

# Enterotyping Danish Wastewater Treatment Plants

Exploring patterns in the microbial community  
composition 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



# Acknowledgements

Tak AAU, MA, PHN



# Preface

This is a report about microbes and ordination. Yes, it is!



# Table of Contents

<b>Chapter 1: Introduction</b>	<b>1</b>
1.1 Wastewater treatment	1
1.2 16S amplicon sequencing	1
<b>Chapter 2: Ordination in Microbial Ecology</b>	<b>3</b>
2.1 Exploratory vs. Explanatory	4
2.2 Niche theory and the double-zero problem	6
2.3 Distance-based ordination	8
2.3.1 Principal Coordinates Analysis	9
2.3.2 non-Metric Multidimensional Scaling	10
2.3.3 Distance- and (dis)similarity metrics	11
2.4 Eigenanalysis-based ordination	13
2.4.1 Principal Components Analysis	14
2.4.2 Redundancy Analysis	16
2.4.3 Correspondence Analysis	17
2.4.4 Canonical Correspondence Analysis	19
2.5 Data transformation	21
<b>Chapter 3: Aims</b>	<b>23</b>
<b>Chapter 4: Materials and Methods</b>	<b>25</b>
4.1 Sampling	25
4.2 Library preparation	25

4.2.1	Polymerase Chain Reaction . . . . .	25
4.2.2	Library pool cleanup . . . . .	25
4.3	DNA sequencing and bioinformatics . . . . .	25
4.3.1	Filtering . . . . .	26
4.4	Data visualisation . . . . .	26
<b>Chapter 5: Results Part 1: Exploring</b>		<b>27</b>
5.1	How does the microbial community composition describe the wastewater treatment plants? . . . . .	31
<b>Chapter 6: Results Part 2: Explaining</b>		<b>37</b>
6.1	The effect of plant design on the microbial community composition .	37
6.1.1	Configuration . . . . .	37
6.1.2	Enhanced Biological Phosphorus Removal (EBPR) vs Biological Nutrient Removal (BNR) . . . . .	37
6.1.3	The effect of primary settling . . . . .	37
6.1.4	The effect of industrial inflow water . . . . .	37
6.2	Plant stability over time . . . . .	37
<b>Chapter 7: Discussion</b>		<b>49</b>
<b>Chapter 8: Conclusion</b>		<b>51</b>
<b>Appendix A: Characteristics of the WWTPs</b>		<b>53</b>
<b>References</b>		<b>55</b>



# List of Tables

2.1	A classification of the most used ordination methods in microbial ecology. Based on (Braak & Prentice, 1988; Ramette, 2007) . . . . .	6
A.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 . . . . .	53



# List of Figures

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.	5
2.2	Species response (or abundance-) curves in most cases show a uni-modal distribution along an environmental gradient. Adapted from (Whittaker, 1967)	7
2.3	A scree plot of the eigenvalues of the 7 axes obtained from the PCA ordination seen in Figure 2.1.	14
2.4	The main structure of two-dimensional data can often be described in one dimension, by positions on a straight line. Adapted from (Pearson, 1901)	15
2.5	Principles of unconstrained analyses	20
5.1	Overview heatmap - updated 14/3-15:20	29
5.2	Overview PCA, hellinger - updated 14/3-15:20	30
5.3	Overview PCoA, Jensen-Shannon Divergence	31
5.4	Overview CCA, hellinger - updated 14/3-15:20	32
5.5	Overview species PCA, hellinger - updated 14/3-15:20	33
5.6	TOP 20 OTU's - updated 14/3-15:20	34
5.7	overview CCA, hellinger - updated 14/3-15:20	35
6.1	EV fit CCA, hellinger - updated 14/3-15:20	38

6.2	EV fit RDA, hellinger - updated 14/3-15:20 . . . . .	39
6.3	Plant design, hellinger - updated 14/3-15:20 . . . . .	40
6.4	EBPR vs BNR, hellinger - updated 14/3-15:20 . . . . .	41
6.5	Primary settling, hellinger - updated 14/3-15:20 . . . . .	42
6.6	Industrial, hellinger - updated 14/3-15:20 . . . . .	43
6.7	Stability . . . . .	44
6.8	Stability . . . . .	45
6.9	Stability . . . . .	46
6.10	Stability Viborg . . . . .	47
6.11	Stability Ejby Moelle . . . . .	48

# Abstract

The preface pretty much says it all.

Second paragraph of abstract starts here.



# 1. Introduction

The importance of access to clean drinking water for human civilisation is without doubt essential. A report by the World Health Organisation (WHO) in 2015 showed that roughly 663 million people still lack improved drinking water resources.

## 1.1 Wastewater treatment

detflot

## 1.2 16S amplicon sequencing

hej





## 2. Ordination in Microbial Ecology

To visualise very large and complex multidimensional data as obtained with next-generation sequencing of sometimes hundreds of samples each containing hundreds of different microorganisms, perhaps the most suited method is ordination. In essence, ordination seeks to reduce the dimensionality of a contingency table into a few, usually 1-3, more important dimensions (hence it is also termed dimensionality reduction techniques) to ease interpretation which makes it perfectly suited for complex ecological data. Through 'dimensional yoga' one obtains  $n - 1$  new dimensions, 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)similarity, distance, correlation 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, 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 below.(P. Legendre & Legendre, 2012)

## 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. However it is still possible to color or shape the points by known environmental variables (see Figure 2.1), 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)

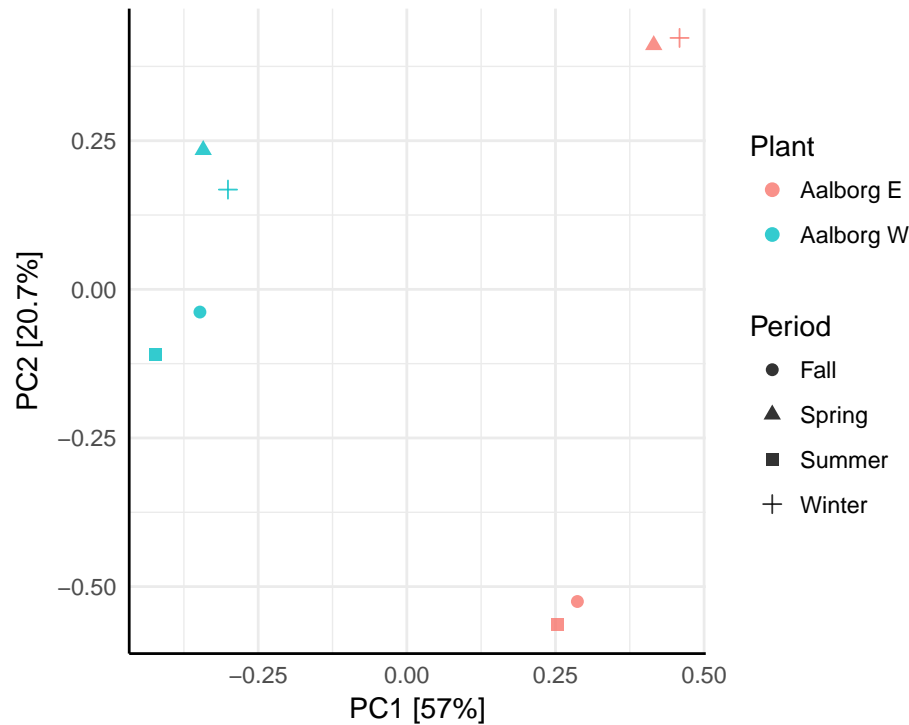


Figure 2.1: A minimal example of exploratory data analysis using Principal Components Analysis. 8 samples from two wastewater treatment plants, Aalborg East and Aalborg West have been analysed and the points have been shaped by when the samples were taken in 2012.

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. 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). (Braak & Prentice, 1988; P. Legendre & Legendre, 2012)

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

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

## 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 2.2, 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 a Euclidean distances like PCA and RDA. (P. Legendre & Legendre, 2012)

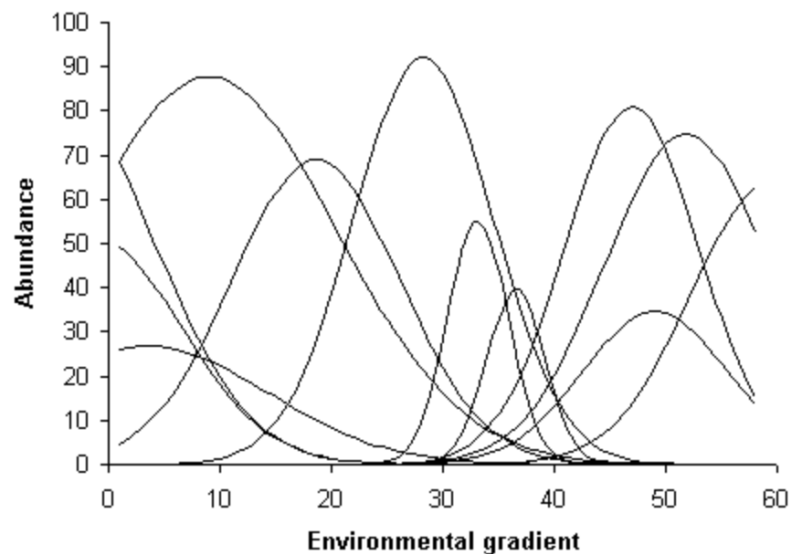


Figure 2.2: Species response (or abundance-) curves in most cases show a unimodal distribution along an environmental gradient. Adapted from (Whittaker, 1967)

The goal of using ordination is often to represent patterns or differences between samples based on species abundances to draw ecologically meaningful conclusions about the sampling site(s) and their corresponding  $\beta$ - or  $\gamma$ -diversity. 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 preferring 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*. 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. Choosing the correct ordination method and, for some, also a distance metric 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 (more on transformation in Chapter 2.5) or manual curation of the data 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 metric explicitly or implicitly to calculate distances, respectively (P. Legendre & Legendre, 2012). As

such, eigenanalysis-based ordination methods use a metric 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 metrics. When calculated, a symmetrical matrix is then obtained containing *distance coefficients*, or more generally *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 metric and choosing it wisely (more on metrics in Chapter 2.3.3).

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). There is also Polar Ordination, however it is rarely used and will not be detailed in the following. (Ramette, 2007)

### 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 metric is Euclidean, the results obtained by PCoA is identical to that of PCA.

