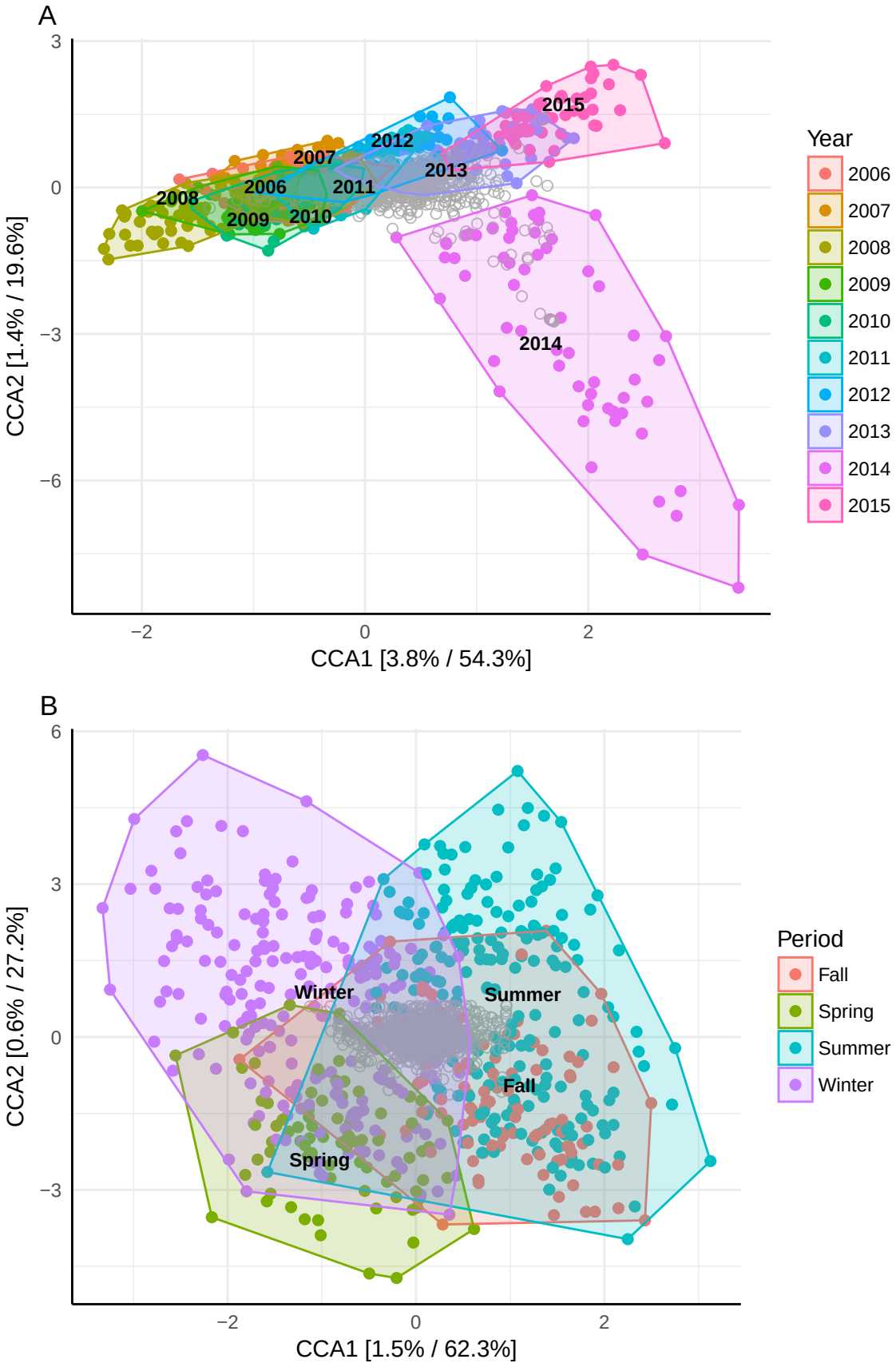


Figure 5.3: Canonical Correspondence Analysis constrained to which year (A) or seasonal period (B) the samples were taken. The labels of each year or seasonal period have been marked at the approximate centroid of the corresponding samples.

Furthermore, it is not possible to conclude whether the observed clusters per year are actually the result of the samples being taken within the same year, or whether they simply are the result of identical experimental processing of the samples from each year. However, the differences are small, as the eigenvalue percentage of the main axis is relatively small (3.8%). It is expected that samples taken at different seasonal periods of the year (summer/winter) would be different as a result of differences in temperature, which is perhaps the environmental factor with the largest impact on the thriving of microbes (J Farrell & A Rose, 1967). As seen in Figure 5.3(B), this seems to generally be the case in all the WWTPs, however there do not seem to be distinct clusters of samples per period, they are largely overlapping.

6. General Discussion

When analysing microbial communities using 16S rRNA Amplicon Sequencing, it is important to note that the obtained OTU abundances are nothing more than relative measures of the *read abundances* of the extracted DNA sequences in the samples relative to the total amount of reads. They do not exactly represent the amount of cells present of a particular microorganism due to various biases. These biases are introduced by for example multiple gene copy numbers (GCN) within each cell, primer specificity, and DNA extraction (M. C. van Loosdrecht et al., 2016). A study in 2004 found that the 76 bacterial genomes sequenced contained between 1-15 operon copies and up to 7 are commonly found. Only 40% of the genomes had 1-2 (Acinas, Marcelino, Klepac-Ceraj, & Polz, 2004). This introduces a significant bias which can be further enhanced by multiple genome copies. Additionally, the primers used to target the 16S rRNA gene are designed based on a consensus sequence matching as many bacteria as possible, and therefore whole taxonomic groups can be overlooked. To provide a more complete picture of the microbial communities, perhaps all 9 variable regions of the 16S rRNA should be sequenced, or alternatively sequencing full-length ribosomal small subunits (SSUs) using the Oxford Nanopore instead (S. M. Karst et al., 2016). Together, these biases may have had the consequence that the true ecological differences between the WWTPs are not exactly as presented here. Because the abundances do not seem to have a large influence on the differences when comparing the results of PCA (Figure 4.1) with PCoA using the Bray-Curtis dissimilarity measure (Figure 4.2), biases in abundances are considered unimportant, however. If neccessary, tools exist to correct the abundances based on known GCNs of specific bacteria (Angly et al.,

2014).

The fact that the WWTPs have many OTUs in common is supported by the work of A. M. Saunders, Albertsen, Vollertsen, & Nielsen (2016), where they found that 13 Danish WWTPs (of which many are the same as those investigated in this report) contained 63 genera making up 68% of the total reads using 16S rRNA amplicon sequencing. In one of the WWTPs, they further related these genera to what was present in the influent wastewater, concluding that 10% of the total reads were due to immigration from the influent wastewater, and that these OTUs participated little to the wastewater treatment based on their growth rate in the AS. This helps elucidate the influence of the bacteria present in the influent wastewater on the microbial community of the activated sludge. The exact influence of the influent wastewater to the WWTPs sampled here have not been investigated and is a subject for further study.

There have not previously been performed many studies of this kind using different ordination methods to describe so many samples at once, and in particular from the AS of different WWTPs. The largest similar study to date seems to be of GenBank records of 202 globally distributed environmental samples from different soil and water environments (C. A. Lozupone, Hamady, Kelley, & Knight, 2007). The authors used PCoA of UniFrac distances, which incorporates phylogenetic distances (Lozupone & Knight, 2005), to identify clusters of samples based on known physical environmental factors. They found that salinity was a major environmental determinant of the microbial community compositions, more impactful than other physical factors like temperature or pH. This supports that the halophilic genera observed in the Ribe WWTP (Figure 4.6) indeed could be due to high salinity in the wastewater. In another study of 14 geographically separated WWTPs in China and US (Zhang, Shao, & Ye, 2012), they performed both Cluster Analysis using the Bray-Curtis dissimilarity measure aswell as PCoA of weighted UniFrac distances on data obtained using Roche 454 pyrosequencing of 16S rRNA amplicons. They were able to roughly identify 3 clusters of samples from the WWTPs, however only 15 samples were analysed. Interestingly, like A. M. Saunders et al. (2016), they also

found that the majority (70 to be exact) of the 744 identified genera present in all 15 samples made up a large amount of the OTU reads (63.7%), which supports that a few genera can be dominant in most WWTPs, as also observed in this study (5 out of all 364 identified genera made up ~20% of the reads). Furthermore, Zhang et al. (2012) reported that some of the most abundant genera identified in the 14 WWTPs were *Zoogloea*, *Trichococcus*, *Prosthecobacter*, and *Dechloromonas* (in decreasing order of abundance), where *Zoogloea* and *Prosthecobacter* are not observed among the 40 most abundant genera in this study (Figure 4.4). This is possibly due to a different climate or other factors, however.

The fact that the WWTPs seem similar and have many OTUs in common is further confirmed by other ordination methods than the ones presented in this report. There are numerous combinations of data transformation and filtering, ordination methods, and distance measures which can be used to analyse the microbial communities of the WWTPs, however only a handful of the most informative have been shown in this report. Several other methods showed somewhat the same patterns, and in general they all indicated that the WWTPs are similar with shared OTUs (a few additional plots can be found in Appendix A). Categorising the microbial communities of WWTPs using ordination was initially inspired by the clustering of human gut microbiomes into so-called *Enterotypes* (Arumugam et al., 2011). Here, the authors particularly used one measure only (Jensen-Shannon Divergence) to cluster the microbiomes with little reasoning noted, and the practice of *Enterotyping* has received criticism (Knights et al., 2014; Koren et al., 2013), because the clustering of microbial communities highly depends on the chosen distance measure. This is also evident in the ordination analyses in this report, where different measures revealed different results from the exact same data, which therefore highlights the importance of knowing what kind of information the particular measure is able to reveal in the data and use this as an advantage. CCA were able to reveal unique OTUs in the WWTPs, while PCA and PCoA instead highlighted differences based on OTUs generally with a high read abundance.

The eigenvalue percentages of the axes plotted are generally low in the ordination

plots, especially in CCA. This is believed to be due to the large amount of similar samples being analysed at the same time. When analysing fewer samples from only one year, 2013 (Figure A.7 in Appendix A), the overall patterns are very similar to the patterns observed when all 622 samples are analysed at once (Figure 4.3). Again, the microbial communities of the Ribe and Esbjerg E+W WWTPs seem to be different from the rest of the WWTPs, but the eigenvalue percentages of the axes are significantly higher (*all samples*: 5.8% and 3.7% - *2013 samples*: 13.9% and 9.5%). This further confirms that there are many similar OTUs between the samples, or else the eigenvalues would not be significantly lower with more samples analysed. In perspective, in the study mentioned earlier of 202 environmental samples (C. A. Lozupone et al., 2007), the axis percentages of their PCoA analyses were similarly low (between 3.1%-5.5%), which indicates that this is not unusual when analysing this many samples at once.

As a last note, it would arguably make sense to filter the OTUs that are only present in one sample to provide a completely representative picture of the microbial communities of each WWTP, as these OTUs are possibly not part of the functional "core" microbial community of the particular WWTP. As mentioned, there are many of these OTUs which are only present in one sample, and filtering them may have revealed a more ecologically meaningful representation of the WWTPs with respect to their performance.

