Predicting fungal community composition based on soil properties

---

Master Thesis in Biostatistics (STA495)

by

Natacha Bodenhausen
97-424-444

supervised by

Prof. Dr. Reinhard Furrer

Zurich, July 2019

Predicting fungal community composition based on soil properties

Natacha Bodenhausen

Version August 27, 2019

Contents

Acknowledgments

First of all, I would like to thank Professor Reinhard Furrer (UZH) for his support during this master thesis. I am very grateful for his patience and his enlightened advice at all the stages of this process. Craig Wang, Professor Furrer's PhD student, also offered some useful tips about R and R Markdown and provided feedback on an earlier version.

I am grateful to Dr. Paul Mäder (FiBL) who gave me the time to work on my thesis. I would also like to thank Dr. Klaus Schläppi (Agroscope and UniBE) and Professor Marcel van der Heijden (Agroscope and UZH) for having the original idea of this project.

At Agroscope, I would like to thank Julia Hess who is charge of the field sampling campaign and inoculation with arbuscular mycorrhizal fungi; Alain Held who did the DNA extraction, PCR and library preparation; Susanne Müller for the soil respiration data; Andrea Bonvicini for the microbial biomass data; and Diane Bürge for the other soil parameters. Finally, I would like to acknowledge the help of Andrea Patrignani (Functional Genomics Center Zurich), who is responsible for the Pacbio sequencing, and Jean-Claude Walser (Genetic Diversity Center), who performed the bioinformatic analysis. Finally, I would like to thank Dominika Kundel (FiBL) for her feedback on an earlier draft of this thesis.

On a personal side, I am grateful to my husband, Lukas Elmiger, who took care of our two children, Alice and Clara, so that I could have more time to write my thesis during week-ends.

Last but not least, I am grateful to the Gebert Rüf Stiftung for funding the project which generated this data.

Natacha Bodenhausen

July 2018

Chapter 1

Introduction

1.1 Why sustainable agriculture?

The world population is increasing (seemingly) "exponentially". This population growth has been accompanied by a concomitant increase in agricultural output, thanks to the so-called "green revolution", which introduced chemical pesticides, mineral fertilizers, and new varieties. However, the green revolution contributed to several environmental problems including pollution and soil erosion. In particular, application of excess fertilizers has led to eutrophication of rivers and lakes, which can lead to excessive algal growth, oxygen depletion and fish death. Moreover, nitrogen fertilizers are energetically costly to produce; they are derived from ammonia whose production necessitates high temperature and pressure. In addition, phosphate fertilizers are mined; known reserves of rock phosphate are expected to become depleted in the next 50–100 years (Cordell *et al.*, 2009). Therefore, we need to find alternatives.

1.2 How sustainable agriculture?

One answer to these problems is called sustainable agriculture, also called conservation agriculture. Novel practices include minimal mechanical soil disturbance, permanent soil cover, and species diversification (Food and Agriculture Organization of the United Nations, 2019). However, yields of sustainable agriculture are lower compared to intensive agriculture. To compensate reduction of yields when reducing the amount of fertilizer, several approaches are possible. For example, farmers can use alternative soil management

such as no-till or reduced tillage in order to improve living conditions for soil microbes participating in key soil functions such as decomposition and nutrient and carbon cycling. A different approach is the direct inoculation with microbes that can better access other resources of nutrients in the soil.

1.3 Arbuscular mycorrhizal fungi

One example of beneficial microbes are arbuscular mycorrhizal fungi (AMF). These are fungi of the phylum Glomeromycota that colonize the inside of plant roots where they form characteristic arbuscules (Figure 1.1). AMF form a symbiosis with about 80% of plant species, where the fungi provide the plants with nutrients, in particular phosphorus (P), in exchange for carbohydrates. When inoculation is successful, plants colonized with mycorrhiza are larger compared to plants grown in the absence of mycorrhiza (Figure 1.1). However, inoculation success is highly variable. Hoeksema *et al.* (2010) performed a meta-analysis of the response to inoculation with mycorrhizal fungi and identified host plant and nitrogen (N) fertilization as the most important factors. On the other hand, Zhang *et al.* (2019) found no evidence for an effect of fertilization on inoculation success with a more recent meta-analysis. By contrast, (Bender *et al.*, 2019) recently shown that phosphorus fertilization was an important factor for inoculation success.

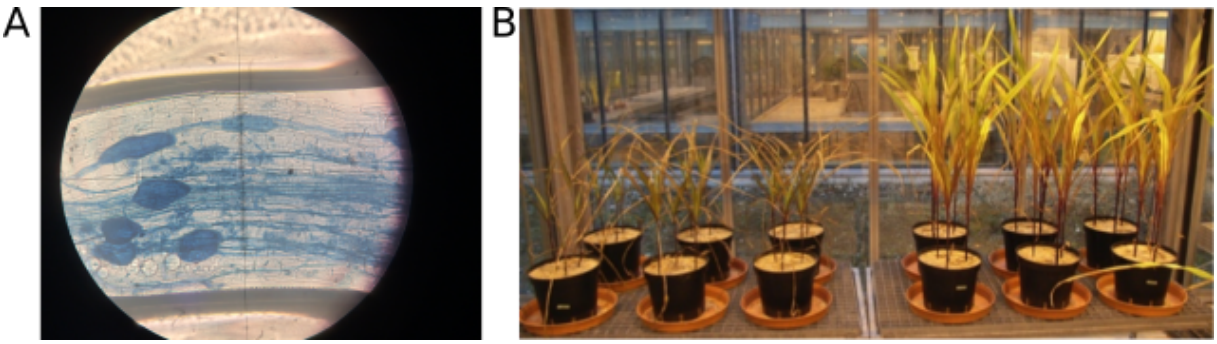

Figure 1.1: (A) Plantago roots colonized with AMF (Photo by J. Hess). (B) Maize plants grown in absence (left) or presence (right) of AMF in pots filled with field soil without fertilizer addition (Photo by Dr. F. Bender).

1.4 Goal of this thesis

AMF inoculate is costly to produce and to apply to crops; therefore, we need to be able to predict whether the inoculation will be a success. Previous research has shown that inoculation success depends on soil properties, but results were variable. Soil analyses are relatively cheap; moreover, farmers regularly measure classical soil properties in order to receive federal subsidies, so the data is readily available. In this thesis, we will attempt to predict AMF community composition from soil properties. We hypothesize that AMF inoculate will better establish in a soil when related species are already present. We call this approach "soil microbiome diagnostics". Based on the results of the soil analysis, we will be able to recommend inoculation with AMF in order to reduce fertilizer application and promote plant growth.

Chapter 2

From soil samples to data

This chapter briefly outlines the field and laboratory procedures used to obtain the data analyzed in this thesis. The goal of this thesis is to understand the relationship between physical, chemical, and biological properties of the soil and the fungal community that inhabits those soils; there are thus two sets of data. The soil physical, chemical, and biological data was obtained using traditional methods from soil science while the fungal community data was obtained using methods from molecular biology and bioinformatics tools.

2.1 Soil sampling

We sampled soils of 22 maize fields in a radius of 30 kilometers around Zurich. Exact GPS coordinates of those fields are available but not provided here for confidentiality reasons. These fields were chosen for a series of inoculation experiments with arbuscular mycorrhizal fungi (AMF). Sampling was performed with a tool called a hand auger, which removes a small cylinder of soil at a depth of 20 cm. Sampling was repeated along a transect on each field in order to collect about 4 kilograms of soil. Back in the lab, the soil was sieved with a 2 mm sieve to remove stones and large plant debris and to homogenize the samples. Finally, each sample was split in different fractions to be processed by several laboratories specialized in the different analyses. The subsamples for nitrogen analysis and DNA extraction were stored at –20°C while the other subsamples were stored at 4°C until processing. The soil samples were analyzed at two laboratories, Labor für Boden- und Umweltanalytik (lbu), a private company also used by farmers for soil analysis, and

Agroscope, which is affiliated with the Federal Office for Agriculture.

2.2 Soil parameters

Soil parameters, i.e., soil properties, are traditionally classified into three categories: physical, chemical and biological variables, see Table 2.1.