Since the choice of metric 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 metric also influences how the results are to be interpreted, as the original data is then a function of the chosen metric, 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 a semi-metric is 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 metric and its value should not be directly referred to when performing PCoA. 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 metric.(Ramette, 2007)

### 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 specific metric. However, nMDS is different from PCoA on numerous aspects. PCA as well 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 as well 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 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. Generally choosing more dimensions will lower the stress value. (Ramette, 2007)

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 metrics

As mentioned, choosing a distance- or (dis)similarity metric that makes ecological sense is crucial for the analysis and subsequent interpretation of the ordination. A distance metric 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 (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:

### The four metric properties

1. The distance between identical objects is 0, which is the lowest possible value:

$$\text{if } a = b, \text{ then } D(a, b) = 0$$

2. When the compared objects are not identical, the coefficient, and the distance, has a positive value:

$$\text{if } a \neq b, \text{ then } D(a, b) > 0$$

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 distance coefficients, they are then termed semi-metric coefficients. When this is the case, distances cannot be ordinated (at least reliably) in Euclidean space and nMDS is the optimal ordination method. Therefore carefully considering whether the chosen metric is suitable for the ordination method used is important, if not then interpretation of the result should be done with caution. When all four properties are met, both nMDS and PCoA can be used. Dissimilarity and similarity coefficients can be both metric or semi-metric. They 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)

LIST OF COMMON METRICS: BRAY-CURTIS, KULCZYNSKI, JACCARD, MANHATTAN/CITYBLOCK METRICS ETC

## 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 function (metric), where distance coefficients are not first calculated manually by the user. They all have a few things in common, however:

- There is always one unique solution to the 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 each other, thus uncorrelated and express their own 'unique' inertia (P. Legendre & Legendre, 2012)

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'.

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 2.3. 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. Below, the remaining four ordination methods listed in 2.1 will be explained briefly.(Ramette, 2007)

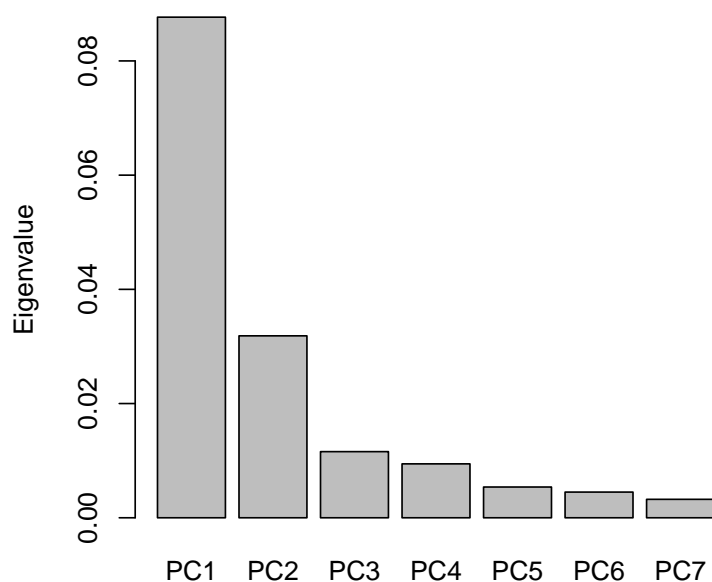


Figure 2.3: A scree plot of the eigenvalues of the 7 axes obtained from the PCA ordination seen in Figure 2.1.

### 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 2.4.

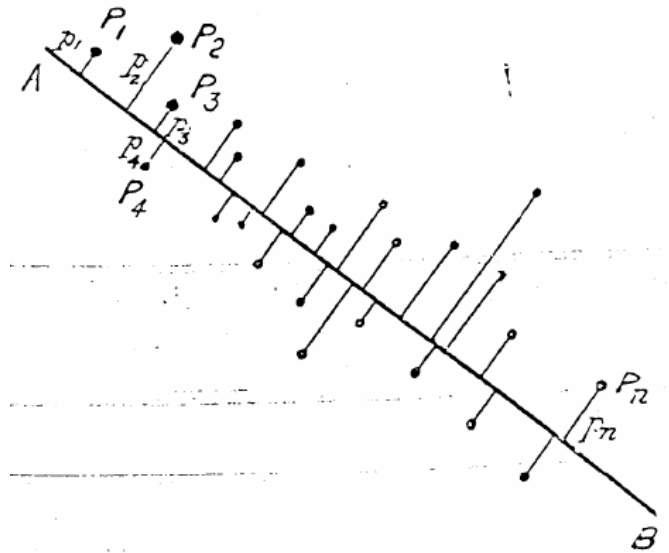


Figure 2.4: The main structure of two-dimensional data can often be described in one dimension, by positions on a straight line. Adapted from (Pearson, 1901)

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 fundamental 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 aligned so that it represents the most inertia in the data. 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.

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 usefulness without appropriate data transformation. Of course, this is not a problem when analysing  $\alpha$ -diversities 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, 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?”. (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

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

where  $Y = [y_{ij}]$  is a species abundance table of the size  $(n \times p)$  with sites (rows)  $i = [1 \dots n]$  and species (columns)  $j = [1 \dots p]$ . (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 are usually indicated on the axis labels to give an indication of the significance of the constrained variable(s). When doing constrained analysis, both with RDA and CCA (Chapter 2.4.4), a plot with not only sites and species are made, but also environmental vectors are plotted together in one plot called a triplot. This is a very convenient way to visualise all three types of information at once to easily draw ecological conclusions about the sites.

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. (Ramette, 2007)

### 2.4.3 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 testing how “this species correspond to that site”. CA is thus considered more of a qualitative representation of the data and is based on the Pearson chi-squared statistic of the form  $\chi^2 = \sum \frac{(\text{observed} - \text{expected})^2}{\text{expected}}$ .

CA is developed by several authors independently 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 sample scores as defined by the  $\chi^2$ -metric described below, where the weights are, with community composition data, specifically species abundances. This means that species abundances to some degree contribute to the calculated distances, however they are also calculated relative to the average abundances

and therefore does not influence the results as much as with the Euclidean-based ordination methods. 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. With CA, and CCA, it is therefore important to consider the importance of low abundant species in the particular study and give less weight if needed (Braak & Prentice, 1988; P. Legendre & Gallagher, 2001).

guldorn:

awesome for  $\beta$ -diversity

during the procedure, the result is matched with a random solution and a P-value can be extracted to reject the NULL-hypothesis of no association

a site with more species than another will have higher chi-squared distances

- low abundances has an unduly high influence, use transformation
- eigenvalues are correlation coefficients between samples and species

*Total inertia* The total inertia of a CA solution conveys the degree to which the values of rows and columns correspond to each other. More specifically, it reflects the degree to which rows and columns deviate from the null hypothesis of “no association”, according to the logic of the Pearson’s  $\chi^2$  statistic.

- Eigenvalues in CA are not equivalent to those of PCA and should not be interpreted in terms of “variation” but “inertia”.

### The chi-squared distance

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

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



#### **2.4.4 Canonical Correspondence Analysis**

Constrained version of CA

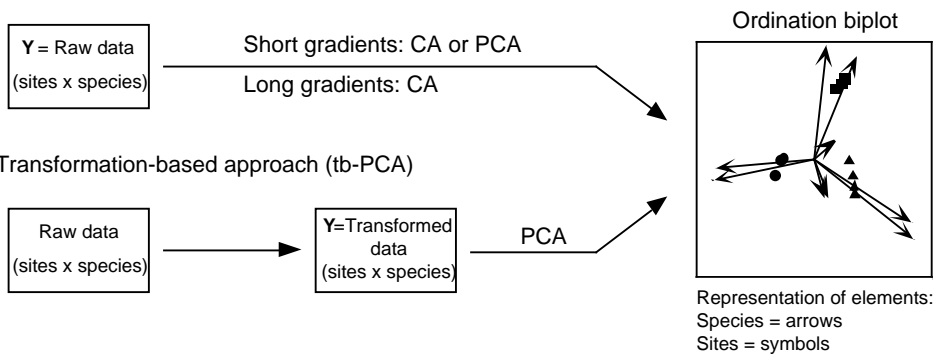
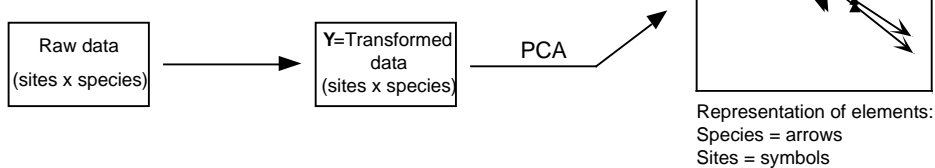
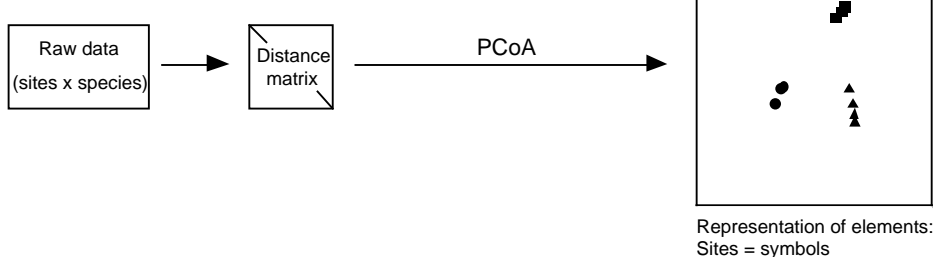
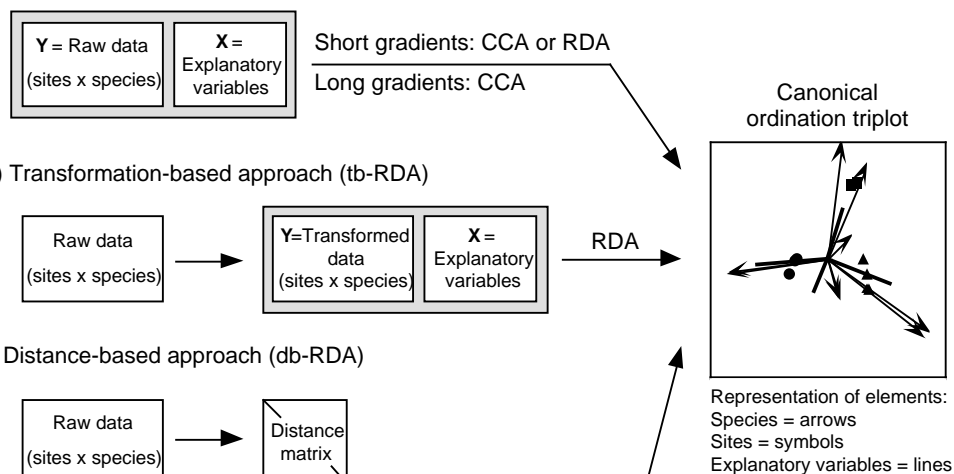
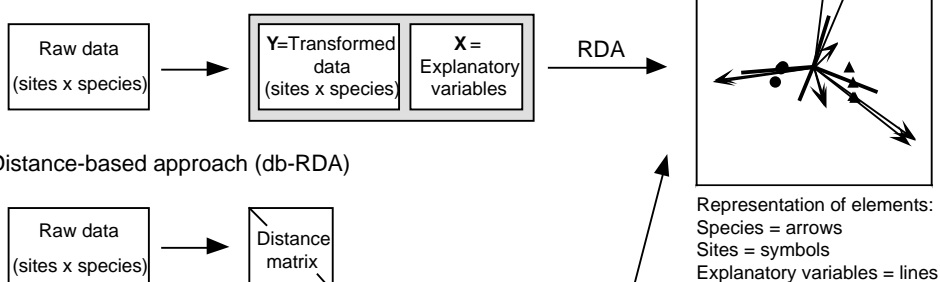
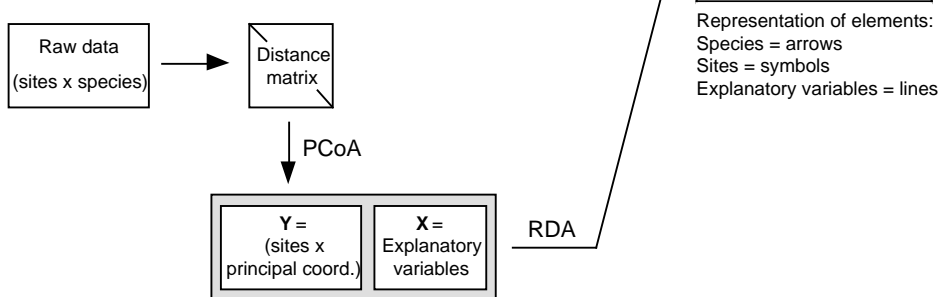
**Unconstrained ordination of species data****(a) Classical approach****(b) Transformation-based approach (tb-PCA)****(c) Distance-based approach (PCoA)****Constrained ordination of species data****(d) Classical approach****(e) Transformation-based approach (tb-RDA)****(f) Distance-based approach (db-RDA)**