Conclusion

By using different ordination methods it was possible to identify general differences in the microbial communities of 32 Danish WWTPs based on 16S rRNA amplicon sequencing. Principal Components Analysis (PCA) and Principal Coordinates Analysis (PCoA) of Bray-Curtis dissimilarities both revealed largely overlapping clusters of samples from the individual WWTPs, indicating considerably similar microbial communities between large groups of samples from different WWTPs. General differences between clusters of WWTPs were evident in most of the ordination methods used to analyse the samples, where groups of similar WWTPs were different from other groups of WWTPs. The differences were considered to be the result of a combination of both quantitative and qualitative differences, where both variation in the read abundances of common OTUs and the presence of unique OTUs were important to consider. The latter was evident in Canonical Correspondence Analysis (CCA), where the Ribe and Esbjerg E+W WWTPs seemed to have several unique OTUs, which were absent in the other WWTPs. The 5 most abundant genera (65 OTUs) in all samples were found to be *Tetrasphaera*, *Candidatus Microthrix*, *Trichococcus*, *Rhodobacter*, *Rhodoferax* which together made up roughly 20% of the total number of reads, and specifically *Tetrasphaera* was generally abundant in all 32 WWTPs.

The correlation between variation in the microbial communities and four different design characteristics of the WWTPs were investigated and were found to be significant ($P<0.001$). According to CCA, WWTPs with a high amount of industrial wastewater in the influent were also found to have a slightly different microbial community composition than those with a lower content. Furthermore, variation in

the microbial communities were observed over the years 2006-2015, where the year 2014 were significantly different from the other years, presumably due to different laboratory handling of the samples. Lastly, general differences in the microbial communities of all 32 WWTPs as explained by which time of the year the samples were taken, were minor with no particular bacteria corresponding to each seasonal period.

A. Supplementary plots

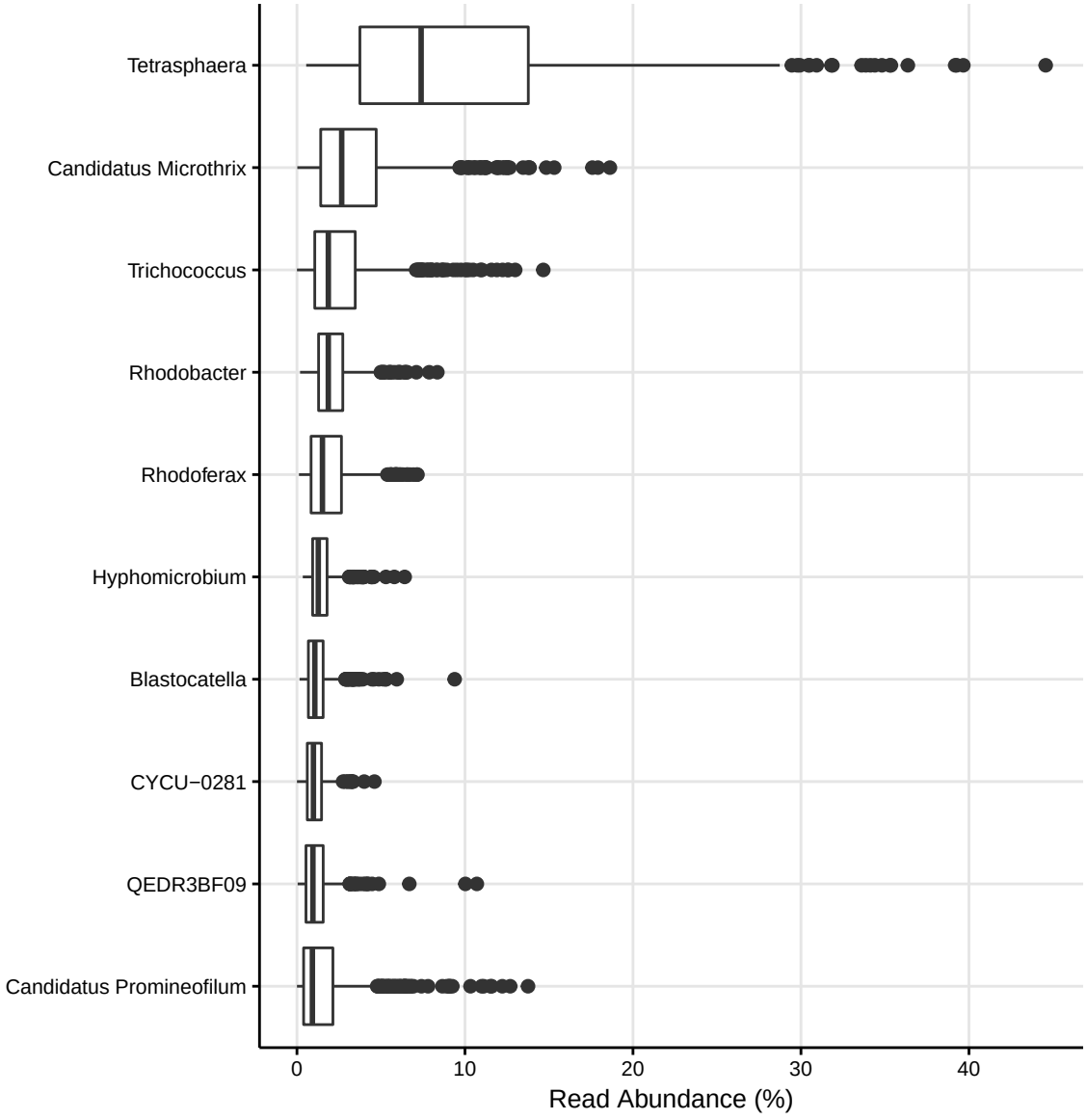


Figure A.1: Boxplot of the 10 most frequent Genera in all samples (ordered by median). No filtering nor transformation has been used.

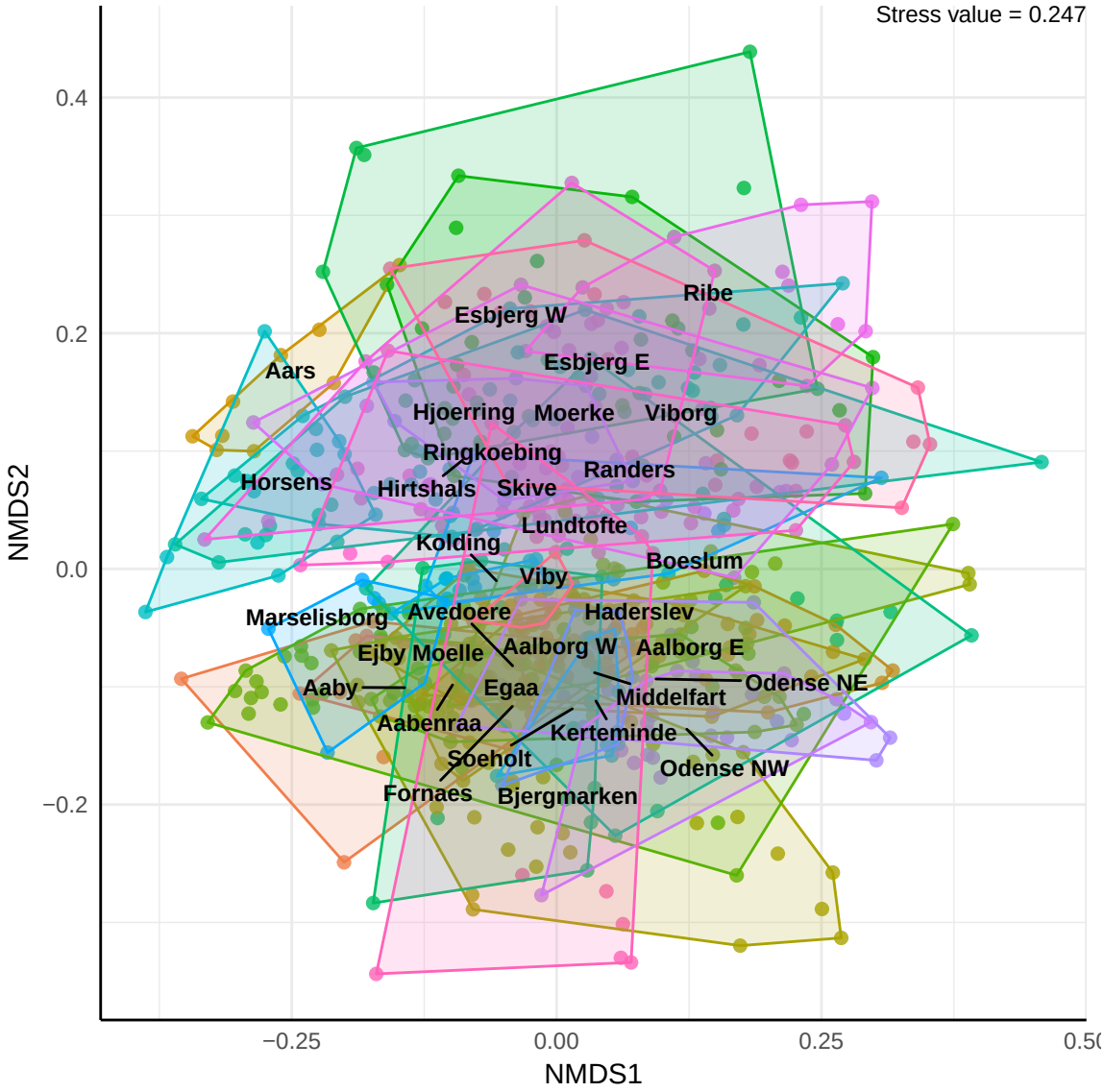


Figure A.2: non-Metric Multidimensional Scaling, Bray-Curtis Dissimilarities. No Transformation.