Physical properties include soil texture, which is the proportion of sand, silt and clay particles (from largest to smallest). Soil texture was determined qualitatively by Labor für Boden- und Umweltanalytik with the finger test whereas Agroscope used a quantitative method based on wet sedimentation fractionation. Water holding capacity is defined as the maximum amount of water retained in the soil, which is measured by comparing the weight of the fully saturated soil sample with the weight of the dried sample. Soil structure is a measurement of the degree of aggregation.

Chemical properties include pH, which is directly related to hydrogen (H) ion concentration ($\text{pH} = -\log(\text{H}^+)$), with $\text{pH} = 7$ called neutral, $\text{pH} < 7$ acidic, and $\text{pH} > 7$ basic. Soil pH is considered a very important parameter because it affects nutrient cycling and determines whether the soil is suitable for plant growth (Blume *et al.*, 2015).

The three most important nutrients for plant growth are nitrogen (N), phosphorous (P), and potassium (K), they are called macronutrients. The amount of nitrogen available for plant growth is measured by quantifying nitrate (NO_3) and ammonia (NH_4) and summing those measurements into a variable called Nmin. Calcium (Ca), sodium (Na) and magnesium (Mg) are also considered macronutrients. Furthermore, other elements important for plant growth are needed at much smaller concentrations and are thus called micronutrients. These include manganese, iron, copper, zinc, and boron. All these nutrients can be extracted from the soil with various methods based on different solvents that have varying affinity for the elements.

Some nutrients are positively charged (for example Ca^{2+}, Mg^{2+}, K^+, ${NH_4}^{2+}$), so they

Table 2.1: Soil physical, chemical, and biological variables

physical	chemical	biological
soil texture	pH	respiration (SIR)
water holding capacity	extractable nutrients (N, P, K, ...)	biomass (C and N)
soil structure	cation exchange capacity (CEC)	
	soil organic matter (humus)	

bind to negatively charged surfaces of the soil by electrostatic forces. The cations can be easily exchanged for other ions; this process is called ion exchange. Cation exchange capacity (CEC) is defined as the sum of the concentrations of sodium (Na), potassium (K), magnesium (Mg), calcium (Ca), and hydrogen (H). According to Blume *et al.* (2015), aluminum is usually included in CEC calculation but this ion was not measured by Agroscope. Base saturation (BS) is the fraction of CEC occupied with base cations. The equation used at Agroscope for BS is $(q_{\mathrm{Ca}} + q_{\mathrm{K}} + q_{\mathrm{Mg}} + q_{\mathrm{Na}})/\mathrm{CEC} \times 100$, where q_i represents the concentration of exchangeable ions in moles per kg. At Agroscope, the CEC was determined in a buffered solution of pH 7 and is thus called potential cation exchange capacity.

Humus plays an important role for soil fertility thanks to its negative charges, which increases the cation exchange capacity. Moreover, humus stabilizes the soil structure and provides a source of nutrients for the soil fauna and microbes. Humus comes from decomposing plant and animal remains and is therefore also called soil organic matter. Humus is quantified by first determining the total organic carbon (TOC), which is measured by combustion and determination of the liberated CO_2. TOC is multiplied by a constant (usually 1.724) to obtain the humus content (Blume *et al.*, 2015).

Soil organisms play an important role in soil functions, such as decomposition of organic matter, structure formation, and nutrient cycling. The most commonly measured biological properties are soil respiration and microbial biomass. Soil respiration can be measured by substrate-induced respiration. First, a soil sample is saturated with water; the sample is then incubated with a substrate, for example glucose, in a closed bottle containing NaOH during 72 hours. The CO_2 dissolves in NaOH and the amount of CO_2 is quantified by titration. Quantification of microbial biomass is determined with a method called chloroform fumigation extraction. The soil samples are fumigated with chloroform during 24 hours; the chloroform kills the cells and solubilizes organic C and N, which are then extracted with potassium sulfate (K_2SO_4) and quantified by titration.

2.3 Fungal community

In this thesis, we used a DNA-based method to analyze the fungal community. When analyzing microbial communities, several questions can be addressed. Who is present in

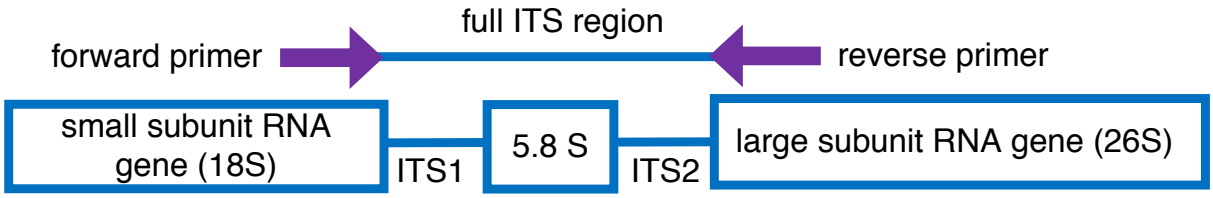

Figure 2.1: Internal transcribed spacer region (ITS) with PCR primers.

the sample (what is the taxonomy of the members of the community)? How many of each species are present (what is the abundance of the different members of the community)? The DNA of the samples was extracted with the soil DNA extraction kit from Macherey-Nagel and then a marker gene was amplified by polymerase chain reaction (PCR). To study fungal community composition, the marker gene most often used is the internal transcribed spacer (ITS) region (Lindahl *et al.*, 2013), which separates the genes encoding the small subunit (18S) and the large subunit (28S) of the ribosome. The ITS region is composed of two variable spacers called ITS1 and ITS2 surrounding the 5.8S gene (Figure 2.1). These ribosomal genes (18S, 5.8S and 28S) are conserved across the phylogenetic tree thus providing ideal targets for PCR primers. On the other hand, ITS1 and ITS2 do not code for a gene and as a result, they are more variable because of decreased selective pressure on non-gene coding DNA; for this reason, ITS1 and ITS2 may allow classification down to the species level.

In this thesis, we chose primers ITS1F (Gardes and Bruns, 1993) and ITS4 (White *et al.*, 1990) which amplify the full ITS region. Each PCR product was barcoded in a second PCR reaction, following Herbold *et al.* (2015), so that each DNA sequence can later be assigned to a sample (Figure 2.2). After purification, the DNA concentration of the PCR products was quantified and the PCR products were pooled into one DNA library in equimolar fashion. Finally, the DNA was sequenced at the Functional Genomics Center Zurich (FGCZ) by single molecule real-time (SMRT) sequencing also called PacBio sequencing. The library is prepared for sequencing and then placed inside a SMRT Cell, which is a nano-fabricated device (or chip) with tiny wells where the real-time sequencing is detected. The yield from the current SMRT sequencing technology (so-called Sequel) is about 300,000 sequences per cell. In this study, the yield from one cell was too low, therefore, the same library was sequenced in a second cell. The data from both sequencing cells was combined before the bioinformatic analysis.

The sequencing data was then transferred to the server from the Genetic Diversity

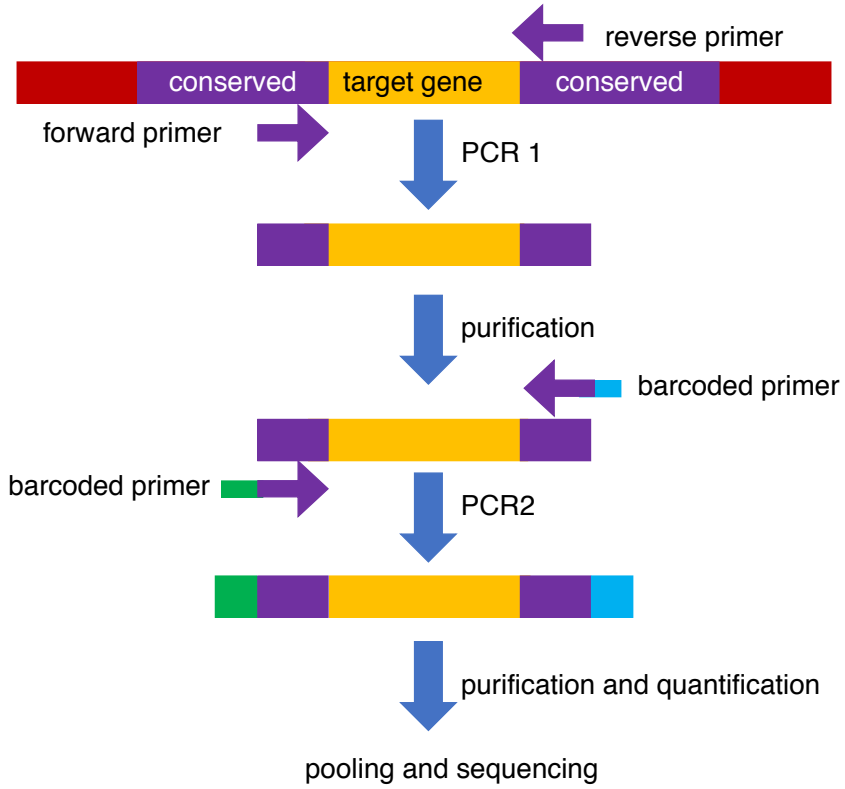

Figure 2.2: Library preparation with two-step barcoding.