Figure 2.5: Principles of unconstrained analyses

## 2.5 Data transformation

Important, weighting low/high abundant species

$$D_{\text{Hellinger}}(x_1, x_2) = \sqrt{\sum_{j=1}^p \left( \sqrt{\frac{y_{1j}}{y_{+j}}} - \sqrt{\frac{y_{2j}}{y_{2+}}} \right)^2} \quad (2.3)$$

(P. Legendre & Gallagher, 2001)



### 3. Aims

- What are the key differences in the microbial community compositions of danish Wastewater Treatment Plants?
- What causes the differences?
  - Underlying environmental gradients
  - Known physical-/operational parameters
  - Weather conditions
  - Plant design
  - Influent microbial community composition
  - Other. . .
- Is the microbial community composition stable over time and how can it be related to plant performance?
- Are the abundances of microbes with similar function(s) correlated?
- Are microbes present in the influent also present in the activated sludge and in the same amount? (Odense, Janni's data)



## 4. Materials and Methods

### 4.1 Sampling

Samples were taken at danish wastewater treatment plants mostly more than two times a year from 2006 to 2015. BLA BLA

### 4.2 Library preparation

Prepared by: (or refer to a protocol?) ### DNA extraction Beadbeating+DNA extraction kit xxyy+centrifuge BLA BLA

#### 4.2.1 Polymerase Chain Reaction

Temperature scheme? Primer region V1-1234 BLA BLA

#### 4.2.2 Library pool cleanup

BLA BLA

### 4.3 DNA sequencing and bioinformatics

Usearch9, MiDAS database version xx samples checked with rarefaction curves, filtered which? Only looking at WWTPs with many samples

### 4.3.1 Filtering

OTU's at or below 0.1% abundance in all samples have been removed.

## 4.4 Data visualisation

Done in R, wrote my own functions. Made a shiny app for quick, brief analysis.

Both available at github: <https://github.com/knaldhat>

FILTRATION:  $\leq 0.1\%$  REMOVED WHICH RESULTED IN 32 WWTPs



## 5. Results Part 1: Exploring

Filtering the samples from 56 Wastewater Treatment Plants (WWTPs) resulted in a remaining 622 samples from 32 WWTPs containing a total of 21728 different OTU's. Because this is a very large amount of data to visualise (even for ordination), the following plots may be best viewed in the interactive online *bookdown* version of this report, where it is possible to zoom in the plots and hover the points for additional information. It is available at <https://github.com/KasperSkytte/MasterThesis>

will in this chapter be described using exploratory ordination methods (as described in Chapter @ref{ordination}). The differences between the WWTPs are represented simply by the distance between sample points if not noted otherwise. The points are colored by a unique color for each WWTP and the corresponding name of the WWTP is written at the approximate center of all the sample points from the particular WWTP. Because there are 622 samples to represent, it can be hard to visualise simpler, so an interactive version of this report can be accessed online, where it is possible to zoom in the plots. **HERE OR IN MM?!**

The following two PCA and NMDS plots show very similar groupings of the WWTPs. With both ordination methods the WWTP's Ribe, Hjoerring, Esbjerg W+E, Viborg, Randers, Aars, Horsens, Hirtshals, Ringkoebing and Moerke seem to be clustered together positioned opposite of Odense NE+NW, Bjergmarken, Egaa, Middelfart, Aalborg E+W, Soenderborg, Avedoere, Aabenraa, Viby, Ejby Moelle, Aaby, Marselisborg and Kolding. However they seem to disagree on Skive and Soeholt, and Fornæs, Boeslum, Haderslev, Soeholt, Kolding and Lundtofte seem to be in between on PCA+RDA, but with NMDS Kolding, Lundtofte and Skive.