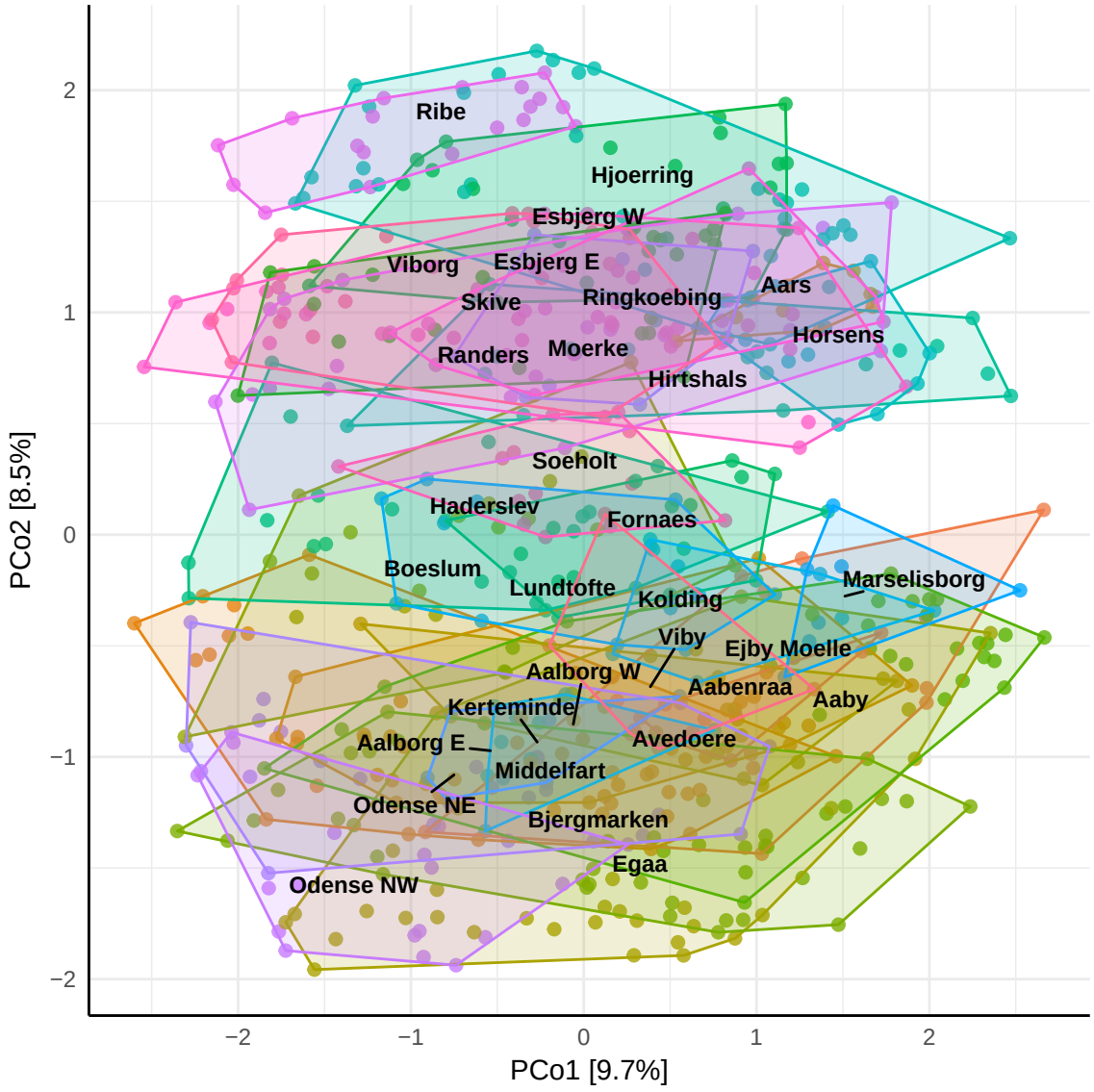


Figure A.3: Principal Coordinates Analysis using the Jensen-Shannon Divergence dissimilarity measure. No transformation.

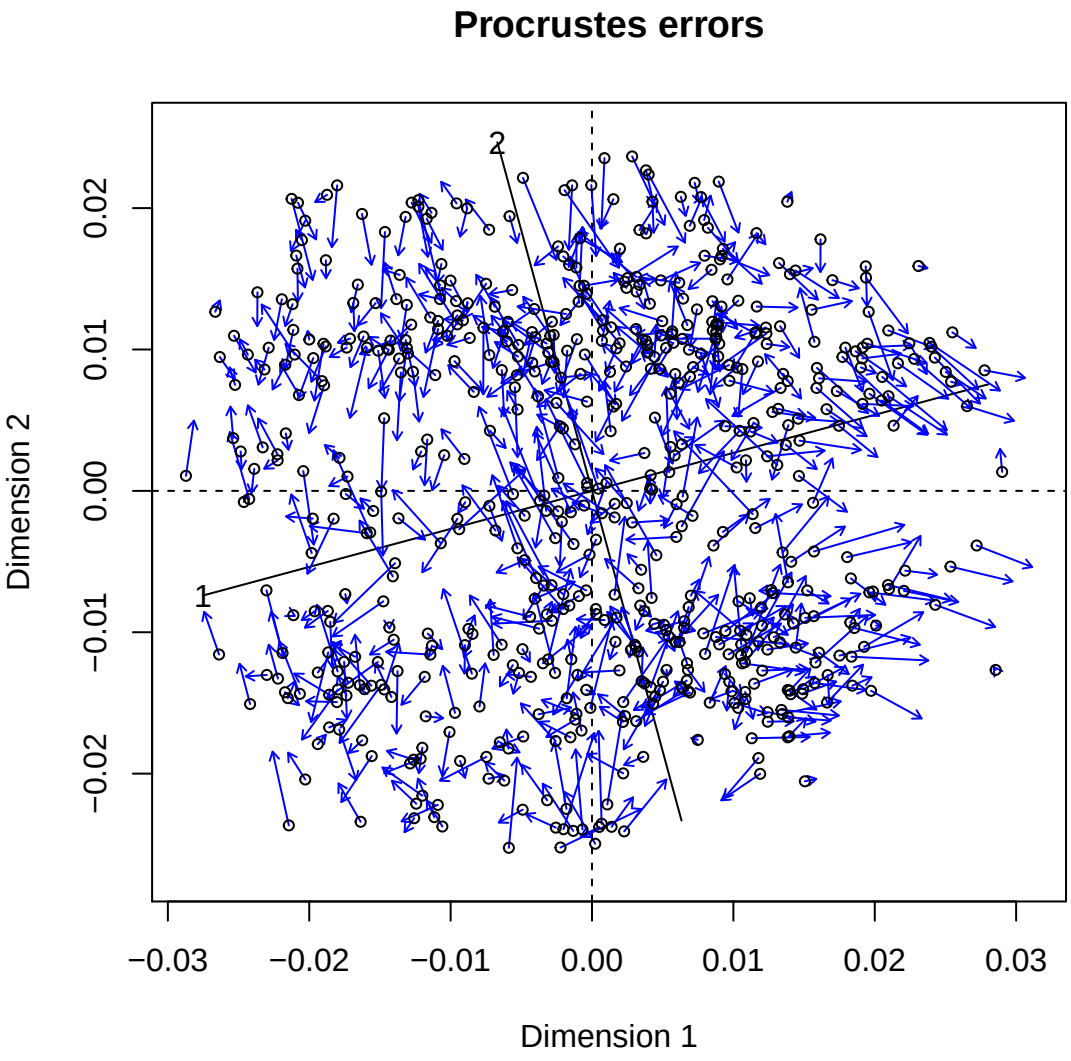


Figure A.4: Procrustes analysis of the ordination results from Figure 4.1 and Figure 4.2. Procrustes sum of squares: 0.80. The blue arrows indicate the differences in the relative positions of the same samples in both figures.

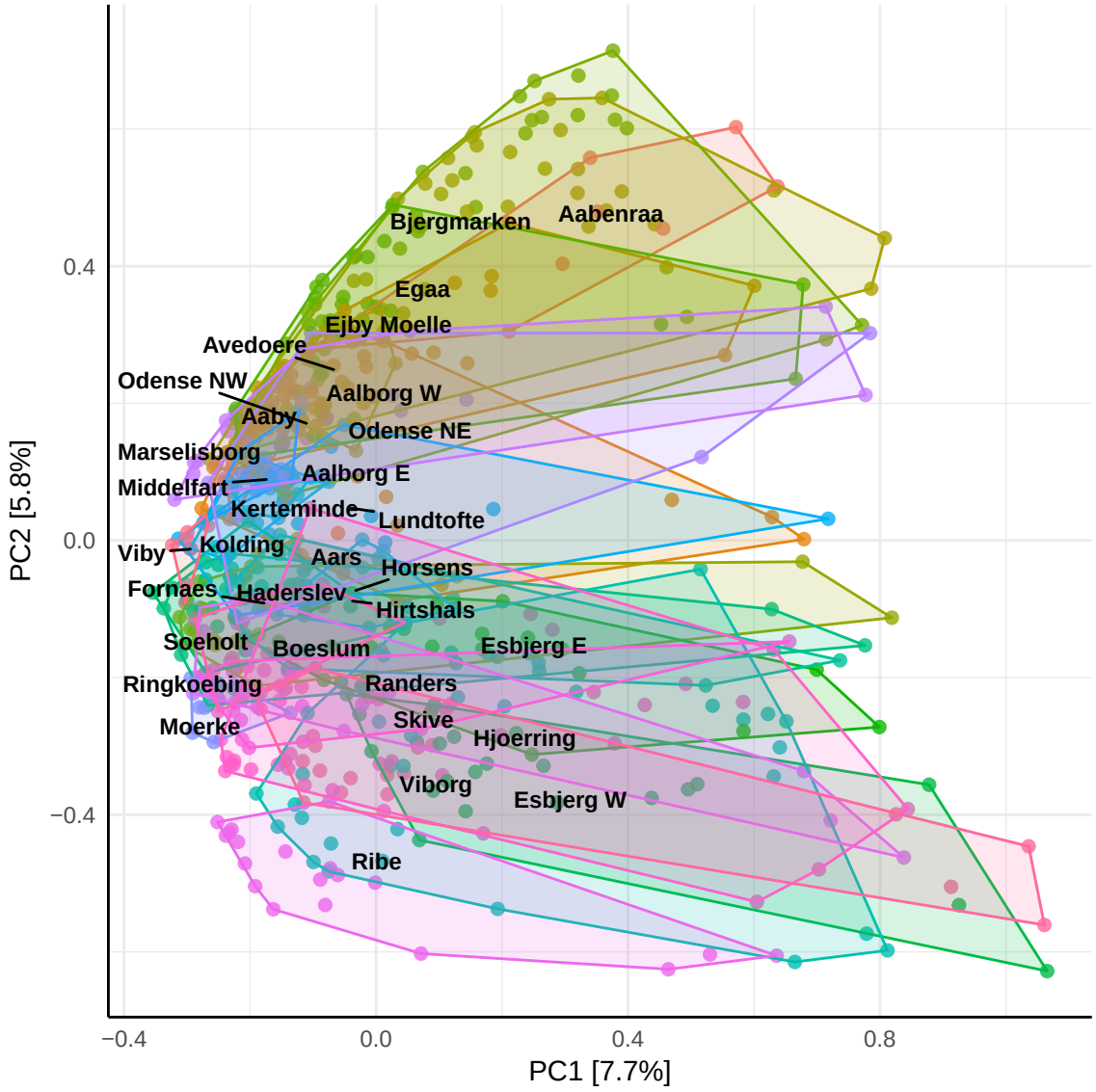


Figure A.5: Principal Components Analysis where all positive abundances in the raw count data have been set to 1. Species with abundances <0.1% are removed beforehand.