Center (GDC). The different steps of the bioinformatic analysis are organized in so-called pipelines, here we used *USEARCH* (Edgar, 2013). First, sequences which did not meet certain quality thresholds were trimmed: for example, sequences with long stretches of homopolymers or ambiguous bases, sequences shorter or longer than set parameters. Remaining sequences were then assigned to samples using the barcodes (this step is called demultiplexing). The next step with *USEARCH* is called denoising, which is equivalent to removing sequencing errors. Finally, sequences were clustered into operational taxonomic units (OTU). These are groups of sequences similar to each other at a certain threshold; typically 97% similarity is said to correspond to the species level (Schloss and Handelsman, 2005). One representative sequence for each OTU was used to predict its taxonomy with *SINTAX* (Edgar, 2016) using the *SILVA* database (Abarenkov *et al.*, 2010). The results of the classifier depends on the chosen confidence value for *SINTAX*. In this thesis, we chose a cutoff of 0.7, which resulted in 33% of the sequences unclassified at the phylum level. Sequences were very rarely classified down to the species level, probably because the databases are not yet complete and many microbes have not been sequenced.

The output from the bioinformatic analysis is two tables: the OTU table, which is a matrix with the number of sequences per sample per OTU, and a taxonomy table,

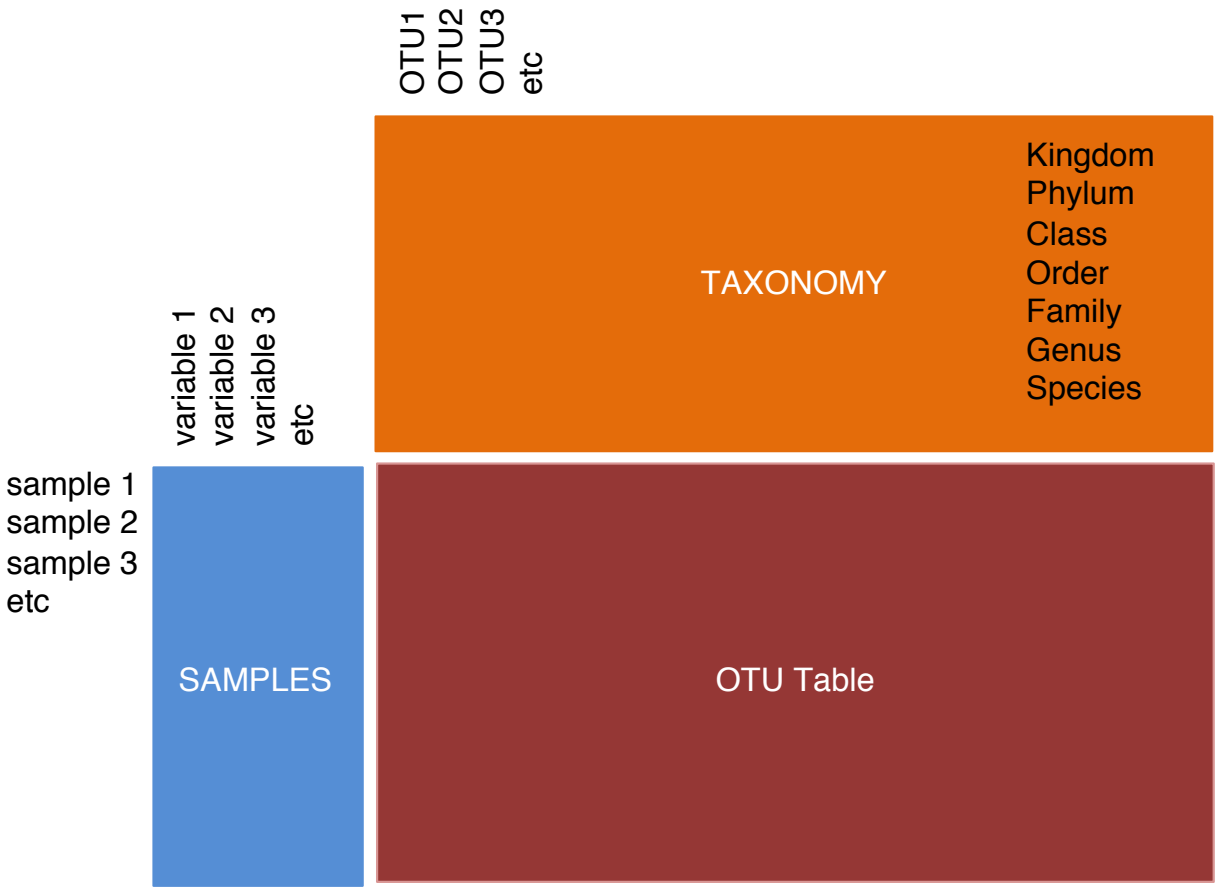

Figure 2.3: Microbiome data is usually organized in three tables.

which provides the information about the classification for each OTU (Figure 2.3). The taxonomy table provides the answer to the first question (what is the taxonomy of the members of the community), while the OTU table provides the answer to the second question (what is the abundance of the different members of the community).

The two tables from the bioinformatic analysis are analyzed together with a third table, containing all the information about the samples, including the barcodes used for the PCR and the environmental data. The three tables can be conveniently combined into one R object with the R package `phyloseq` (McMurdie and Holmes, 2019a).

Chapter 3

Soil data and variable reduction

This chapter focuses on the soil variables. We analyzed the soil physical, chemical and biological properties of 22 agricultural soils in two different laboratories, Labor für Boden- und Umweltanalytik (lbu) and Agroscope. We obtained 50 variables; however, there was some redundancy because several variables were analyzed by both laboratories and most nutrients were analyzed according to different protocols. In addition, some variables were found to be a function of the other variables. In this chapter, the number of variables is reduced from 50 to 9 with different methods, first by understanding the type of data and then using literature, multivariate analysis and, finally, using the R package `varrank`.

3.1 Reducing the number of variables by understanding the data

In this section, we will describe how the number of variables was reduced from 50 to 17 variables based on the information about the data. The variables are named by the measurement followed by the laboratory (`agro` stands for Agroscope and `lbu` stands for Labor für Boden- und Umweltanalytik) separated by an underscore. Further, for the chemical element analysis, the name of the method used was abbreviated and placed between the name of the element and the name of the laboratory. For example, `phosphorus_H2O_lbu` stands for P analysis extracted with H_2O by Labor für Boden- und Umweltanalytik.

3.1.1 Physical variables

Soil texture is defined by the size of the particles, from the largest particles called sand, then silt, and finally clay. In this study, two measurements for silt and clay were obtained: Labor für Boden- und Umweltanalytik used the finger test, while Agroscope used a quantitative method, therefore the measurements from Agroscope were retained.

We measured two other physical variables: aggregation and water holding capacity (WHC). Aggregation is an important physical property related to soil structure. Higher aggregation is an indicator of stronger biological activity; for example, soil microbes that produce mucus or soil fauna such as earthworms that promote the formation of organo-mineral compounds in their guts (Blume *et al.*, 2015). WHC is also an important soil property for agriculture because plants benefit from a soil with larger WHC by having access to more water in times of drought. Therefore, both these variables were kept in the data set.