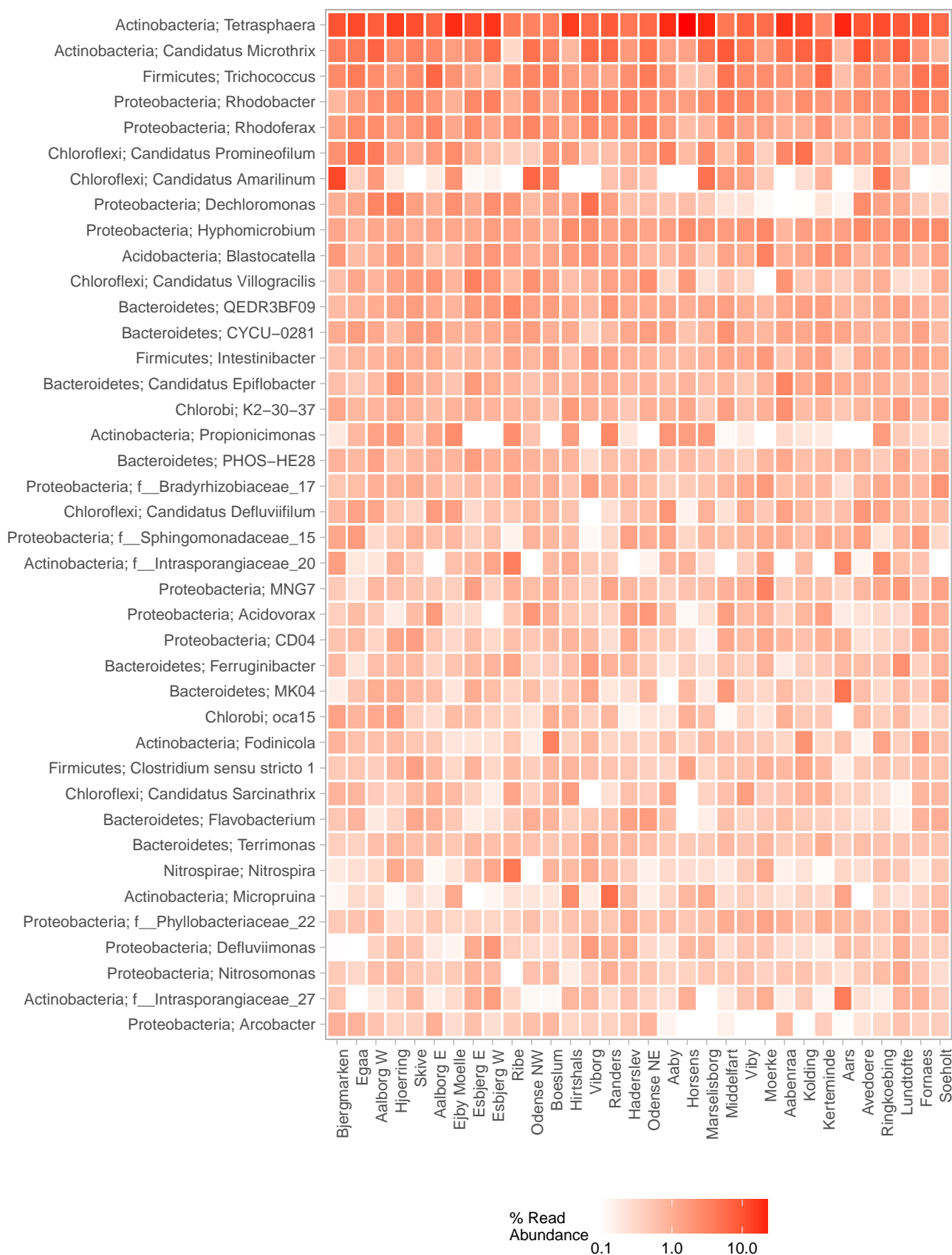


Figure 5.1: Overview heatmap - updated 14/3-15:20

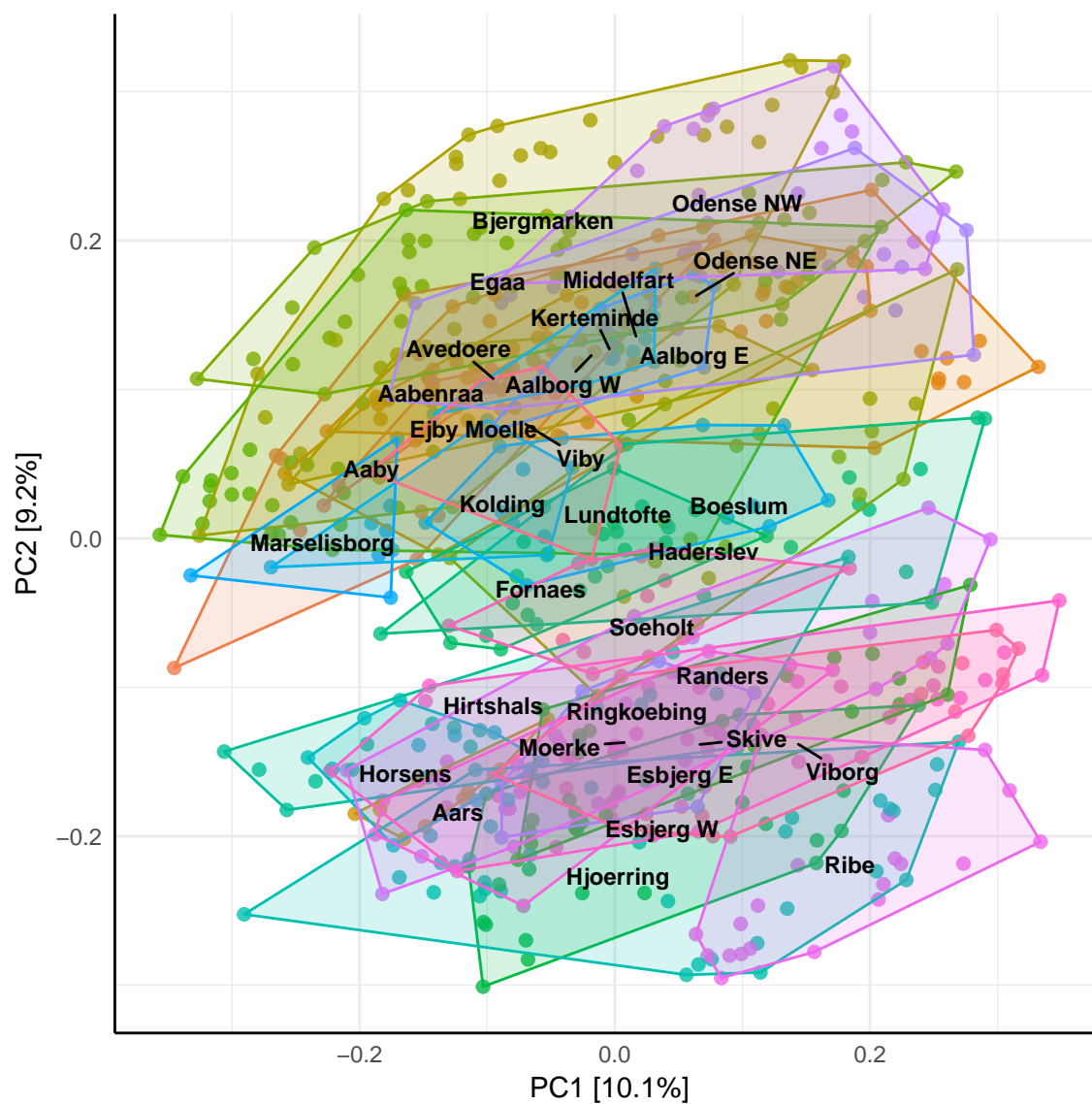


Figure 5.2: Overview PCA, hellinger - updated 14/3-15:20

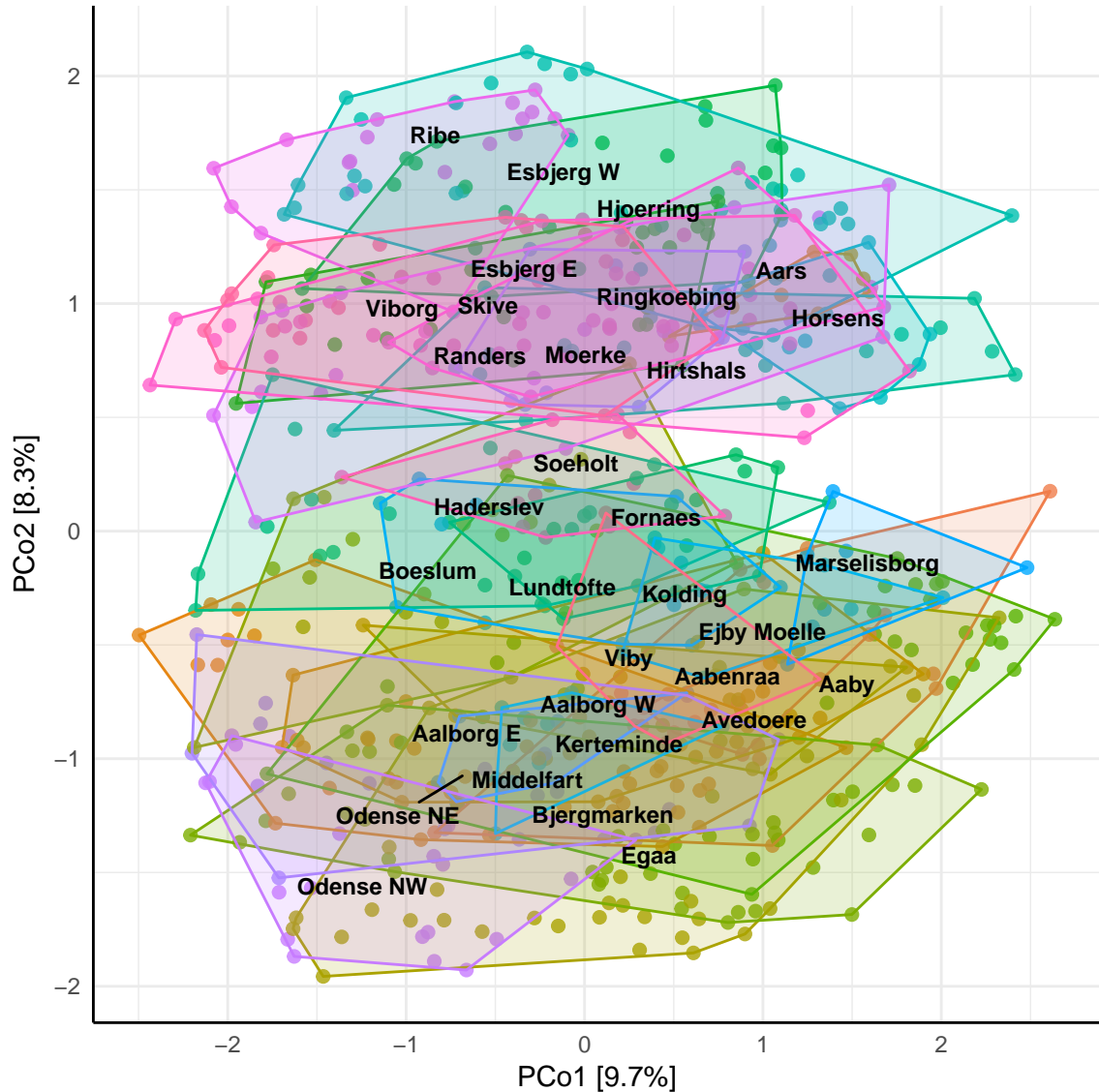


Figure 5.3: Overview PCoA, Jensen-Shannon Divergence

test: As seen on Figure 5.4 bla bla...