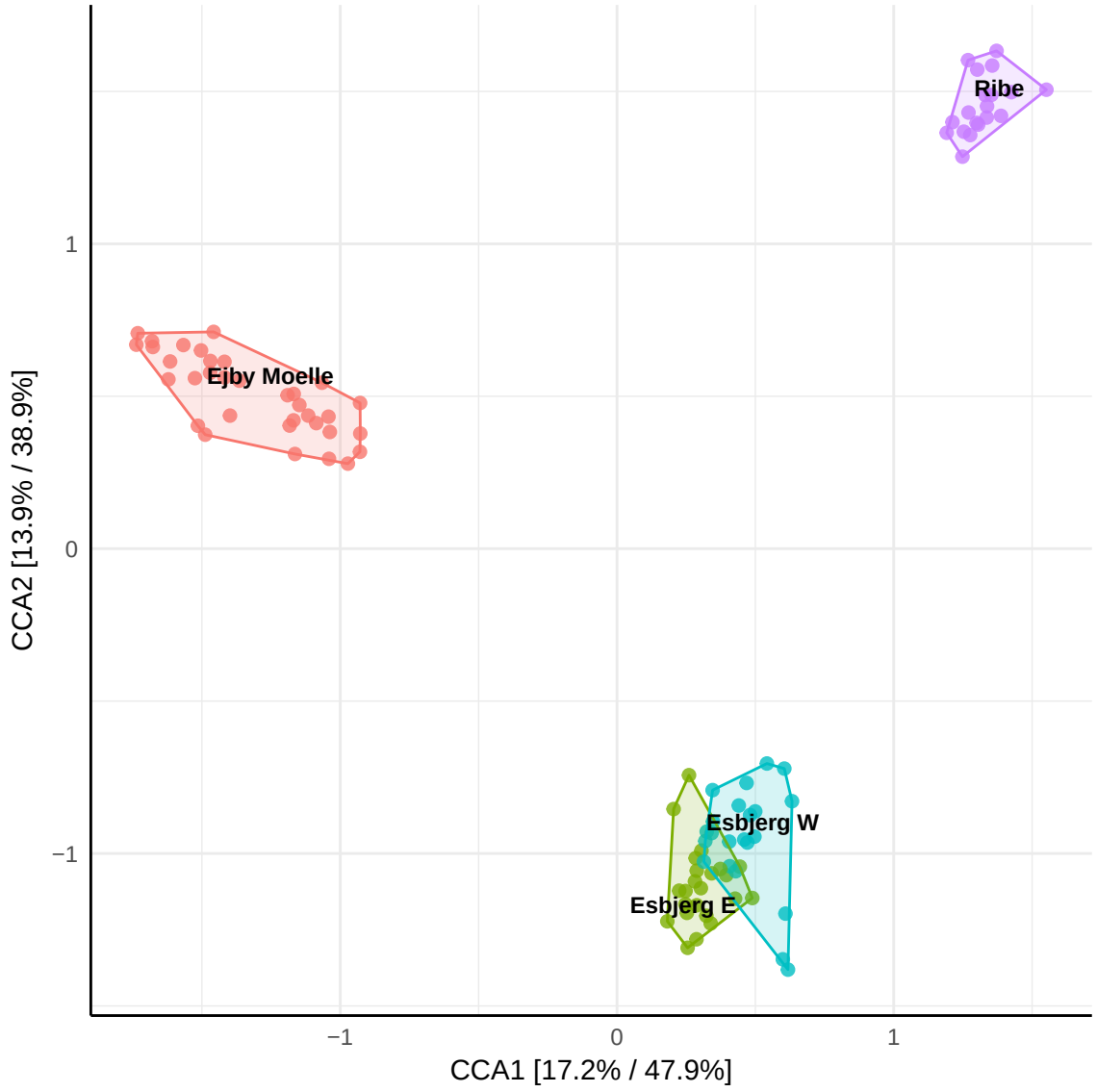


Figure A.6: Canonical Correspondence Analysis of samples from 4 of the WWTPs. Hellinger transformed.

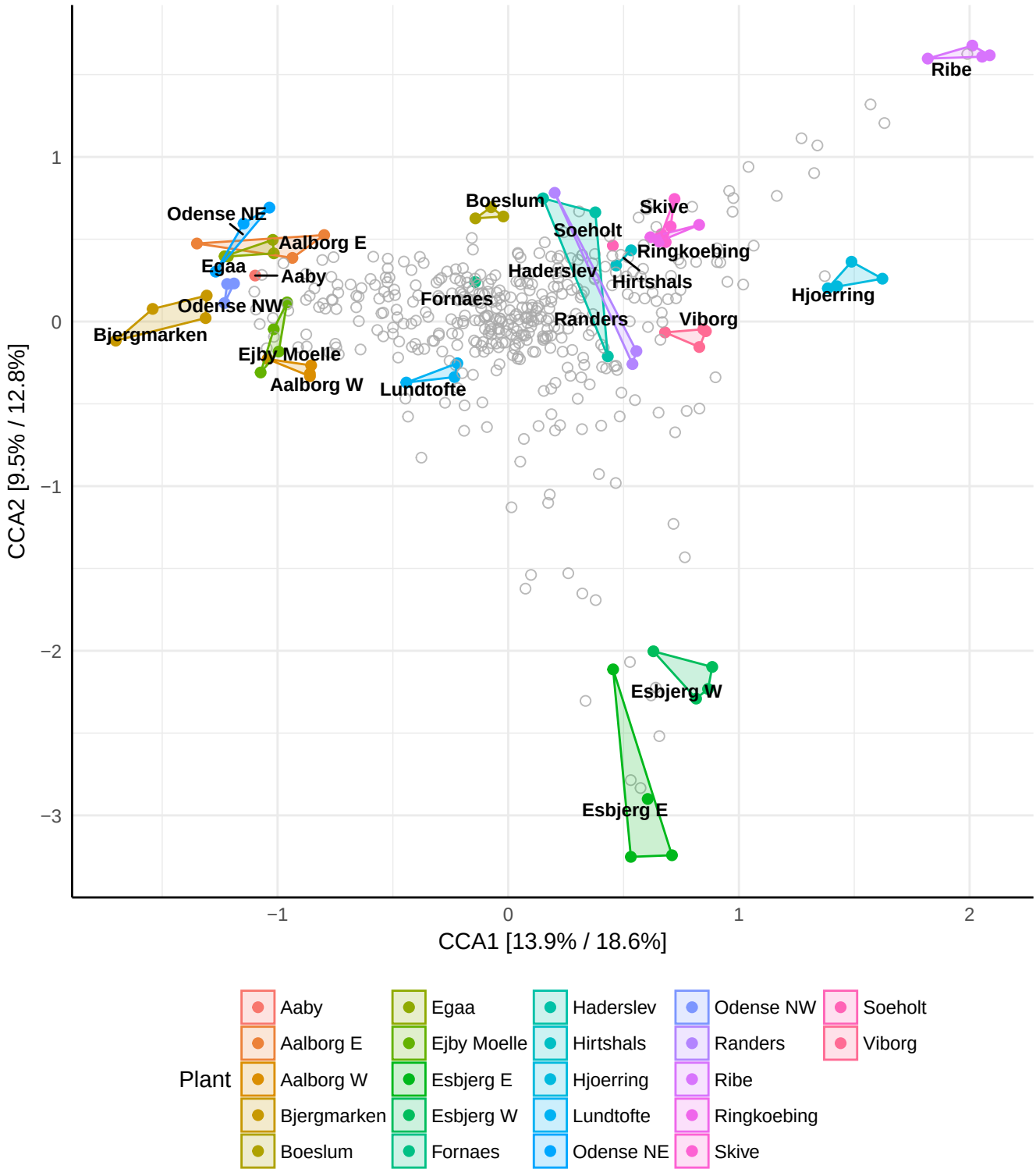


Figure A.7: Canonical Correspondence Analysis of all samples from 2013. Hellinger transformed.

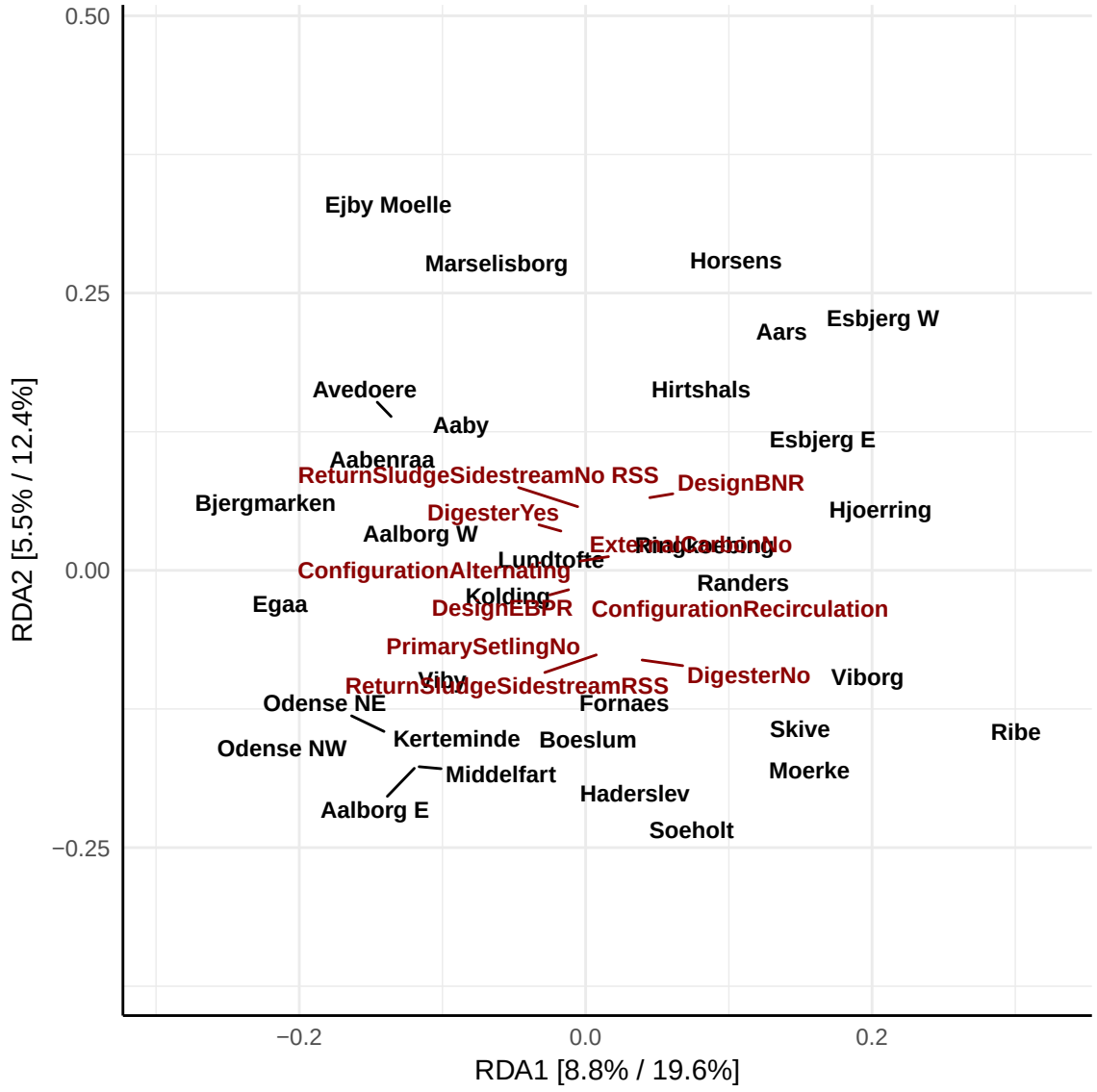


Figure A.8: Redundancy Analysis biplot with all possible plant design parameters plotted in red. Constrained to WWTP.