3.1.2 Chemical variables

pH is directly related to the concentration of hydrogen: $\mathrm{pH} = -\log(\mathrm{H}^+)$. Field 6 (F06) was identified as outlier by plotting pH and `hydrogen_agro` (Figure 3.1). Since the concentration of hydrogen is taken into account for the calculation of cation exchange capacity (CEC) and base saturation (BS), the outlier was replaced with the estimate from the linear regression and CEC and BS were calculated based on the corrected data.

Five variables are related to nitrogen (N) content. Ammonium and nitrate were measured at both laboratories. The samples measured by Agroscope were stored at –20°C before extraction, while the samples measured by Labor für Boden- und Umweltanalytik were stored at room temperature; we assume that the values measured by Agroscope are closer to the values in the field. Nmin is the sum of ammonium and nitrate; therefore we keep only Nmin.

Five variables are related to phosphorus (P) content. Total P was measured at Agroscope, but this variable is not relevant for agriculture because most of this phosphorus is not available for plants nor microbes. In addition, P was measured with four different methods by Labor für Boden- und Umweltanalytik. Traditionally, Anglo-Saxons countries favor the Olsen method (Blume *et al.*, 2015), which extracts P with carbonate. The other methods are named after the solvent used to extract the phosphorus. In Switzerland,

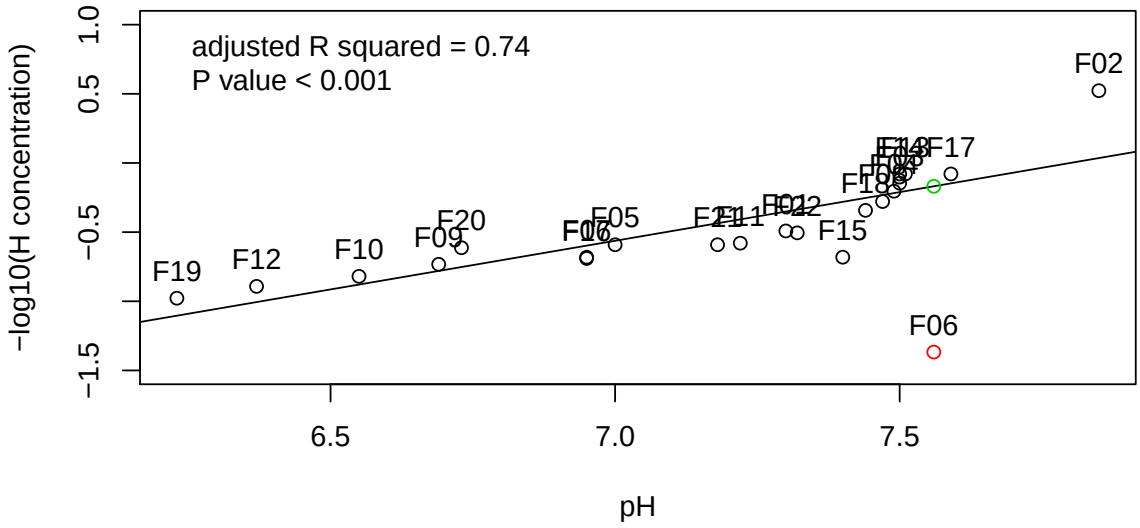

Figure 3.1: Regression of hydrogen concentration and pH. The outlier (Field 06) is shown in red. The estimate from the regression is shown in green.

extraction with CO_2-saturated water is recommended for cereal crops, while H_2O extraction is recommended for fruits and vegetables (Richner *et al.*, 2017). The extraction with the solvent containing EDTA (the full recipe is: 0.5 M ammonium-acetate, 0.5 M acetic acid, 0.025 M ethylenediaminetetraacetic acid) gives variable results depending on the type of soil; for example, in calcareous soil, results are not reproducible and therefore this method is not recommended (Richner *et al.*, 2017). In our sample set, the five measurements for P correlated strongly with each other (Figure 3.2), so the analysis of extracts with CO_2-saturated water is retained. Similarly, we also keep measurements extracted with CO_2-saturated water for potassium, whereas for magnesium, the extraction with $CaCl_2$ is kept, and for calcium and sodium the extractions with H_2O.

In addition to macronutrients, several micronutrients were also measured by Labor für Boden- und Umweltanalytik: manganese, iron, copper, zinc, and boron. Micronutrients are also called trace elements because plants only need minute amounts of them. A meta-analysis showed that there was no effect of micronutrients on AMF inoculation success (Zhang *et al.*, 2019) so micronutrients were dropped from the reduced data set.

CEC is the sum of exchangeable cations (sodium, potassium, magnesium, calcium, hydrogen) (Blume *et al.*, 2015). At Agroscope, the measurement is done is a buffered solution of pH 7, is is therefore called potential CEC. BS is defined by Agroscope as

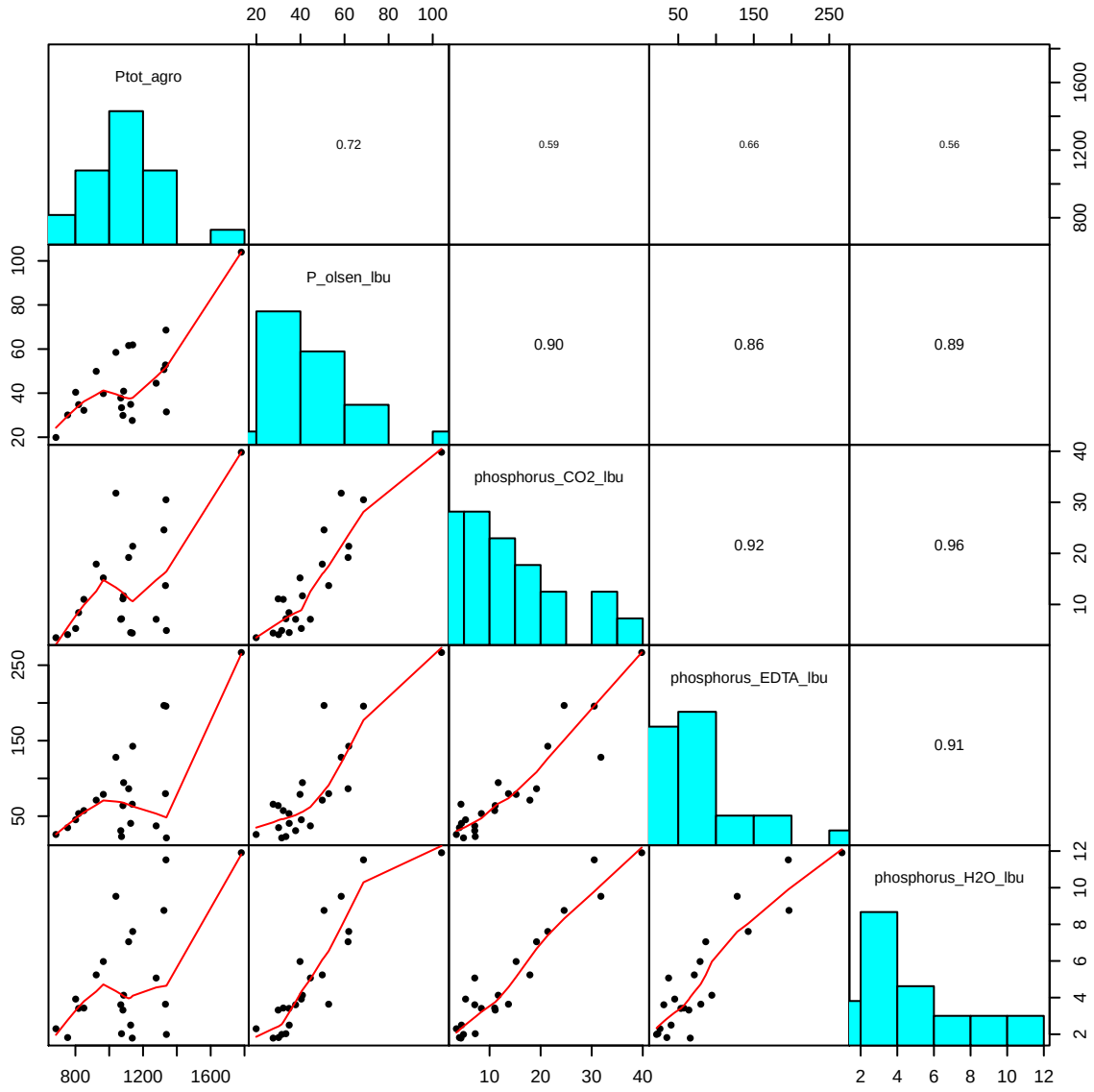

Figure 3.2: Scatterplots of five phosphorus measurements. Histograms are shown in the diagonal. Correlations are shown in the upper panel. Scatterplots are shown in the lower panel with red line LOWESS smoother.

the concentration of sodium, potassium, magnesium, and calcium over CEC. Both these variables are kept in the reduced data set.