## 5.1 How does the microbial community composition describe the wastewater treatment plants?

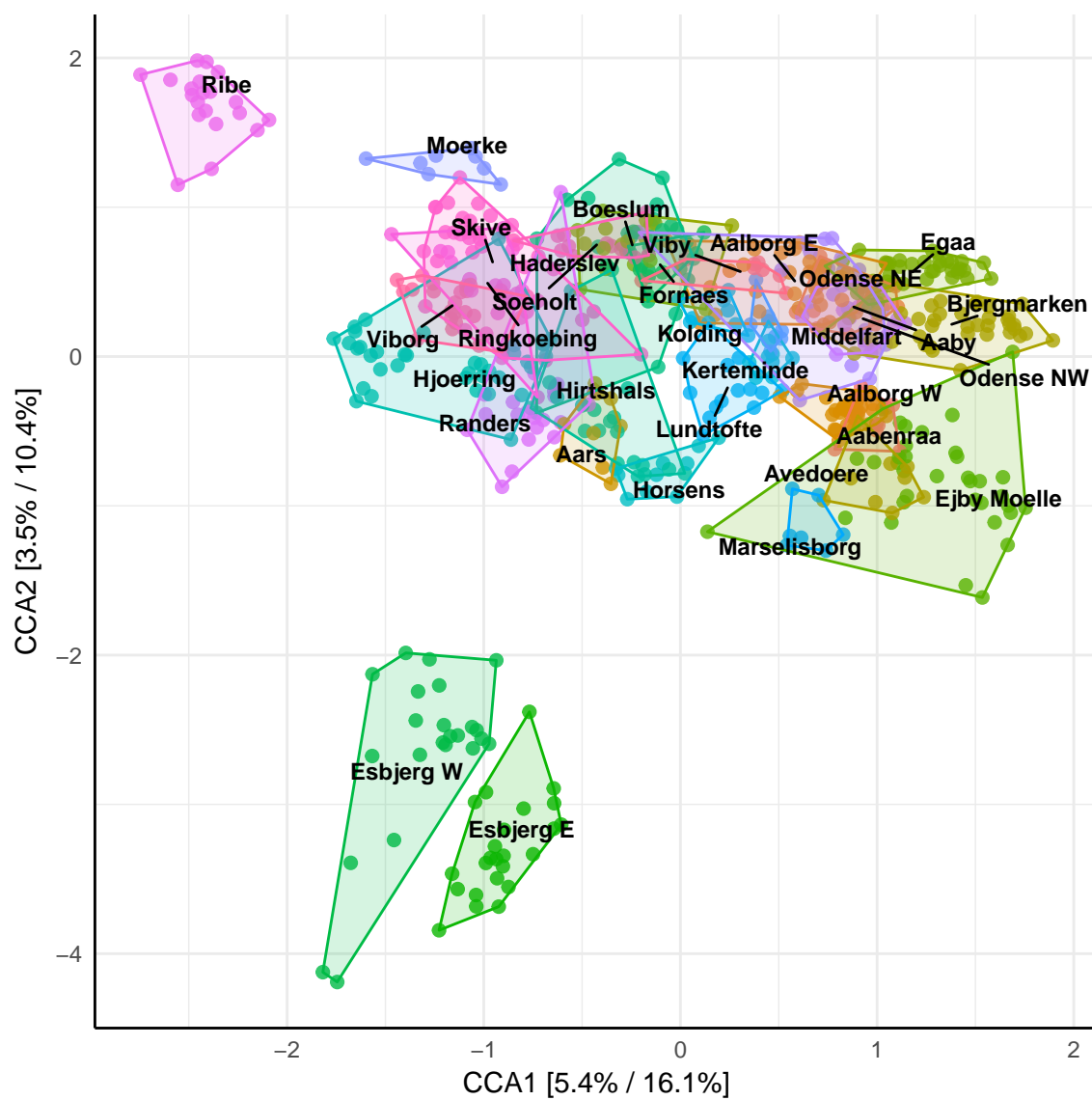


Figure 5.4: Overview CCA, hellinger - updated 14/3-15:20





Figure 5.6: TOP 20 OTU's - updated 14/3-15:20



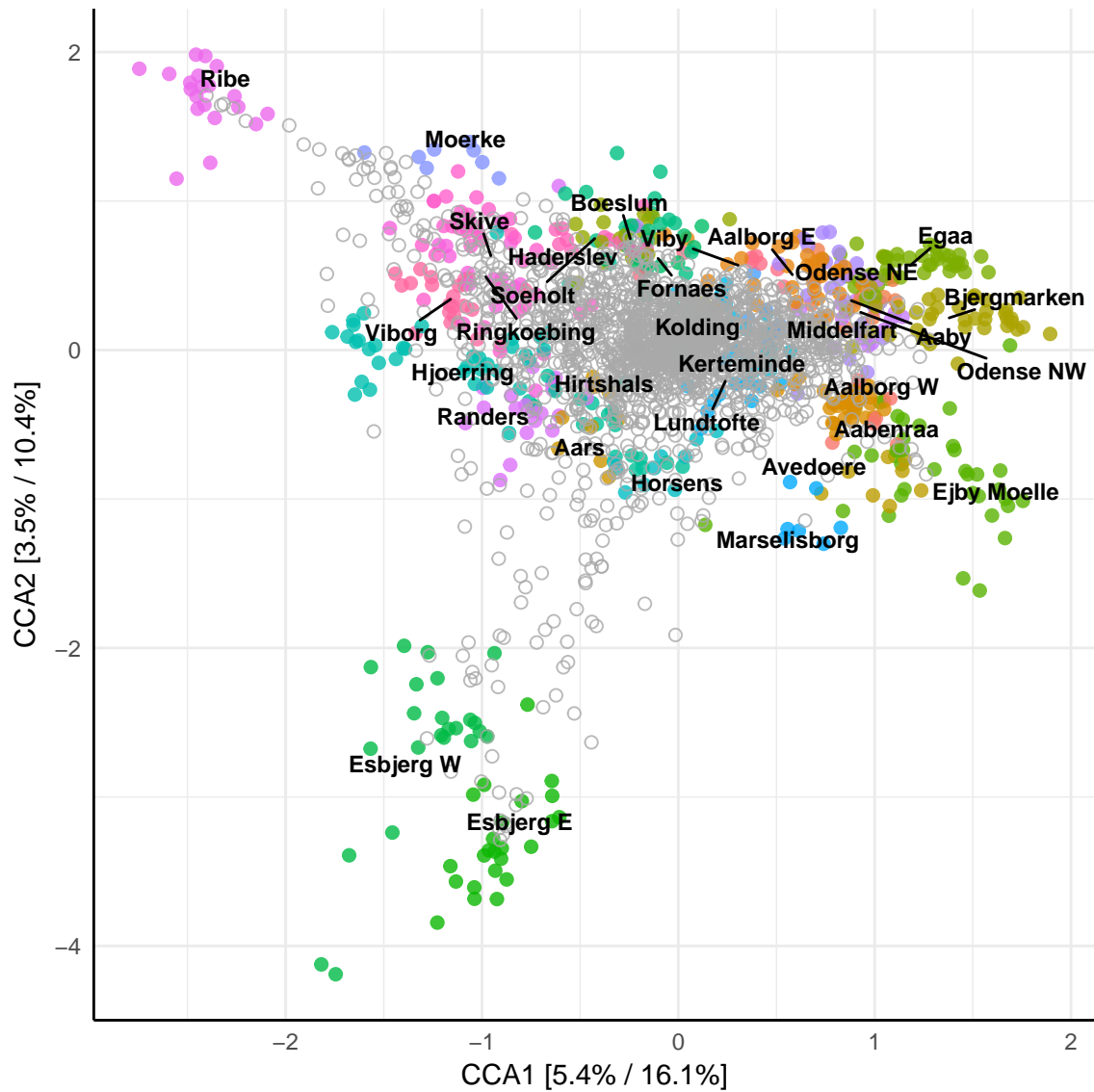


Figure 5.7: overview CCA, hellinger - updated 14/3-15:20



## 6. Results Part 2: Explaining

Soeholt: Digester fra summer 2009 Viborg: RSS og EBPR fra early 2012

### 6.1 The effect of plant design on the microbial community composition

#### 6.1.1 Configuration

[1] "Eigenvalue of RDA1: 1.2%"

#### 6.1.2 Enhanced Biological Phosphorus Removal (EBPR) vs Biological Nutrient Removal (BNR)

[1] "Eigenvalue of RDA1: 0.9%"

#### 6.1.3 The effect of primary settling

[1] "Eigenvalue of RDA1: 1%"

#### 6.1.4 The effect of industrial inflow water

[1] "Eigenvalues of RDA: 1.1%, 0.3%"