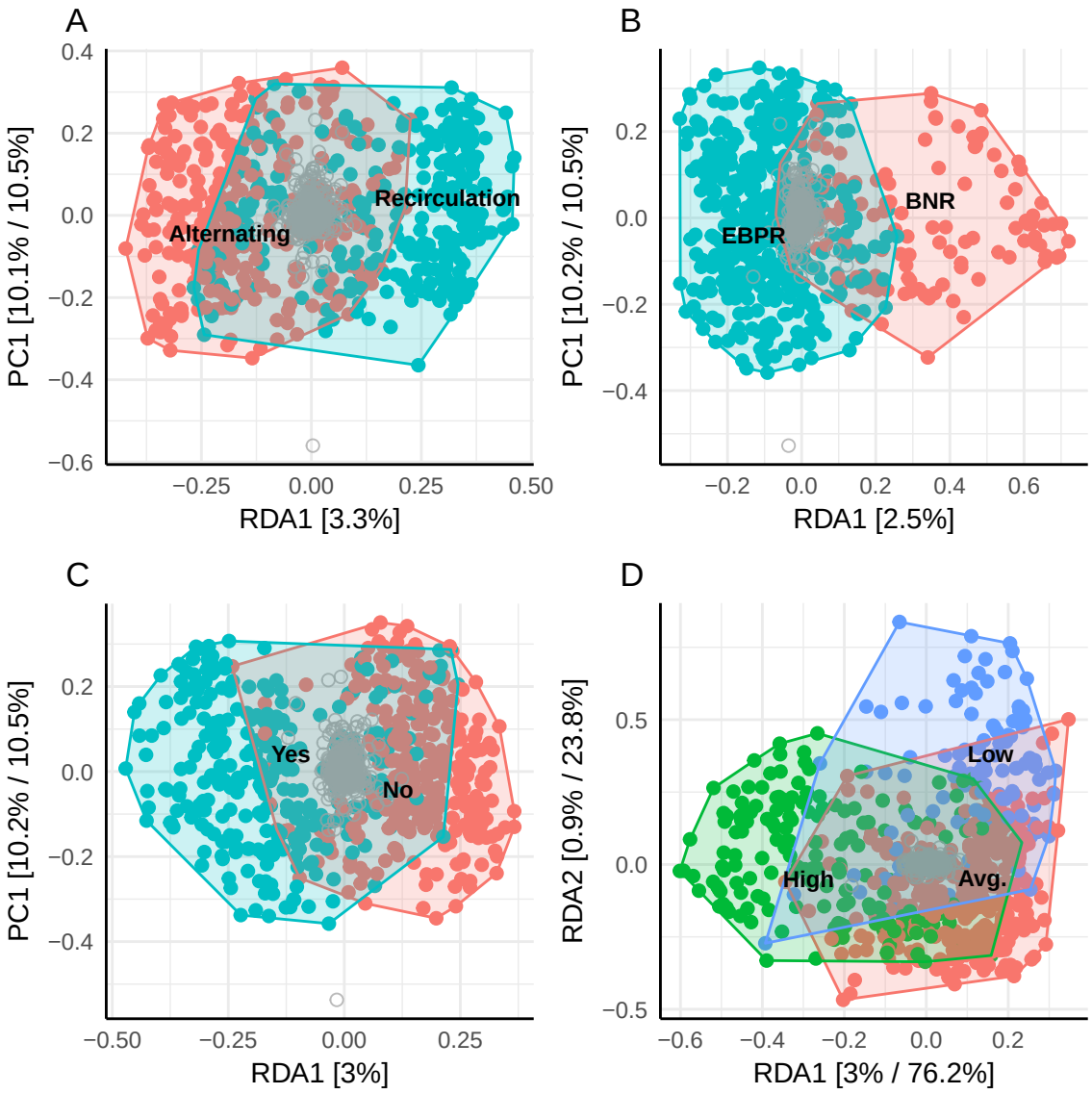


Figure A.9: Redundancy Analysis with the constraints (A): Alternation vs Recirculation, (B): Enhanced Biological Phosphorous Removal (EBPR) vs Biological Nutrient Removal (BNR), (C): Primary Setling, and (D): the amount of industrial wastewater, where Low is 5%<10%, Avg. is 10%<35% and High is 35%<100%.

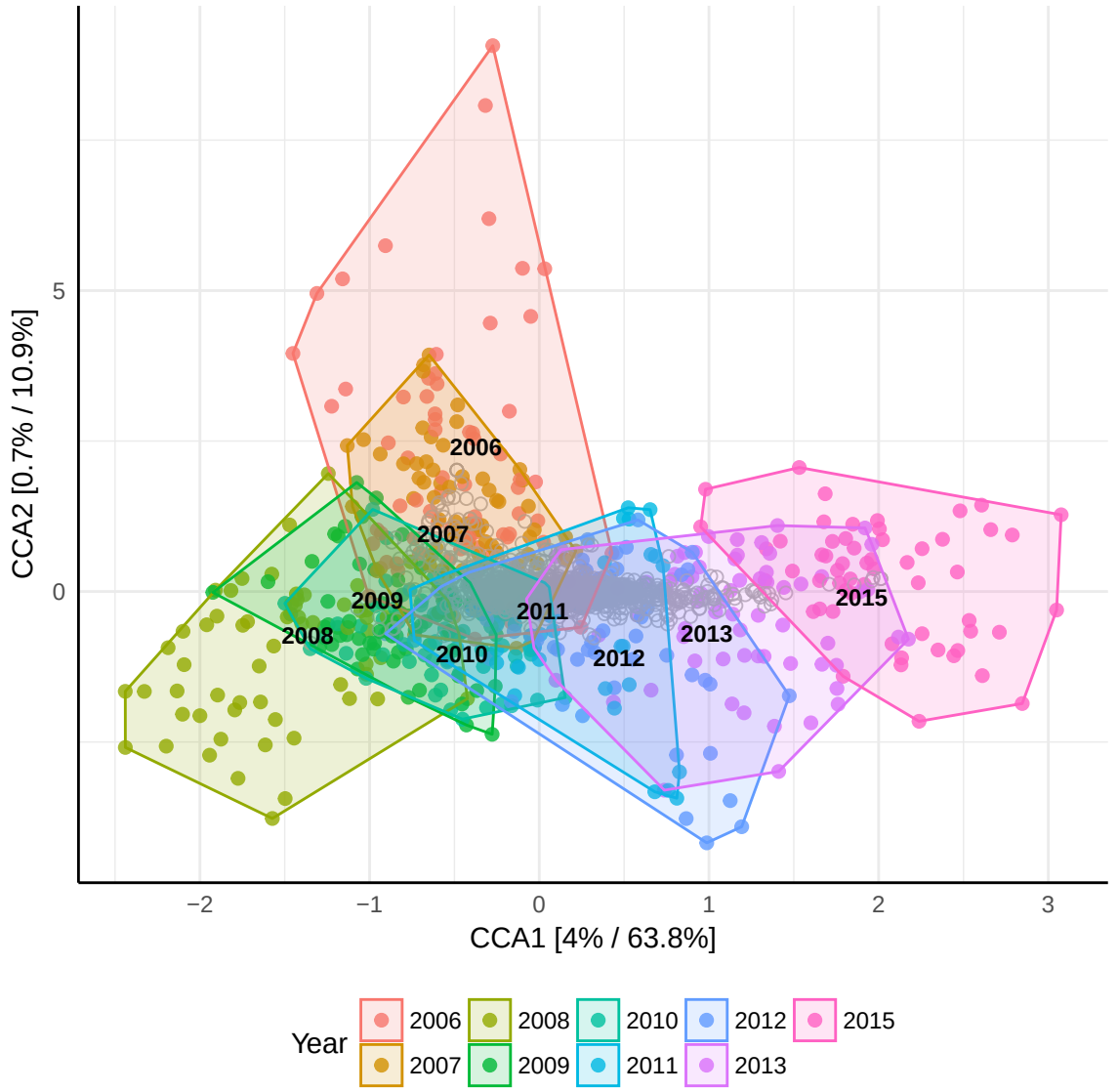
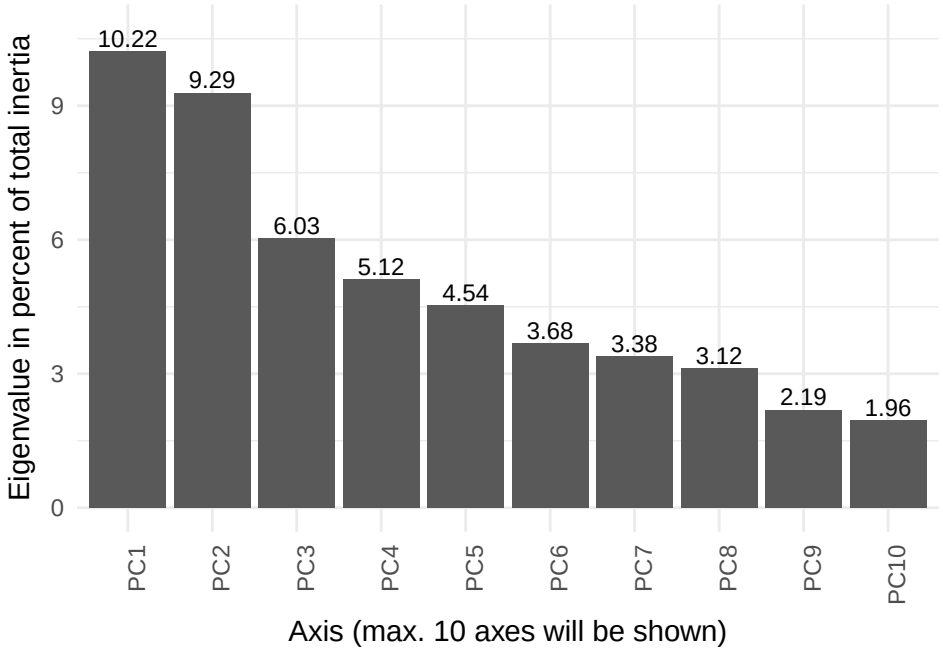


Figure A.10: Canonical Correspondence Analysis constrained to which year the samples were taken, except the year 2014. The different years are marked with a label at the centroid of the corresponding samples and colored according to the legend.

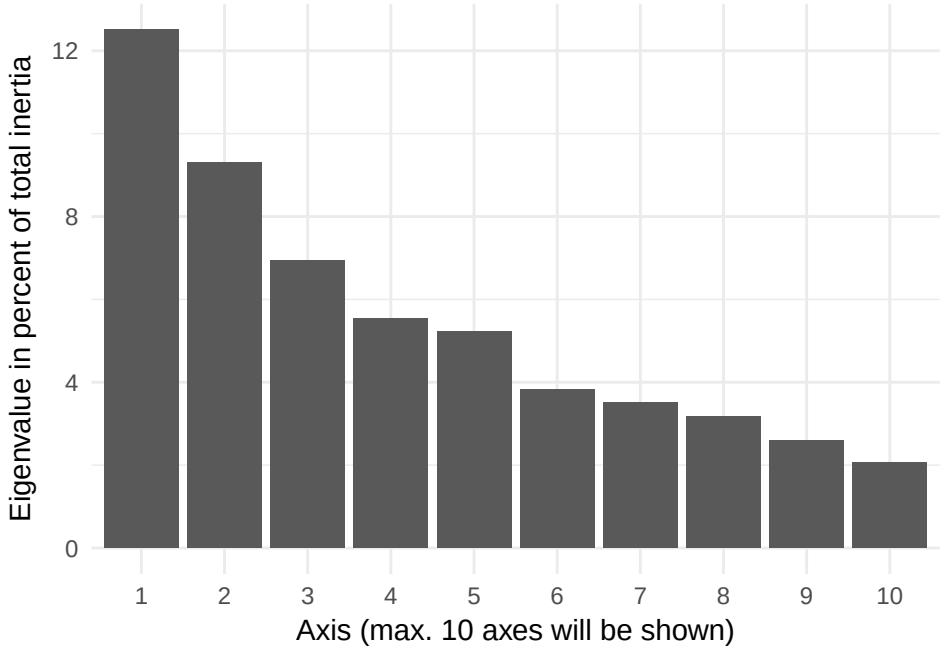
B. Scree plots

Scree plots for all plots in the two results chapters are shown here.