Three variables are related to organic matter: organic carbon, only provided by Agroscope, and humus, provided by both laboratories. Organic carbon and humus are redundant because humus is obtained from organic carbon measurement by multiplying by a constant, usually 1.724 (Blume *et al.*, 2015). The two values for humus measured by both laboratories are very similar, so `humus_agro`, measured by Agroscope is kept in the reduced data set.

3.1.3 Biological variables

Microbial biomass can be measured by quantifying either the amount of organic carbon (Cmic) and organic nitrogen (Nmic). In our study, both measurements correlated quite well; therefore, we keep only Cmic. In addition, respiration is also kept in the reduced data set because it is an important soil function.

3.1.4 SOLVITA variables

Labor für Boden- und Umweltanalytik sells a new product called SOLVITA, which includes three variables: *vast_lbu*, *respiration_lbu*, *slan_lbu* and as well as a fourth variable, a fertility index which is a combination of those three variables (*fertility_lbu*). SLAN stands for "Solvita Labile Amino-Nitrogen" and is said to report the amount of organic nitrogen reserves. Solvita respiration, called C-Burst, measures CO_2 release after re-wetting, similarly to the respiration assay from Agroscope; however, instead of titration, the CO_2 is quantified with a digital color reader. From the SOLVITA variables, only *vast_lbu* is retained in the reduced data set because it is a physical variable which was not measured otherwise. The other two SOLVITA variables are directly related to variables measured by Agroscope, and since the protocol used by Agroscope is more quantitative, *vast_lbu*, *respiration_lbu* and *fertility_lbu* variables were dropped.

3.2 Reducing the number of variables with literature and varrank

At the end of the first section, 17 variables were selected (Table A.1). In this section, the number of variables is further reduced to 9. The first step in the selection is based on the literature. The second step is performed with principal component analysis and the *varrank* approach implemented by package *varrank* (Kratzer and Furrer, 2018).

3.2.1 Literature research

Soil texture is an important soil parameter because it correlates with many other physical properties such as load-bearing capacity, pore size distribution, and air storage capacity, all of which are important for plant growth (Blume *et al.*, 2015). These other parameters

Table 3.1: Soil texture data classified with United States Department of Agriculture classification

	class name	number of samples
Cl	clay	1
SiCl	silty clay	0
SaCl	sandy clay	0
ClLo	clay loam	10
SiClLo	silty clay loam	0
SaClLo	sandy clay loam	2
Lo	loam	9
SiLo	silty loam	0
SaLo	sandy loam	0
Si	silt	0
LoSa	loamy sand	0
Sa	sand	0

were not assessed here, in part because they are difficult to measure. Therefore, we decide to keep soil texture in the final data set. However, since the three variables of soil texture together with humus add up to 100%, only silt, sand and humus were retained.

Soil texture data is often represented with a ternary plot with `clay` at the top, `sand` on the left and `silt` on the right (Figure 3.3). Here we used the `soiltexture` package (Moeys, 2018) to plot and classify the data. Most of the soil samples analyzed in this study are found in the middle of the ternary plot, indicating that they are not extreme soils (Figure 3.3). Most of the fields were clay loam or loam fields with three exceptions, Field 15, a clay soil, and Field 5 and Field 18, sandy clay loam (Table 3.1). `Humus` content can be plotted on top of the soil texture ternary plot, here with the size of the bubbles (Figure 3.3). The mean for humus in the 22 fields was 3.07%. Field 15 was the field with the highest SOM. Two other fields also had a humus content larger than 5% (Field 15 and Field 16), while only one field had a SOM content smaller than 1.5% (Field 20).

Soil pH is relevant for agriculture because of its direct and indirect effect on plant growth. For example, `pH` affects `CEC`, and thus the availability of macronutrients (Blume *et al.*, 2015). In addition, pH is also known to be a major driver of bacterial community composition (Lauber *et al.*, 2009), as well as AMF community composition (Van Geel *et al.*, 2018). For all theses reasons, pH is retained in the final data set.

We chose to keep `phosphorus_CO2_lbu` and `Nmin_agro` in the data set because they are the major plant nutrients. In addition, P is also important for AMF colonization. Countless studies with pot experiments have shown that AMF colonization is inhibited

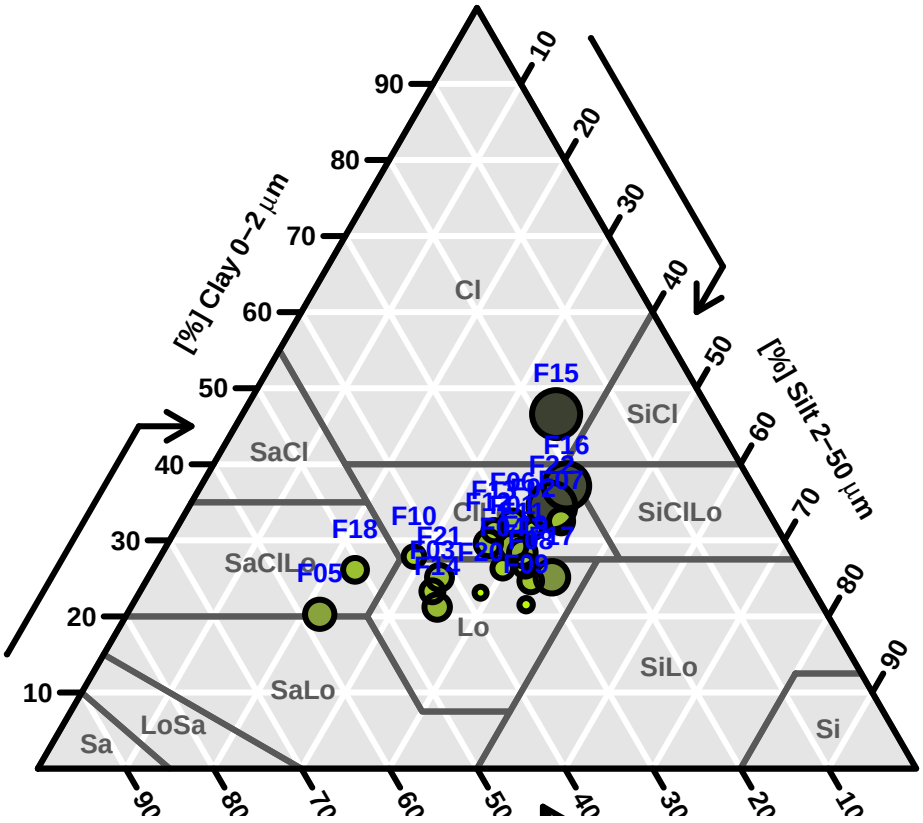

Figure 3.3: Ternary plot of soil texture. The size of the bubbles represent the concentration of soil organic matter. The name of the fields is indicated next to the bubble in blue.

by high phosphorus concentration (for example, see Breuillin *et al.*, 2010). More recently, Bender *et al.* (2019) have shown that AMF colonization is affected by P content in the soil in eight field experiments in Switzerland. In addition, N fertilization is an important predictor of AMF inoculation success (Hoeksema *et al.*, 2010).