### 6.2 Plant stability over time

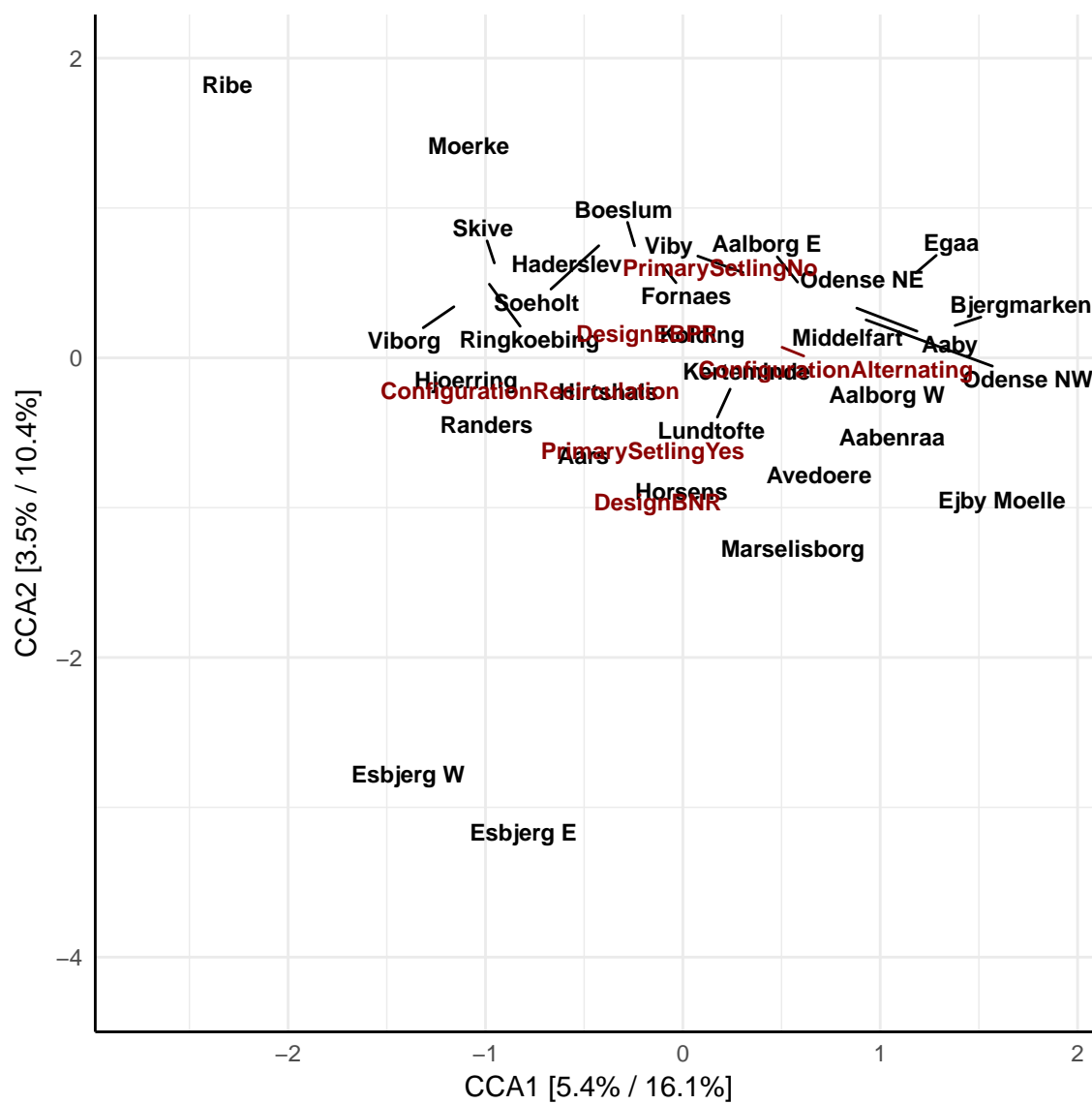


Figure 6.1: EV fit CCA, hellinger - updated 14/3-15:20

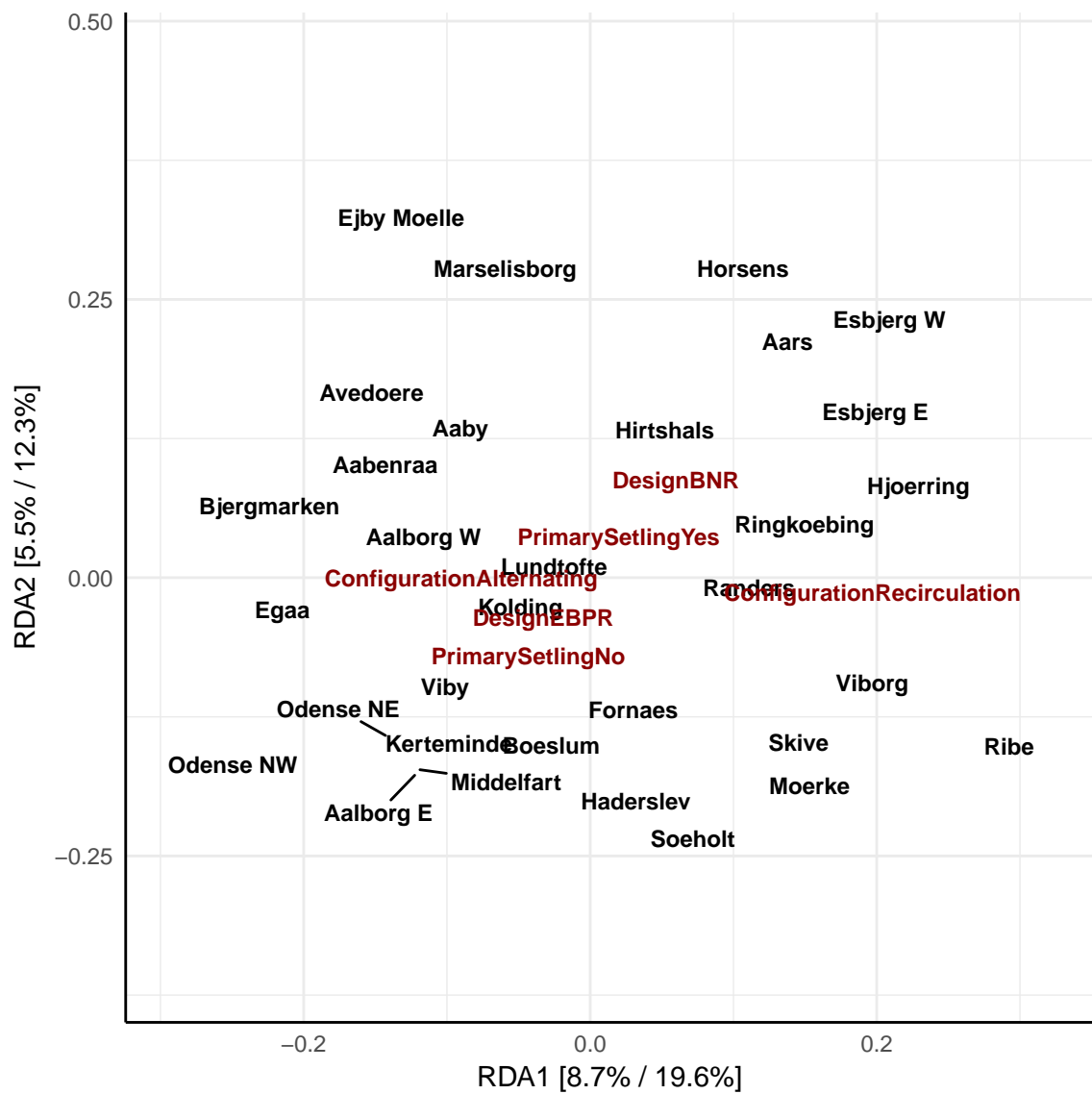


Figure 6.2: EV fit RDA, hellinger - updated 14/3-15:20

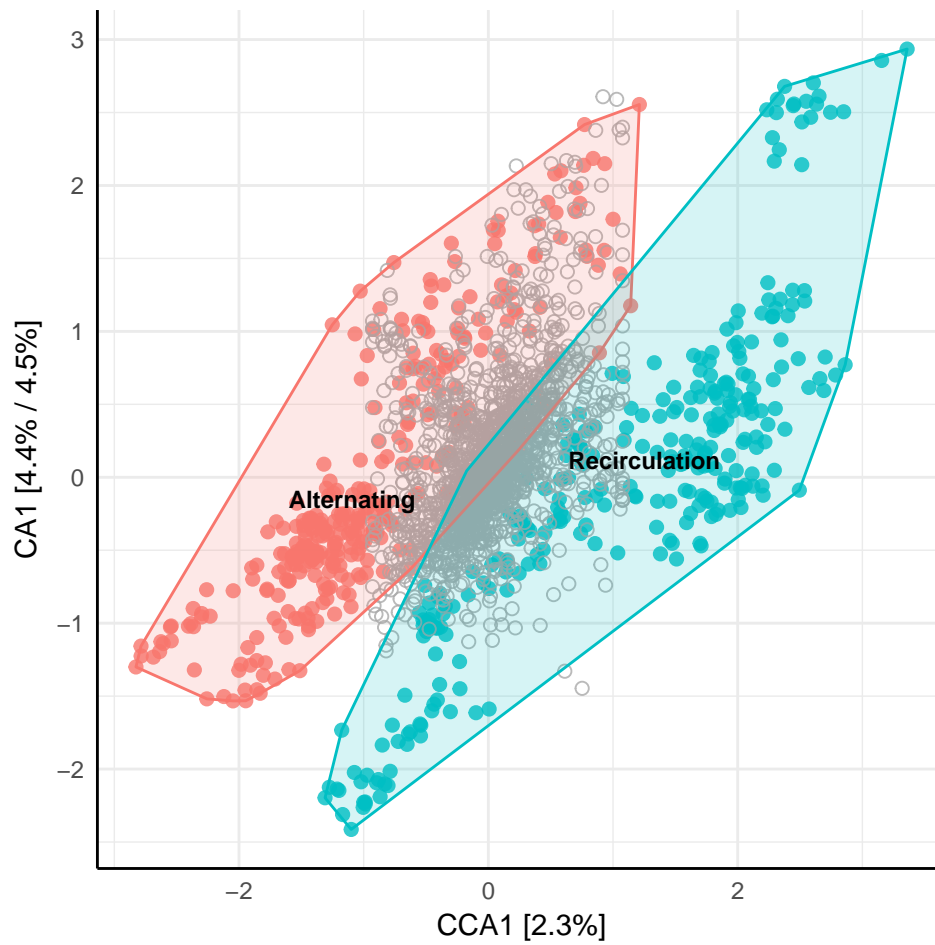


Figure 6.3: Plant design, hellinger - updated 14/3-15:20

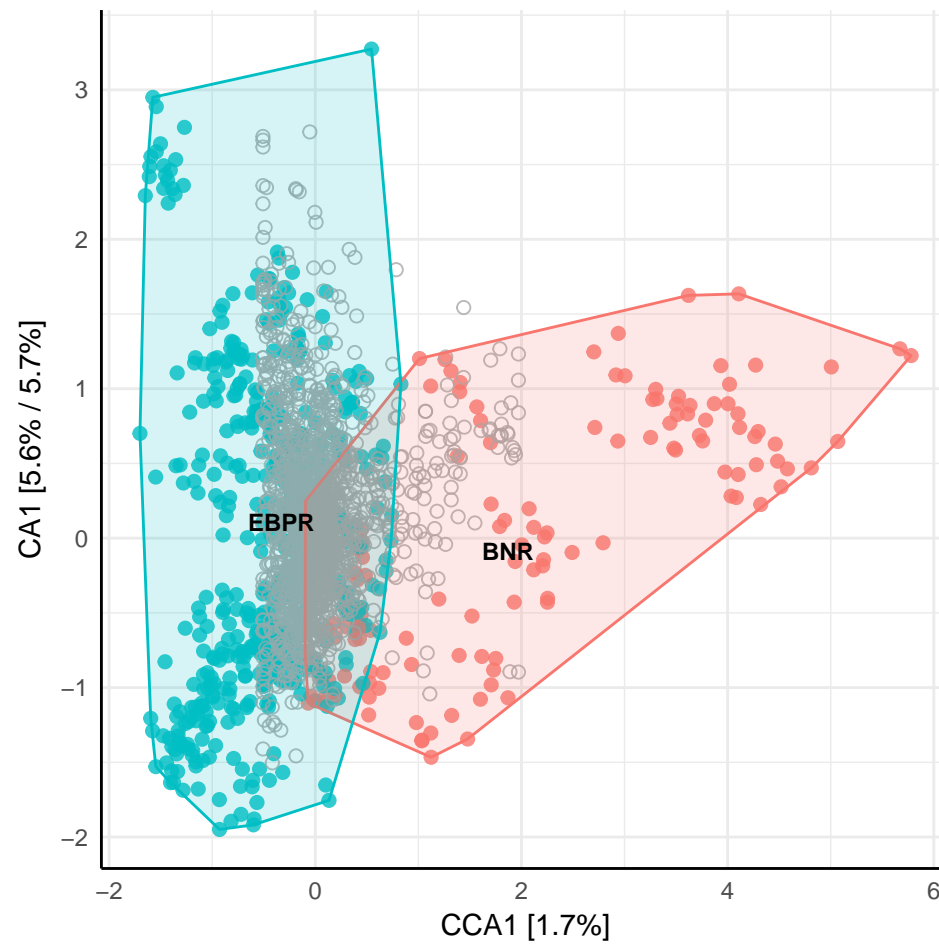


Figure 6.4: EBPR vs BNR, hellinger - updated 14/3-15:20

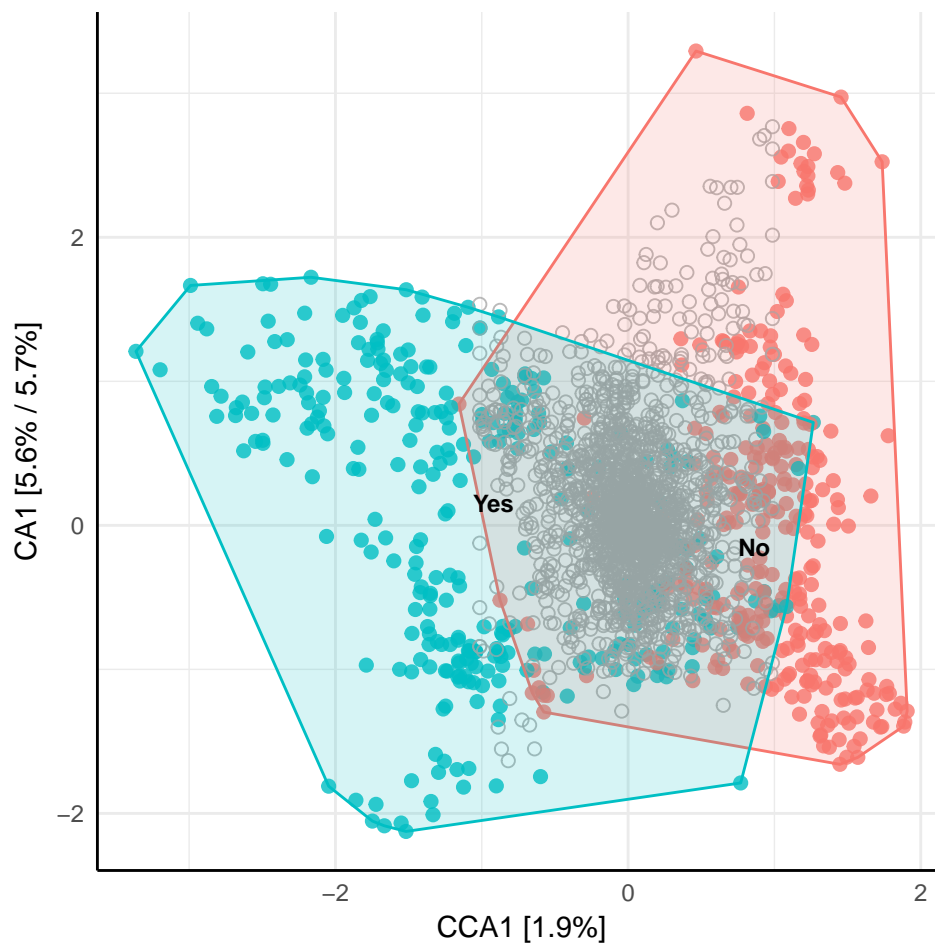


Figure 6.5: Primary settling, hellinger - updated 14/3-15:20



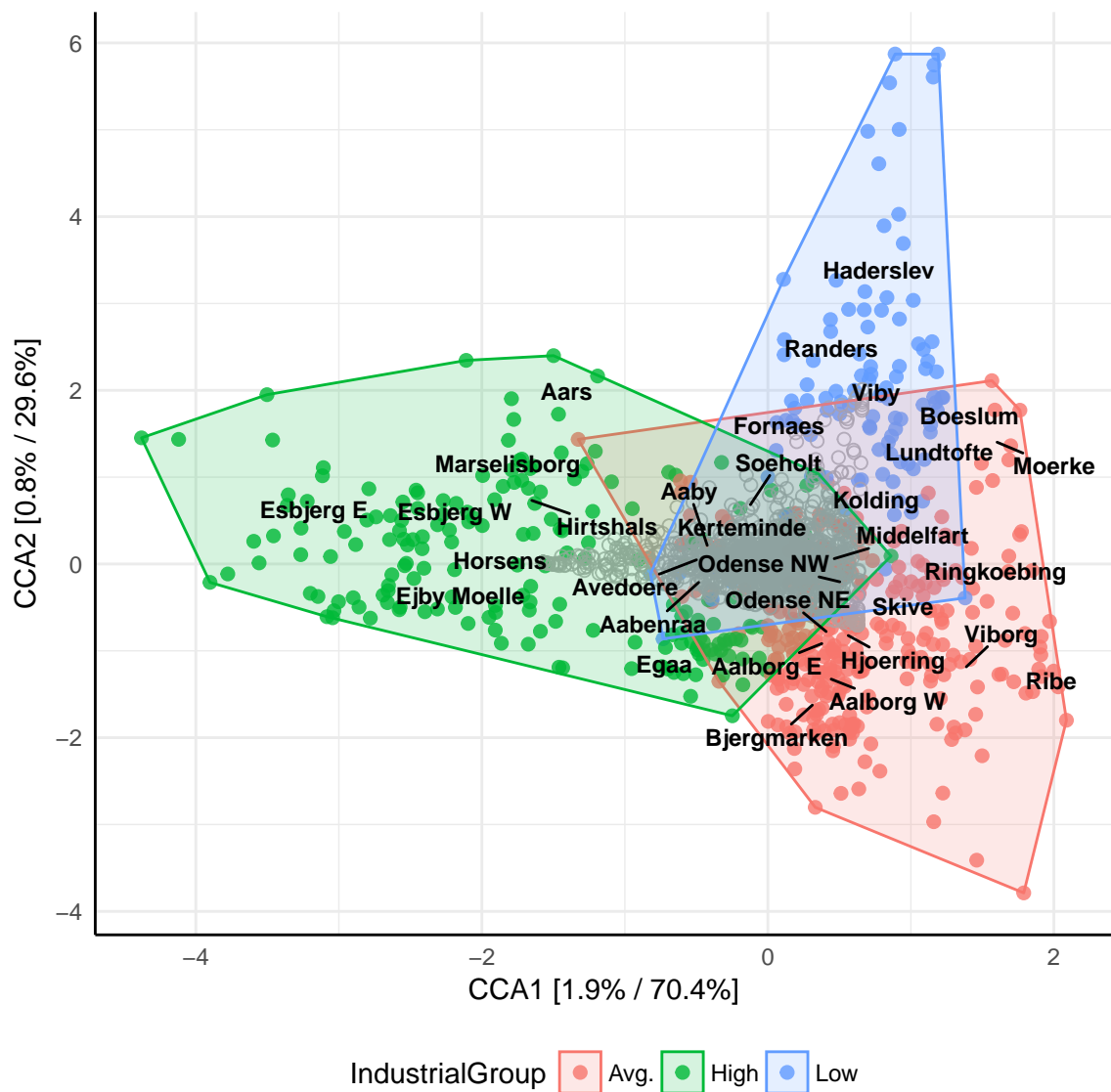


Figure 6.6: Industrial, hellinger - updated 14/3-15:20

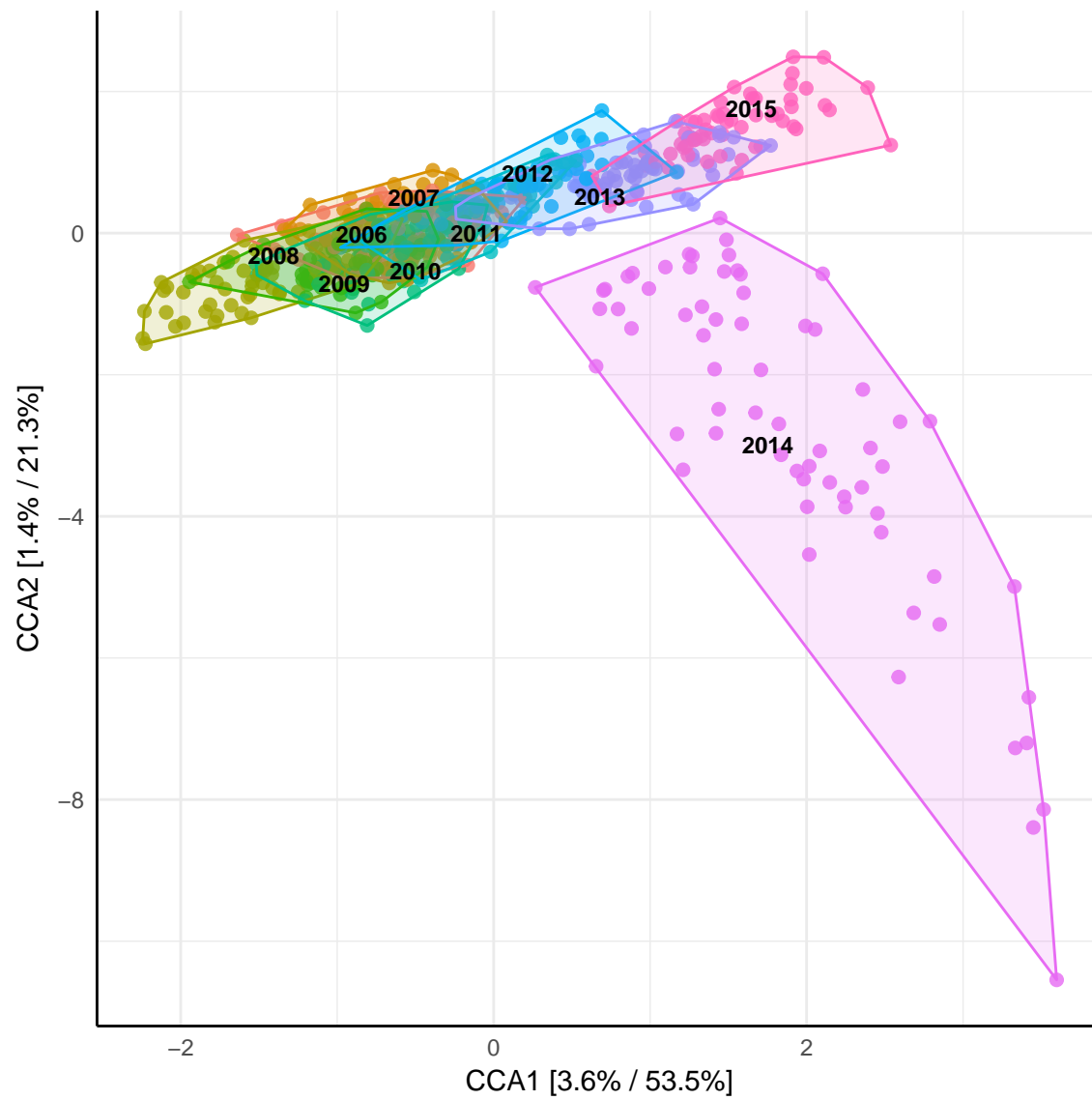


Figure 6.7: Stability