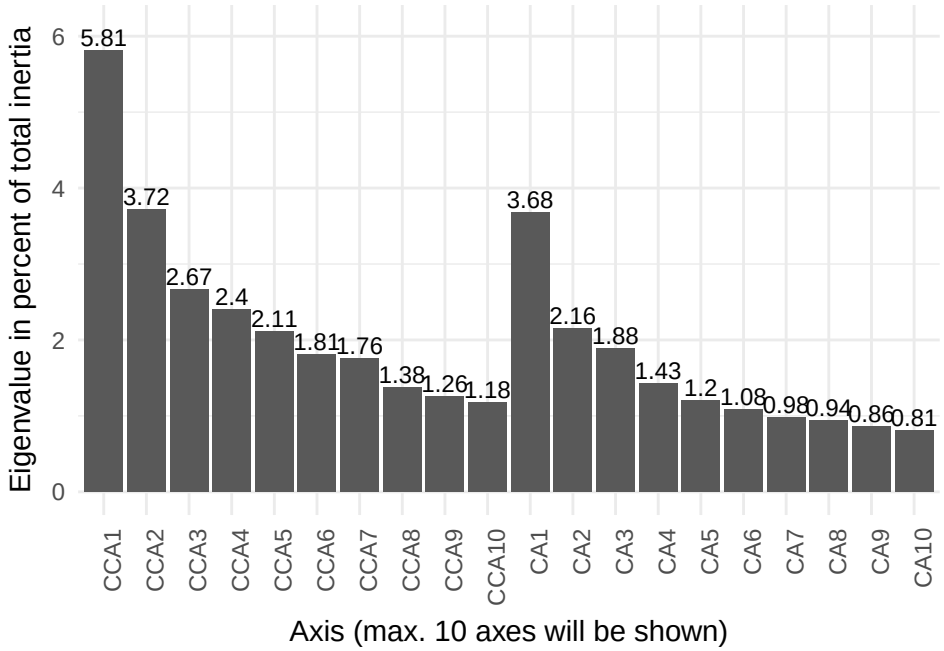
B.1 Scree plot of Figure 4.1



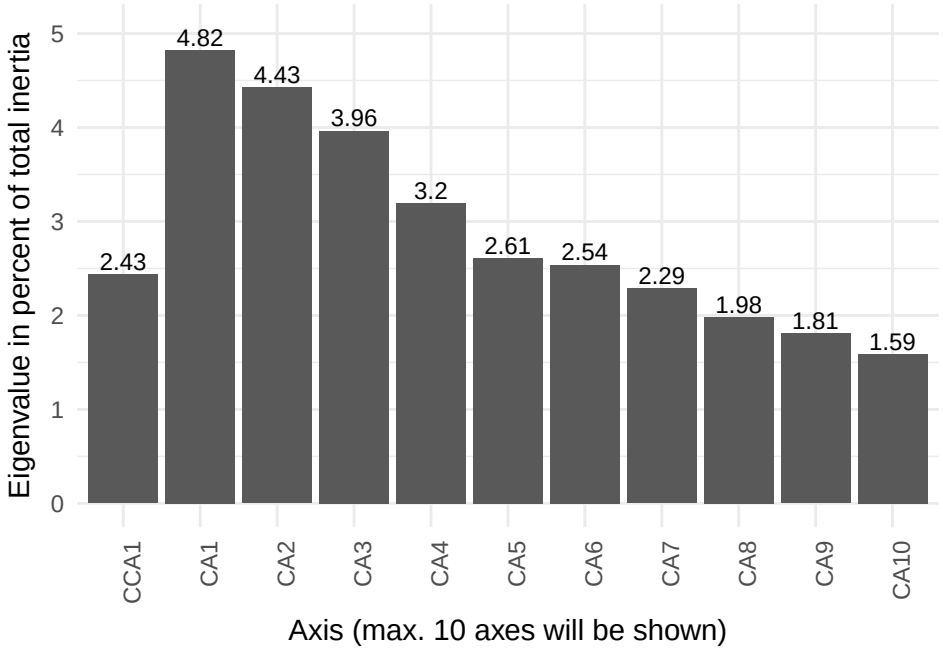
B.2 Scree plot of Figure 4.2

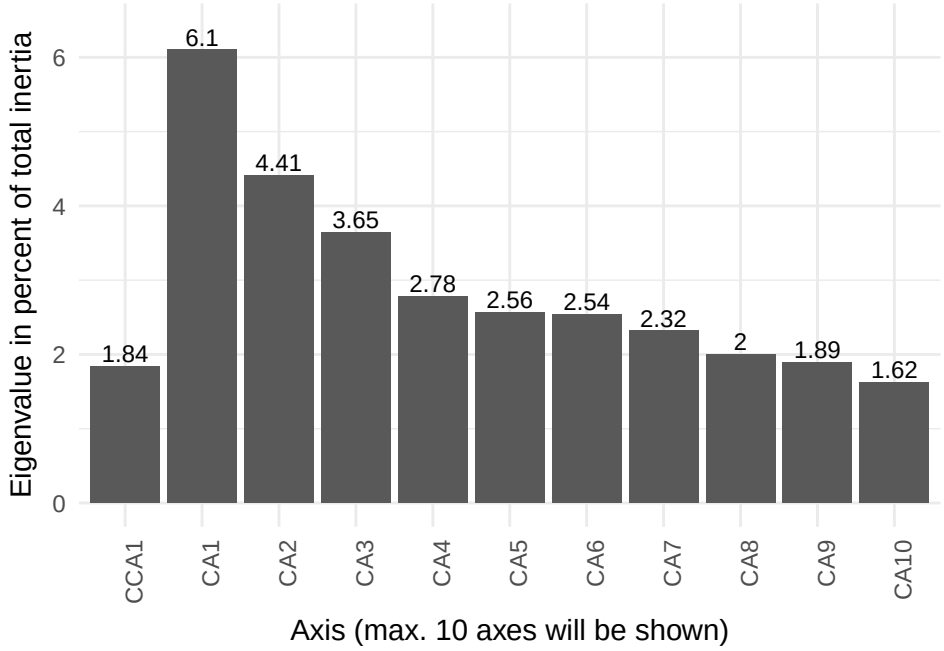


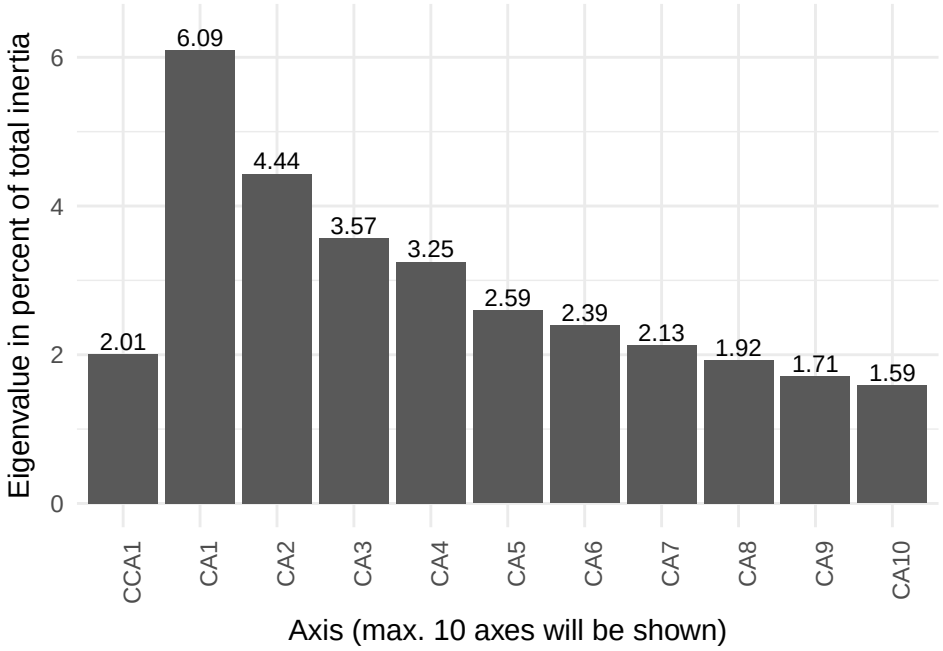
B.3 Scree plot of Figure 4.3

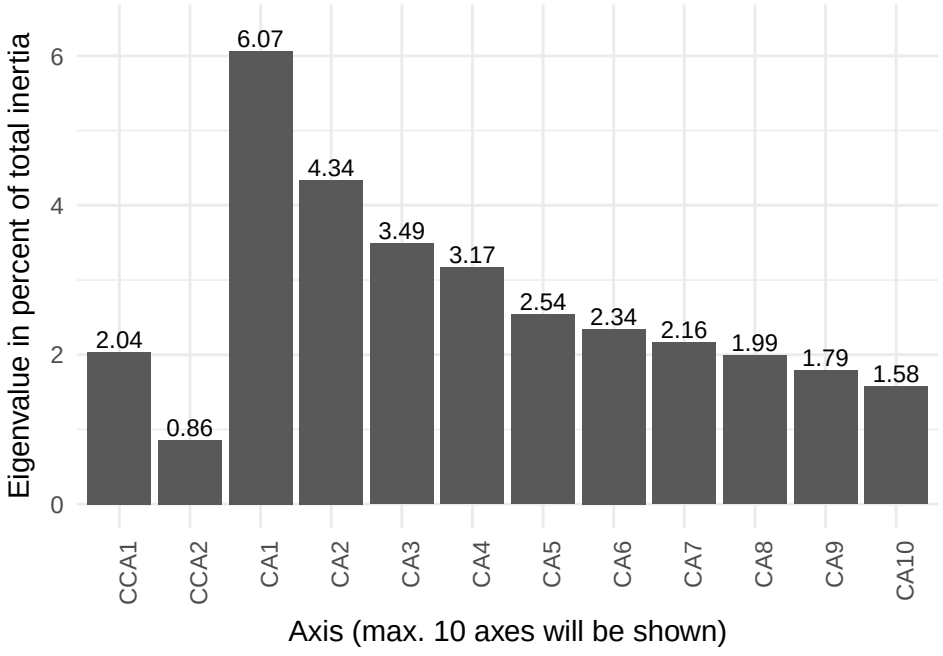


B.4 Scree plot of Figure 5.2(A-D)