3.2.2 Multivariate analysis

We used principal component analysis (PCA) to examine the relationship between the 17 variables remaining after the first step in variable reduction. The goal of PCA is to reduce the number of variables by transforming the original variables with a set of linear combinations to a new set of principal components (PCs), retaining as much variability

as possible. The PCs are uncorrelated and ordered, so that the first PC accounts for the most variation, the second PC accounts for the maximal proportion of the remaining variance, etc. The number of components needed to summarize a given data set can be chosen based on different methods (Everitt and Hothorn, 2011): here we decided to keep the number of components that explain at least 70% of the total variation of the original variables. In our study, we found that the first three components explain 73% of the total variation of the original variables, so we present the results with PC1, PC2 and PC3 (Figure 3.4).

PC1 is negatively correlated with many variables typical of soils with high fertility including *humus*, *Nmin*, *respiration*, *CEC*, *Cmic*, *clay*, *aggregation*, and *WHC*. By contrast, PC1 is positively correlated with *sand*. PC1 can be interpreted as measure of water availability. PC2 is positively correlated with *pH*, *phosphorus*, *potassium*, and *BS*, while it is negatively correlated with sodium and magnesium. PC2 can be interpreted as a measure of nutrient availability. Finally, PC3 is positively correlated with *phosphorus*, *magnesium*, and *potassium* while it is negatively correlated with silt. PC3 can also be interpreted as a measure of nutrient availability. In conclusion, PCA further confirmed the importance of *pH*, *silt*, *sand*, *humus*, *Nmin*, and *P*, which are the variables identified to be important based on the literature.

3.2.3 Varrank

Finally, we used *varrank::varrank()* for variable selection. This package allows for variable ranking with respect to a subset of important variables (Kratzer and Furrer, 2018). Here, the *variable.important* argument receives as input the list of variables selected based on the literature. We chose to return only three variables (argument *n.var=3*). The result from *varrank* was that *Ca*, *Mg* and *WHC* are the most relevant variables after removing the redundancy shared with the selected variables.

3.3 Summary

In the first part of the chapter, the number of variables was reduced from 50 to 17 based on the information about the data. In the second part of the chapter, the number of variable was further reduced with two methods. Based on literature, six variables were

chosen: sand, silt, humus, pH, P, Nmin. Based on `varrank` analysis, `calcium`, `magnesium` and `WHC` were selected. The final soil data set is thus reduced to 9 variables (Table 3.2). The variables are renamed to remove the name of the laboratory in order to simplify subsequent plotting and analysis.

Table 3.2: 9 variables retained in the variable selection. "P-Testzahl" unit, where 1 P-Testzahl = 0.155 mg P/kg soil (Richner *et al.*, 2017).

	sand	silt	humus	P	Nmin	pH	Mg	Ca	WHC
unit	%	%	%	P-Testzahl	mg/kg	unitless	mg/kg	mg/kg	unitless
F01	30.3	38.2	3.17	24.6	38.5	7.3	16.9	124.6	0.607
F02	27.3	39.8	2.15	4.1	24.9	7.85	11.9	208.3	0.617
F03	42.3	32.4	2.58	11.1	34.1	7.5	6.4	188.1	0.582
F04	33.1	38.8	2.41	11.7	26.7	7.49	15.5	167.8	0.627
F05	55.8	21.2	3.44	15.2	23.9	7	7.5	75.38	0.678
F06	28.9	36.8	2.86	17.9	25.4	7.56	13.9	164.6	0.684
F07	23.5	42.1	2.83	21.4	25.4	6.95	16.9	118.4	0.593
F08	30.7	42.6	2.66	31.8	34.7	7.47	8.8	162.8	0.579
F09	33.1	44.1	1.59	5.3	9.2	6.69	10.8	52.81	0.471
F10	42.2	28.2	2.43	4.5	24.6	6.55	9.9	96.85	0.558
F11	29.7	39.7	3.12	8.4	33.2	7.22	12.9	122.8	0.657
F12	32.9	35.4	2.96	7.1	34.3	6.37	22	104.6	0.732
F13	31.6	35.4	2.51	4.4	22.7	7.51	5.2	194.5	0.62
F14	42.6	33.8	2.99	19.2	40.9	7.5	8.6	218.2	0.623
F15	16.7	33.7	5.62	4.9	47.8	7.4	15.7	308.1	1.068
F16	20.1	39.4	5.3	39.8	62.5	6.95	19	226.2	0.899
F17	27.8	44.2	3.82	7.2	51.5	7.59	8.9	266.9	0.805
F18	49.4	22.4	2.8	11	29.6	7.44	19.2	121	0.64
F19	30.4	41.1	2.55	7.1	30.4	6.23	10.9	88.87	0.644
F20	37.5	38.3	1.43	3.5	9.5	6.73	7	56.24	0.465
F21	40.5	32.2	2.92	30.5	37.5	7.18	13.2	121.9	0.618
F22	22.9	39	5.47	13.7	60.6	7.32	12.7	266.4	0.873

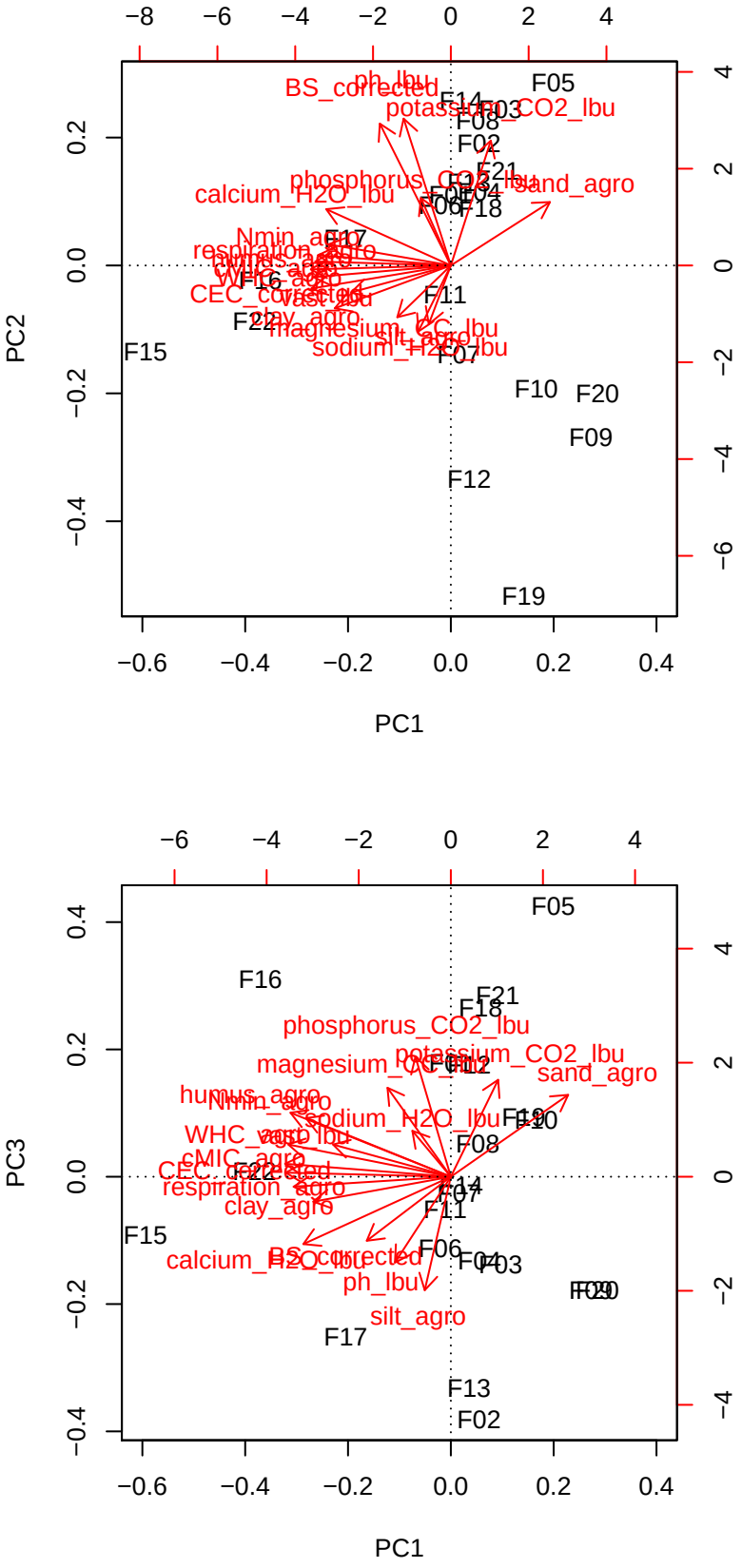

Figure 3.4: Principal component analysis of 17 variables remaining after first step of variable selection. Top: Biplot of the first and second principal components. Bottom: Biplot of the first and third principal components.