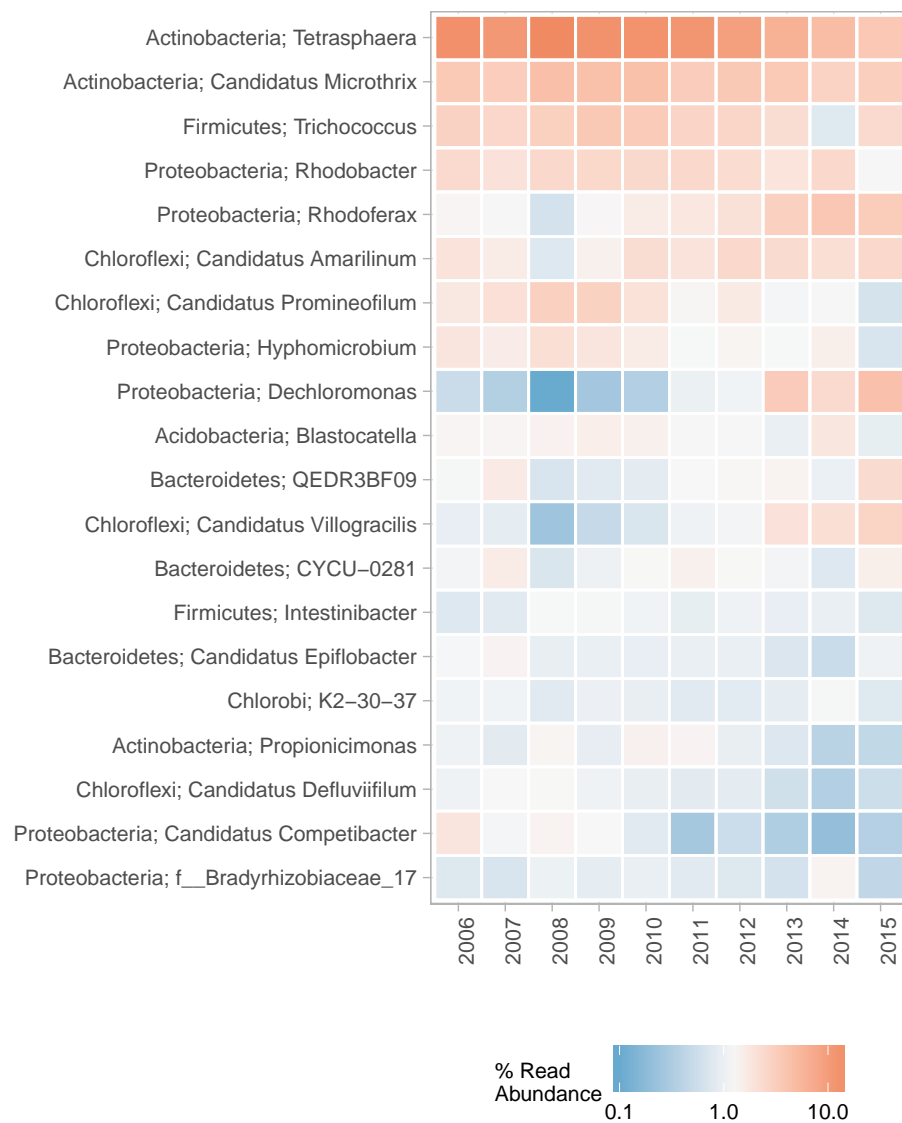


Figure 6.8: Stability

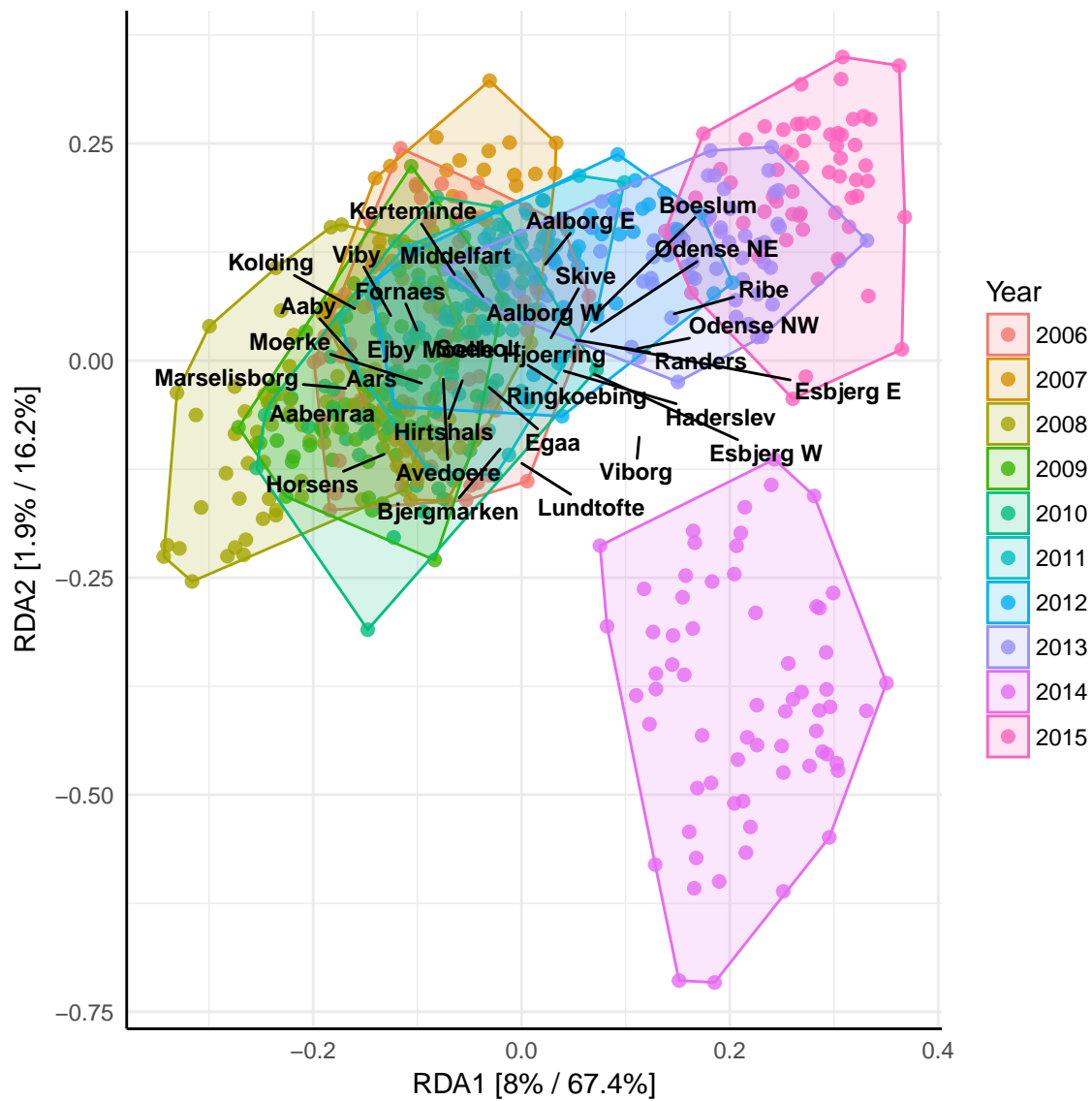


Figure 6.9: Stability

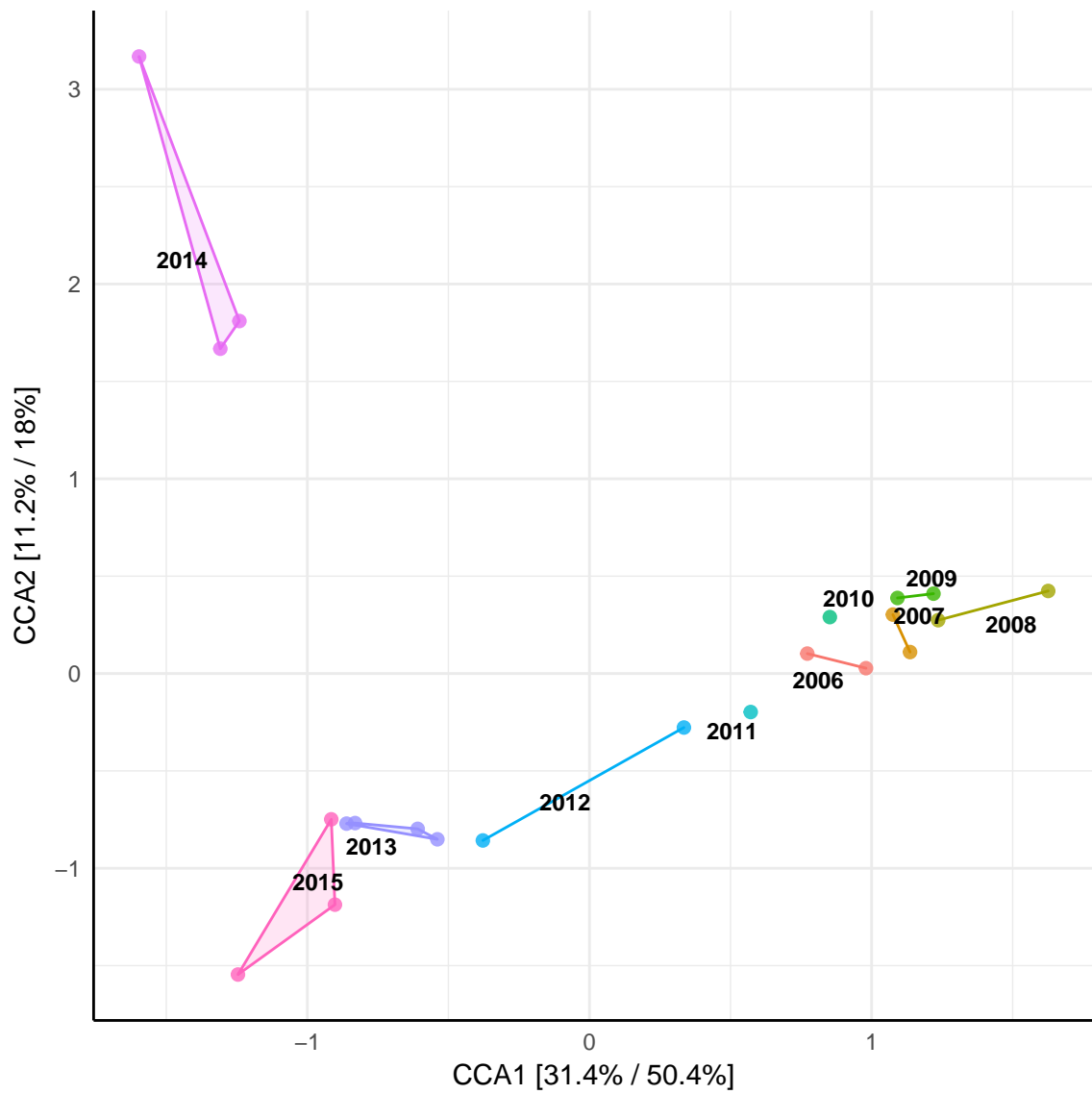


Figure 6.10: Stability Viborg

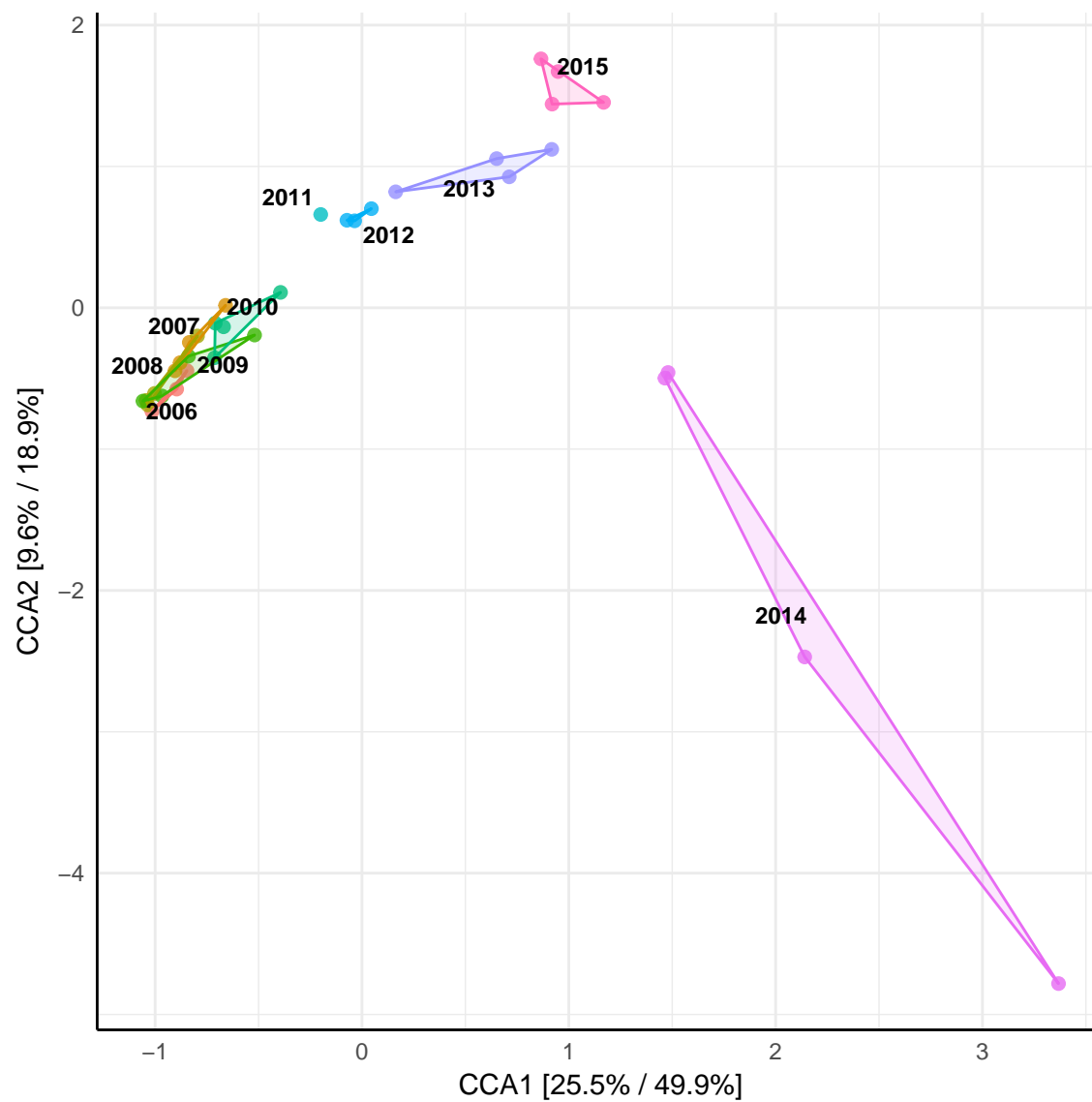


Figure 6.11: Stability Ejby Moelle

## 7. Discussion

test





## 8. Conclusion

If we don't want Conclusion to have a chapter number next to it, we can add the `{-}` attribute.

### **More info**

And here's some other random info: the first paragraph after a chapter title or section head *shouldn't be* indented, because indents are to tell the reader that you're starting a new paragraph. Since that's obvious after a chapter or section title, proper typesetting doesn't add an indent there.



## A. Characteristics of the WWTPs

Table A.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

- 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](http://doi.org/10.1016/S0065-2504(08)60183-X)
- 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>
- 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>
- 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>
- 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=6ZB0A-iDviQC>

- 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>
- 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>
- 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](http://doi.org/10.1016/0022-2496(66)90017-4)
- 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>