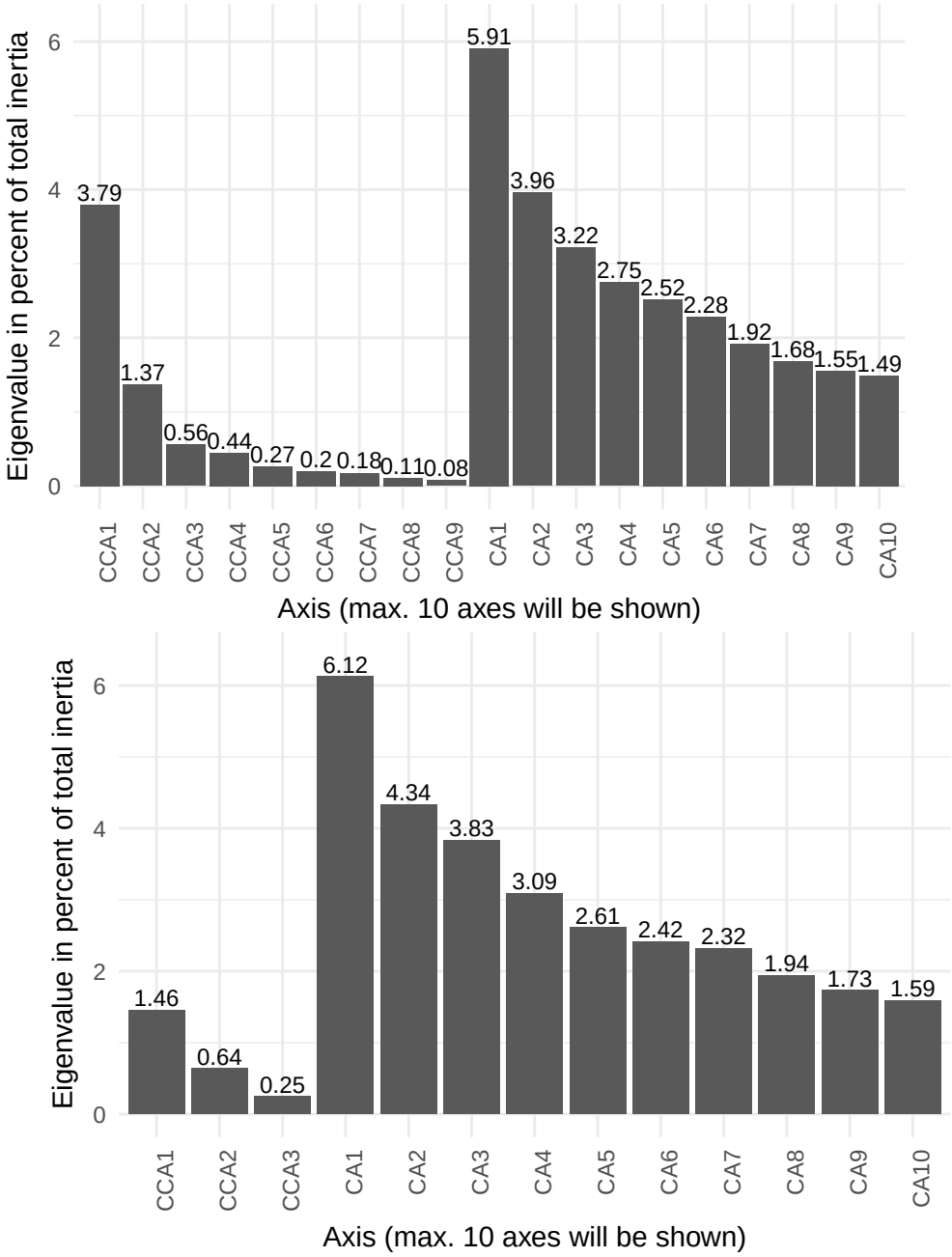





Eigenvalue in percent of total inertia
0
2
4
6
2.01
6.09
4.44
3.57
3.25
2.59
2.39
2.13
1.92
1.71
1.59
CCA1
CA1
CA2
CA3
CA4
CA5
CA6
CA7
CA8
CA9
CA10
Axis (max. 10 axes will be shown)


Eigenvalue in percent of total inertia
0
2
4
6
2.04
0.86
6.07
4.34
3.49
3.17
2.54
2.34
2.16
1.99
1.79
1.58
CCA1
CCA2
CA1
CA2
CA3
CA4
CA5
CA6
CA7
CA8
CA9
CA10
Axis (max. 10 axes will be shown)

B.5 Scree plot of Figure 5.3(A+B)



C. Characteristics of the WWTPs

Table C.1: Summary of the design of the 32 wastewater treatment plants. EBPR: Enhanced Biological Phosphorous Removal, BNR: Biological Nutrient Removal, RSS: Return Sludge Sidestream, Industrial Inf.: Industrial Inflow

Plant	Configuration	Design	RSS	Primary Setling	Digester	Industrial Inf.
Aabenraa	Alternating	EBPR	RSS	No	Yes	Low
Aaby	Alternating	EBPR	NO	No	Yes	Avg.
Aalborg E	Alternating	EBPR	RSS	No	Yes	Avg.
Aalborg W	Alternating	EBPR	RSS	Yes	Yes	Avg.
Aars	Alternating	BNR	NO	No	No	High
Avedoere	Alternating	BNR	NO	No	Yes	Avg.
Bjergmarken	Alternating	EBPR	NO	No	Yes	Avg.
Boeslum	Recirculation	EBPR	NO	No	No	Low
Egaa	Recirculation	EBPR	RSS	No	No	High
Ejby Moelle	Alternating	EBPR	NO	Yes	Yes	High
Esbjerg E	Recirculation	BNR	NO	Yes	Yes	High
Esbjerg W	Recirculation	BNR	NO	Yes	Yes	High
Fornaes	Recirculation	BNR	NO	No	Yes	Low
Haderslev	Alternating	EBPR	RSS	No	No	Low
Hirtshals	Alternating	EBPR	NO	No	No	High
Hjoerring	Recirculation	EBPR	NO	Yes	Yes	Avg.

Horsens	Recirculation	BNR	NO	Yes	Yes	High
Kerteminde	Recirculation	EBPR	NO	No	No	Avg.
Kolding	Alternating	EBPR	NO	Yes	Yes	Avg.
Lundtofte	Alternating	EBPR	NO	Yes	Yes	Low
Marselisborg	Alternating	BNR	NO	Yes	Yes	High
Middelfart	Recirculation	BNR	NO	Yes	Yes	Avg.
Moerke	Alternating	EBPR	NO	No	No	Avg.
Odense NE	Alternating	EBPR	NO	Yes	Yes	Avg.
Odense NW	Alternating	BNR	NO	Yes	Yes	Avg.
Randers	Recirculation	EBPR	RSS	Yes	Yes	Low
Ribe	Recirculation	EBPR	RSS	No	No	Avg.
Ringkoebing	Alternating	EBPR	RSS	Yes	Yes	Avg.
Skive	Recirculation	EBPR	RSS	No	No	Avg.
Skive	Recirculation	EBPR	RSS	No	No	High
Soeholt	Alternating	EBPR	RSS	Yes	Yes	High
Viborg	Recirculation	EBPR	RSS	Yes	Yes	Avg.
Viby	Recirculation	EBPR	RSS	No	Yes	Low

References

Acinas, S. G., Marcelino, L. A., Klepac-Ceraj, V., & Polz, M. F. (2004). Divergence and Redundancy of 16S rRNA Sequences in Genomes with Multiple rrn Operons. *Journal of Bacteriology*, *186*(9), 2629–2635. http://doi.org/10.1128/JB.186.9.2629-2635.2004

Albertsen, M., Karst, S. M., Ziegler, A. S., Kirkegaard, R. H., & Nielsen, P. H. (2015). Back to basics - The influence of DNA extraction and primer choice on phylogenetic analysis of activated sludge communities. *PLoS ONE*, *10*(7), 1–15. http://doi.org/10.1371/journal.pone.0132783

Amann, R. I., Binder, B. J., Olson, R. J., Chisholm, S. W., Devereux, R., & Stahl, D. A. (1990). Combination of 16S rRNA-targeted oligonucleotide probes with flow cytometry for analyzing mixed microbial populations. *Applied and Environmental Microbiology*, *56*(6), 1919–25. http://doi.org/0099-2240/90/061919-07

Angly, F. E., Dennis, P. G., Skarshewski, A., Vanwonterghem, I., Hugenholtz, P., & Tyson, G. W. (2014). CopyRighter: a rapid tool for improving the accuracy of microbial community profiles through lineage-specific gene copy number correction. *Microbiome*, *2*(1), 11. http://doi.org/10.1186/2049-2618-2-11

Ardern, E., & Lockett, W. T. (1914). Experiments on the Oxidation of Sewage without the Aid of Filters. *Journal of Society of Chemical Industry*, *33*(10), 523–539. http://doi.org/10.1002/jctb.5000331005

Arumugam, M., Raes, J., Pelletier, E., Le Paslier, D., Yamada, T., Mende, D. R., ...

Bork, P. (2011). Enterotypes of the human gut microbiome. *Nature*, *473*(7346), 174–180. http://doi.org/10.1038/nature09944

Ashelford, K. E., Chuzhanova, N. A., Fry, J. C., Jones, A. J., & Weightman, A. J. (2005). At Least 1 in 20 16S rRNA Sequence Records Currently Held in Public Repositories Is Estimated To Contain Substantial Anomalies. *Applied and Environmental Microbiology*, *71*(12), 7724–7736. http://doi.org/10.1128/AEM.71.12.7724-7736.2005

Bae, H. S., Moe, W. M., Yan, J., Tiago, I., Costa, M. S. da, & Rainey, F. A. (2006). Brooklawnia cerclae gen. nov., sp. nov., a propionate-forming bacterium isolated from chlorosolvent-contaminated groundwater. *International Journal of Systematic and Evolutionary Microbiology*, *56*(8), 1977–1983. http://doi.org/10.1099/ijs.0.64317-0

Braak, C. J. F. ter, & Prentice, I. C. (1988). A Theory of Gradient Analysis. *Advances in Ecological Research*, *18*(C), 271–317. http://doi.org/10.1016/S0065-2504(08)60183-X

Bray, J. R., & Curtis, J. T. (1957). An Ordination of the Upland Forest Communities of Southern Wisconsin. *Ecological Monographs*, *27*(4), 325–349. http://doi.org/10.2307/1942268

Buttigieg, P. L., & Ramette, A. (2014). A guide to statistical analysis in microbial ecology: A community-focused, living review of multivariate data analyses. *FEMS Microbiology Ecology*, *90*(3), 543–550. http://doi.org/10.1111/1574-6941.12437

Clark, H. W., & Adams, G. O. (1914). Sewage treatment by aeration and contact in tanks containing layers of slate. *Record*, *69*, 158–159.

Clarridge, J. E. (2004). Impact of 16S rRNA gene sequence analysis for identification of bacteria on clinical microbiology and infectious diseases. *Clinical Microbiology Reviews*, *17*(4), 840–62, tableof contents. http://doi.org/10.1128/CMR.17.4.

840-862.2004

Cole, J. R., Wang, Q., Fish, J. A., Chai, B., McGarrell, D. M., Sun, Y., ... Tiedje, J. M. (2014). Ribosomal Database Project: data and tools for high throughput rRNA analysis. *Nucleic Acids Research*, *42*(D1), D633–D642. http://doi.org/10.1093/nar/gkt1244

Dahle, H., & Birkeland, N. K. (2006). Thermovirga lienii gen. nov., sp. nov., a novel moderately thermophilic, anaerobic, amino-acid-degrading bacterium isolated from a North Sea oil well. *International Journal of Systematic and Evolutionary Microbiology*, *56*(7), 1539–1545. http://doi.org/10.1099/ijs.0.63894-0

DeLong, E., Wickham, G., & Pace, N. (1989). Phylogenetic stains: ribosomal RNA-based probes for the identification of single cells. *Science*, *243*(4896), 1360–1363. http://doi.org/10.1126/science.2466341

Edgar, R. C. (2013). UPARSE: highly accurate OTU sequences from microbial amplicon reads. *Nature Methods*, *10*(10), 996–8. http://doi.org/10.1038/nmeth.2604

Eikelboom, D. H. (1975). Filamentous organisms observed in activated sludge. *Water Research*, *9*(4), 365–388. http://doi.org/http://dx.doi.org/10.1016/0043-1354(75)90182-7

Fahrbach, M., Kuever, J., Meinke, R., Kämpfer, P., & Hollender, J. (2006). Denitratisoma oestradiolicum gen. nov., sp. nov., a 17beta-oestradiol-degrading, denitrifying betaproteobacterium. *International Journal of Systematic and Evolutionary Microbiology*. http://doi.org/10.1099/ijs.0.63672-0

Fudou, R., Jojima, Y., Iizuka, T., & Yamanaka, S. (2002). Haliangium ochraceum gen. nov., sp. nov. and Haliangium tepidum sp. nov.: novel moderately halophilic myxobacteria isolated from coastal saline environments. *The Journal of General and Applied Microbiology*, *48*(2), 109–116. http://doi.org/10.2323/jgam.48.109

Fuller, G. W. (1915). RECENT DEVELOPMENTS IN SEWAGE DISPOSAL. *The Public*

Health Journal, *6*(3), 103–106. Retrieved from http://www.jstor.org/stable/41996799

Goodall, D. (1953). Objective methods for the classification of vegetation. I. The use of positive interspecific correlation. *Australian Journal of Botany*, *1*(1), 39–63.