Chapter 4

Methods to study ecological communities

In this chapter, different methods to assess alpha diversity (within sample) and beta diversity (between samples) are presented. The concept of alpha and beta diversity was first defined by Whittaker in 1960 and further refined by him in 1972 (Whittaker, 1972). According to Whittaker, alpha (α) diversity is the local diversity, beta (β) diversity the spatial diversity and gamma (γ) is the regional diversity. They are related with this equation: $\beta = \gamma/\alpha$.

In addition, different ordination methods are presented in this chapter. First, unconstrained methods are explained and limitations of those methods for community data are described. Finally, three methods of constrained ordination, also called canonical ordination, are described.

4.1 Alpha diversity

Alpha diversity is the diversity within sample. The most intuitive index of alpha diversity is q, the richness or number of species (in the case of microbiome data, OTUs). However, this number is actually the observed richness, because richness increases with the size of the sample. In traditional ecology, this was the size of a lake or a quadrant; in the case of microbiome data, it is the number of sequences per sample. Therefore, indices that are not dependent on the size of the sample are preferred. The most used alpha diversity

index is called Shannon diversity:

$$H = -\sum_{i=1}^{q} p_i \log p_i,$$

where p_i is the proportional abundance of species i and q represents the number of species. This index takes into account both the number of species and the evenness of the species frequency distribution (Borcard *et al.*, 2018). Evenness is another important alpha diversity measure as it describes the shape of the rank abundance plot, where the species are ranked by order of decreasing abundance on the x-axis and the log-transformed abundance is plotted on the y-axis. In the most extreme case, a few species are dominant and the others are rare, this community would be called uneven. At the other extreme (which never happens in nature), all species have the same abundance and evenness equals one. The maximum of H is when all species are presented at equal abundances:

$$H_{max} = -\sum_{i=1}^{q} \frac{1}{q} \log \frac{1}{q} = \log q.$$

The most commonly used index of evenness is called Pielou evenness:

$$J = H/H_{max} = H/\log q.$$

However, this index is biased because it is dependent on species richness (Borcard *et al.*, 2018).

Jost (2007) proposed to use the "number equivalents" of diversity indices. They have several advantages, including that they are more easily interpretable and are preferred for linear modelling (Borcard *et al.*, 2018). These numbers are also called Hill's diversity numbers. In this notation, N_0 is species richness. The number equivalent of Shannon diversity is:

$$N_1 = \exp(H), \tag{4.1}$$

and the number equivalent of Shannon's evenness is

$$E_1 = N_1/N_0, \tag{4.2}$$

which is also called Sheldon evenness (Sheldon, 1969).

4.2 Beta diversity

Beta diversity indices can be computed with presence-absence data, see Koleff *et al.* (2003) for the definition of 24 commonly used indices. An example is the Sørensen-Dice coefficient:

$$\text{SDC} = \frac{A + B - 2J}{A + B},$$

where A and B are the numbers of species at the two sites and J is the number of species that occur at both sites.

Alternatively, dissimilarity indices can be computed with quantitative data. The percentage difference, also called Bray-Curtis dissimilarity, is the most popular among ecologists:

$$D_{12} = \Big(\sum_{k=1}^{q} |y_{1k} - y_{2k}|\Big) \Big/ \Big(\sum_{k=1}^{q} (y_{1k} + y_{2k})\Big), \tag{4.3}$$

where y_{1k} is the counts for site 1 of species k and y_{2k} is the counts for site 2 of species k. Both these indices are bounded between zero (the two sites share all the species) and one (no species in common).

4.3 Unconstrained ordination

Multivariate analysis is often used to visualize and explore changes in community composition. In this section, three commonly used methods will be described: principal component analysis (PCA), correspondence analysis (CA), and principal coordinate analysis (PCoA).

4.3.1 Principal component analysis

PCA is generally not well adapted to the study of species abundance because species rarely respond to environmental gradients in a linear fashion (Borcard *et al.*, 2018). Moreover, the so-called "double zero" problem occurs for PCA because PCA relies on Euclidean

distance. The equation for Euclidean distance between sites 1 and 2 across q species is:

$$D_{12} = \sqrt{\sum_{k=1}^{q} (y_{1k} - y_{2k})^2},$$

where y_{1k} is the counts for site 1 of species k and y_{2k} is the counts for site 2 of species k. If a species is present at two sites, this indicates that the sites are similar and that they provide conditions that allow both species to survive. On the other hand, if a species is absent at two sites (double zero), this could be due to many different causes (Borcard *et al.*, 2018): for example, the species is present but was not detected (rare OTUs), the conditions at both sites are not optimal (conditions could be either below the optimum or above the optimum), the species has not reached the site even though the conditions are suitable, the species's niche is occupied by another equivalent species. All in all, absence of one species at two sites may not mean that the two sites are similar because the absence could be due to a different reason at the two sites.

To avoid the double zero problem and the issue of non-linear response to environmental gradients, certain transformations can be used. Legendre and Gallagher (2001) showed that both the Chord and the Hellinger distance are suitable for community data. To calculate the Chord distance, the site vectors are normalized to 1 before calculating the Euclidean distance (Borcard *et al.*, 2018); for details, see Legendre and Gallagher (2001). The Hellinger distance is equivalent to the Euclidean distance of data transformed with the Hellinger transformation:

$$y'_{ik} = \sqrt{\frac{y_{ik}}{y_{i+}}}, \tag{4.4}$$

where y_{ik} is the counts for site i of species k and y_{i+} is the sum of all species at site i (in the case of microbiome data, it is the number of sequences per sample).

The results from PCA can be presented with a biplot, which is a graphical representation of the data in two dimensions; the "bi" in the word does not stand for those two dimensions but because the plot displays both the variances and the covariances of the variables as well as the distances between samples, for details see Everitt and Hothorn (2011). The samples are traditionally shown with points, the distance between points indicates the distance between the samples, so samples which are more similar to each

other are clustered closer to each other. On the other hand, the variables are often shown with vectors, the length of the vector indicates the variance of that variable, while the angle between two vectors reflect the correlation between them, so if the angle is small, the two variables are strongly correlated.

4.3.2 Correspondence analysis

CA is recommended when species display a uni-modal relationship with the environmental gradient. Furthermore, CA is useful to address the question of whether certain species occur at certain sites (Ramette, 2007). Correspondence analysis is an iterative method. First, the data is transformed into a χ^2 distance matrix and then singular value decomposition is performed. The equation for χ^2 distance between sites 1 and 2 across q species is:

$$D_{12} = \sqrt{y_{++}} \sqrt{\sum_{k=1}^{q} \frac{1}{y_{+k}} \left(\frac{y_{1k}}{y_{1+}} - \frac{y_{2k}}{y_{2+}} \right)^2},$$

where y_{++} is the sum of abundances, y_{+k} is the sum of counts for species k, y_{1+} is the sum of counts for site 1 and y_{2+} is the sum of counts for site 2. Legendre and Gallagher (2001) point out that the inner part of the χ^2 distance is the same as the Euclidean distance calculated with relative abundance and weighted by the inverse of the species sum. Therefore, a rare species, with small y_{+k}, will contribute to a greater extent to the sum of squares. On the other hand, χ^2 distance is not sensitive to double zeros, so there is no need for additional transformation. Note that in correspondence analysis, the overall variance is called inertia.