Hamlin, C. (1988). William Dibdin and the Idea of Biological Sewage Treatment. *Technology and Culture*, *29*(2), 189–218. http://doi.org/10.2307/3105523

Hill, M. O. (1973). Reciprocal Averaging: An Eigenvector Method of Ordination. *The Journal of Ecology*, *61*(1), 237. http://doi.org/10.2307/2258931

Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, *24*(6), 417–441. http://doi.org/10.1037/h0071325

Hutchinson, G. E. (1957). Concluding remarks. *Cold Spring Harbor Symposia on Quantitative Biology*, *22*, 415–427. http://doi.org/10.1101/SQB.1957.022.01.039

Ibarbalz, F. M., Figuerola, E. L., & Erijman, L. (2013). Industrial activated sludge exhibit unique bacterial community composition at high taxonomic ranks. *Water Research*, *47*(11), 3854–3864. http://doi.org/10.1016/j.watres.2013.04.010

Ihaka, R., & Gentleman, R. (1996). R: A Language for Data Analysis and Graphics. *Journal of Computational and Graphical Statistics*, *5*(3), 299–314. http://doi.org/10.1080/10618600.1996.10474713

J Farrell, & A Rose. (1967). Temperature Effects on Microorganisms. *Annual Review of Microbiology*, *21*(1), 101–120. http://doi.org/10.1146/annurev.mi.21.100167.000533

Karst, S. M., Dueholm, M. S., McIlroy, S. J., Kirkegaard, R. H., Nielsen, P. H., & Albertsen, M. (2016). Thousands of primer-free, high-quality, full-length SSU rRNA sequences from all domains of life. *BioRxiv*. http://doi.org/10.1101/

070771

Knights, D., Ward, T. L., McKinlay, C. E., Miller, H., Gonzalez, A., McDonald, D., & Knight, R. (2014). Rethinking "Enterotypes". *Cell Host & Microbe*, *16*(4), 433–437. http://doi.org/10.1016/j.chom.2014.09.013

Koren, O., Knights, D., Gonzalez, A., Waldron, L., Segata, N., Knight, R., ... Ley, R. E. (2013). A Guide to Enterotypes across the Human Body: Meta-Analysis of Microbial Community Structures in Human Microbiome Datasets. *PLoS Computational Biology*, *9*(1). http://doi.org/10.1371/journal.pcbi.1002863

Kristiansen, R., Nguyen, H. T. T., Saunders, A. M., Nielsen, J. L., Wimmer, R., Le, V. Q., ... Nielsen, P. H. (2013). A metabolic model for members of the genus Tetrasphaera involved in enhanced biological phosphorus removal. *The ISME Journal*, *7*(3), 543–554. http://doi.org/10.1038/ismej.2012.136

Kruskal, J. B. (1964). Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika*, *29*(1), 1–27. http://doi.org/10.1007/BF02289565

Kuczynski, J., Liu, Z., Lozupone, C., McDonald, D., Fierer, N., & Knight, R. (2010). Microbial community resemblance methods differ in their ability to detect biologically relevant patterns. *Nature Methods*, *7*(10), 813–819. http://doi.org/10.1038/nmeth.1499

Legendre, P., & Gallagher, E. (2001). Ecologically meaningful transformations for ordination of species data. *Oecologia*, *129*(2), 271–280. http://doi.org/10.1007/s004420100716

Legendre, P., & Legendre, L. (2012). *Numerical Ecology*. Elsevier Science. Retrieved from https://books.google.dk/books?id=6ZBOA-iDviQC

Loosdrecht, M. C. van, Nielsen, P., Lopez-Vazquez, C., & Brdjanovic, D. (2016). Experimental procedures in wastewater treatment. *Experimental Procedures in Wastewater Treatment*. Retrieved from http://www.iwapublishing.com/books/

9781780404745/experimental-methods-wastewater-treatment

Lozupone, C. A., Hamady, M., Kelley, S. T., & Knight, R. (2007). Quantitative and qualitative ?? diversity measures lead to different insights into factors that structure microbial communities. *Applied and Environmental Microbiology*, *73*(5), 1576–1585. http://doi.org/10.1128/AEM.01996-06

Lozupone, C., & Knight, R. (2005). UniFrac : a New Phylogenetic Method for Comparing Microbial Communities UniFrac : a New Phylogenetic Method for Comparing Microbial Communities. *Applied and Environmental Microbiology*, *71*(12), 8228–8235. http://doi.org/10.1128/AEM.71.12.8228

McDonald, D., Price, M. N., Goodrich, J., Nawrocki, E. P., DeSantis, T. Z., Probst, A., ... Hugenholtz, P. (2012). An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. *The ISME Journal*, *6*(3), 610–618. http://doi.org/10.1038/ismej.2011.139

McIlroy, S. J., Saunders, A. M., Albertsen, M., Nierychlo, M., McIlroy, B., Hansen, A. A., ... Nielsen, P. H. (2015). MiDAS: The field guide to the microbes of activated sludge. *Database*, *2015*(2), 1–8. http://doi.org/10.1093/database/bav062

Metzker, M. L. (2010). Sequencing technologies — the next generation. *Nature Reviews Genetics*, *11*(1), 31–46. http://doi.org/10.1038/nrg2626

Minchin, P. R. (1987). An evaluation of the relative robustness of techniques for ecological ordination. *Theory and Models in Vegetation Science*, 89–107. http://doi.org/10.1007/978-94-009-4061-1_9

Mino, T., Loosdrecht, M. C. M. van, & Heijnen, J. J. (1998). Review paper - Microbiology and biochemistry of the enhanced biological phosphate removal process. *Wat. Res.*, *11*(32), 3193–3207.

Mullis, K., Faloona, F., Scharf, S., Saiki, R., Horn, G., & Erlich, H. (1986). Specific Enzymatic Amplification of DNA In Vitro: The Polymerase Chain Reaction. *Cold Spring Harbor Symposia on Quantitative Biology*, *51*(1), 263–273. http:

//doi.org/10.1101/SQB.1986.051.01.032

Nielsen, P. H., & McMahon, K. D. (2014). *CHAPTER 2. MICROBIOLOGY AND MICROBIAL ECOLOGY OF THE ACTIVATED SLUDGE PROCESS*. IWA Publishing. Retrieved from https://books.google.dk/books?id=J7cDBAAAQBAJ

Nielsen, P. H., Mielczarek, A. T., Kragelund, C., Nielsen, J. L., Saunders, A. M., Kong, Y., ... Vollertsen, J. (2010). A conceptual ecosystem model of microbial communities in enhanced biological phosphorus removal plants. *Water Research*, *44*(17), 5070–5088. http://doi.org/10.1016/j.watres.2010.07.036

Orhon, D. (2015). Evolution of the activated sludge process: the first 50 years. *Journal of Chemical Technology & Biotechnology*, *90*(4), 608–640. http://doi.org/10.1002/jctb.4565

Pearson, K. (1901). On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, *2*(1), 559–572. http://doi.org/10.1080/14786440109462720

Podani, J., & Miklós, I. (2002). Resemblance Coefficients and the Horseshoe Effect in Principal Coordinates Analysis. *Ecology*, *83*(12), 3331–3343. Retrieved from http://www.jstor.org/stable/3072083

Qin, D., Abdi, N. M., & Fredrick, K. (2007). Characterization of 16S rRNA mutations that decrease the fidelity of translation initiation. *RNA (New York, N.Y.)*, *13*(12), 2348–55. http://doi.org/10.1261/rna.715307

Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., ... Glockner, F. O. (2013). The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Research*, *41*(D1), D590–D596. http://doi.org/10.1093/nar/gks1219

Quinn, G. R., & Skerman, V. B. D. (1980). Herpetosiphon—Nature's scavenger? *Current Microbiology*, *4*(1), 57–62. http://doi.org/10.1007/BF02602893

Ramette, A. (2007). Multivariate analyses in microbial ecology. *FEMS Microbiology*

Ecology, *62*(2), 142–60. http://doi.org/10.1111/j.1574-6941.2007.00375.x

Sanger, F., & Coulson, A. (1975). A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *Journal of Molecular Biology*, *94*(3), 441–448. http://doi.org/10.1016/0022-2836(75)90213-2

Sanger, F., Nicklen, S., & Coulson, A. R. (1977). DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences*, *74*(12), 5463–5467. http://doi.org/10.1073/pnas.74.12.5463

Saunders, A. M., Albertsen, M., Vollertsen, J., & Nielsen, P. H. (2016). The activated sludge ecosystem contains a core community of abundant organisms. *The ISME Journal*, *10*(1), 11–20. http://doi.org/10.1038/ismej.2015.117

Seviour, R., & Nielsen, P. H. (2010). *Microbial Ecology of Activated Sludge*. IWA Publishing. Retrieved from https://books.google.dk/books?id=z9ohFDYcAsgC

Shepard, R. N. (1966). Metric structures in ordinal data. *Journal of Mathematical Psychology*, *3*(2), 287–315. http://doi.org/10.1016/0022-2496(66)90017-4

Valenzuela-Encinas, C., Neria-González, I., Alcántara-Hernández, R. J., Estrada-Alvarado, I., Zavala-Díaz de la Serna, F. J., Dendooven, L., & Marsch, R. (2009). Changes in the bacterial populations of the highly alkaline saline soil of the former lake Texcoco (Mexico) following flooding. *Extremophiles*, *13*(4), 609–621. http://doi.org/10.1007/s00792-009-0244-4

Whittaker, R. H. (1967). Gradient Analysis of Vegetation*. *Biological Reviews*, *42*(2), 207–264. http://doi.org/10.1111/j.1469-185X.1967.tb01419.x

Whittaker, R. H. (1972). Evolution and Measurement of Species Diversity. *Taxon*, *21*(2/3), 213–251. Retrieved from http://www.jstor.org/stable/1218190

Woese, C. R., & Fox, G. E. (1977). Phylogenetic structure of the prokaryotic domain: The primary kingdoms. *Proceedings of the National Academy of Sciences*, *74*(11),

5088–5090. http://doi.org/10.1073/pnas.74.11.5088

Zhang, T., Shao, M.-F., & Ye, L. (2012). 454 Pyrosequencing reveals bacterial diversity of activated sludge from 14 sewage treatment plants. *The ISME Journal*, *6*(6), 1137–1147. http://doi.org/10.1038/ismej.2011.188