4.3.3 Principal coordinate analysis

PCoA, also called multidimensional scaling (MDS), can be used to explore differences in beta diversity because it can take any distance or dissimilarity matrix, for example Bray-Curtis dissimilarity (Equation (4.3)). Similarly to PCA and CA, PCoA also produces a set of orthogonal axes, with eigenvalues that provide information about how much variation is explained by each axis. However, since the dissimilarity matrix only has the site names, a biplot cannot be drawn directly from the output of the procedure but the weighted

averages need to be calculated *a posteriori*, for details, see Legendre and Gallagher (2001). With non-Euclidean distances, PCoA may produce negative eigenvalues (Borcard *et al.*, 2018). Several corrections have been developed that solve this problem. The Lingoes correction adds a constant to the squared dissimilarities while the Caillez correction add a constant to the dissimilarities.

4.4 Canonical ordination

Three methods are described in this section: redundancy analysis (RDA), canonical correspondence analysis (CCA) and distance-based redundancy analysis (dbRDA). The family of methods is called constrained ordination.

4.4.1 Redundancy Analysis

RDA combines regression and principal component analysis (Borcard *et al.*, 2018). RDA works with a matrix $\boldsymbol{Y}$ of centered response data (in our case the community data) and a matrix $\boldsymbol{X}$ of centered explanatory variables (in our case the environmental data). First, each centered y variable is regressed on the explanatory matrix $\boldsymbol{X}$ and for each, the $\hat{y}$ (fitted values) is computed. All $\hat{y}$ vectors are combined in matrix $\hat{\boldsymbol{Y}}$. A test should be computed to test the relationship $\boldsymbol{Y} \sim \boldsymbol{X}$. If $\boldsymbol{X}$ variables explain the variation of $\boldsymbol{Y}$ more than random data would, a PCA of the matrix $\hat{\boldsymbol{Y}}$ is computed. As usual, the PCA yields eigenvalues (called here canonical) and a matrix $\boldsymbol{U}$ of canonical eigenvectors. The "site constraints (linear combination of constrained variables)" are computed with $\hat{\boldsymbol{Y}}\boldsymbol{U}$ (in `vegan` they are coded `'lc'`). The "site scores (weighted sums of site scores)" are computed with $\boldsymbol{Y}\boldsymbol{U}$ (in `vegan` they are coded `'wa'`). To summarize, the axes of RDA are a linear combination of the explanatory variables (Borcard *et al.*, 2018).

A triplot can be drawn to show the results of RDA for the first pair of axes. A triplot represents the sites and the species, both usually with points, and the environmental variables, usually with arrows.

4.4.2 Canonical Correspondence Analysis

CCA was first introduced by Ter Braak (1986), it combines regression with CA. The abundance of a species can be described by a response function which relates its abun-

dance to environmental scores (Zhang and Thas, 2016). The species scores from a CA estimate the optima, which are the values on the axis for which the maximum is obtained (Ter Braak, 1986). The ordination axis can be related to environmental variables by multiple regression. In CCA, the species optima and the regression coefficients are simultaneously estimated, for details see Ter Braak (1986). This is an iterative process: 1) start with arbitrary initial site scores, 2) the species scores are calculated by weighted averaging of the site scores, 3) the site scores are calculated by weighted averaging of the species score, 4) regression coefficients are obtained by multiple regression of the site scores on the environmental variables, 5) new site scores are calculated (fitted values of the regression), 6) site scores are centered and standardized, 7) stop when the new site scores are close to the site scores obtained with the previous iteration.

Since it preserves the χ^2 distance, CCA is well-adapted to species response curve which are usually not expected to be linear in response to the environmental gradient (Borcard *et al.*, 2018). One of the major drawbacks of this method is that rare species influence the results disproportionately; therefore, it is recommended to remove rare species before applying CCA.

4.4.3 Distance-based redundancy analysis

PCA relies on Euclidean distance while CA relies on χ^2 distance. However, many dissimilarity indices are better suited for comparison of community composition; collectively, these indices are called beta diversity indices (section 4.2), because they measure the diversity between samples. One example is the Bray-Curtis dissimilarity (Equation (4.3)). Several ordination methods were developed to work with those dissimilarity indices, including distance-based redundancy analysis and canonical analysis of principal coordinates.

dbRDA was initially developed for testing multispecies responses in multifactorial ecological experiments (Legendre and Anderson, 1999). First, a dissimilarity matrix is computed using any one of the beta diversity indices. Second, PCoA is applied to this matrix using the Lingoes correction for negative eigenvalues. Finally, RDA is run with those principal coordinates instead of the usual species data and with $\boldsymbol{X}$ (explanatory variables) as above. The output of the RDA (a matrix of principal coordinates) can be used for testing multivariate hypothesis in multifactorial experiments. In their initial

paper, Legendre and Anderson (1999) did not propose to do ordination with the results of the RDA. Later, dbRDA was developed as an ordination method. Since the dissimilarity matrix does not include the species, it was initially not possible to directly draw a triplot. However, species score can be added to the plot as weighted average.

A few years later, Anderson and Willis (2003) presented canonical analysis of principal coordinates (CAP). This method was specifically proposed as a constrained ordination method. The first two steps are similar (dissimilarity matrix followed by PCoA); however, the last step of CAP uses canonical discriminant analysis if there is any factor variable or canonical correlation analysis for quantitative variables. Another difference to dbRDA is that for CAP, the number m of axes needs to be chosen before the last step.

4.4.4 Permutation test

The function `vegan::anova.cca()` uses permutation test to assess the significance of the constraints for all three constrained ordination methods described in this chapter. The test statistic (called *pseudo-F*) is defined as follows (Borcard *et al.*, 2018):

$$F = \frac{SS(\hat{\boldsymbol{Y}}/m)}{RSS/(n-m-1)},$$

where n is the number of objects, m is the number of canonical eigenvalues (degrees of freedom), $SS(\hat{\boldsymbol{Y}})$ is the sum-of-squares of the table of fitted values (explained variation), and RSS is the residual sum of squares. A reference distribution of the chosen statistic (in our case the *pseudo-F*) is generated by randomly permuting the data (the rows) a large number of times (by default 999) and recomputing the statistic. The observed value is compared to the reference distribution and the p-value is the proportion of permuted values equal to or larger than the observed value. If this p-value is equal or smaller than the significance level α, the null hypothesis is rejected.

4.4.5 Partitioning of variance

In constrained ordination, the variance is partitioned into constrained and unconstrained. As explained by Borcard *et al.* (2018): "The constrained fraction is the amount of variance of the $\boldsymbol{Y}$ matrix explained by the explanatory variables". It is called R^2 and is similar to the R^2 from multiple regression For the same reason, this R^2 is biased: adding ex-

planatory variables inflates the R^2. Ezekiel's formula to adjust the R^2 is also valid for the multivariate case:

$$R^2_{adj} = 1 - \frac{n-1}{n-m-1}(1-R^2),$$

where n is the number of objects, m is the number of degrees of freedom (the number of quantitative explanatory variables). For CCA, variance is called inertia. The proportion of inertia explained by the explanatory variables is also called R^2; however, Ezekiel's formula cannot be used, so a bootstrap method has been developed to adjust the R^2 (Borcard *et al.*, 2018).

4.5 Summary

In this chapter, statistical methods to analyze microbial communities were discussed. Both alpha diversity and beta diversity indices were defined. Three methods of unconstrained and three methods of constrained ordination were briefly described.

The most useful R package for analysis of community data is called `vegan` (Oksanen *et al.*, 2019), which stands for ***veg***etation ***an***alysis. The methods of `vegan` were initially developed to analyze community data at the scale of plants or mites but the same methods are also appropriate to study communities at the scale of microbes.

Chapter 5

Description of the fungal community data

In this chapter, the fungal community data obtained with the DNA sequencing approach is described. First, rarefaction plots are produced to represent the sequencing effort. Second, changes in alpha diversity across the different fields are investigated. Third, the reasons for filtering of microbiome data are stated and community composition at the phylum level is described. Finally, multivariate analysis of community composition and beta diversity are shown.

5.1 Summary of sequencing results

To analyze the fungal community, the full ITS region was sequenced after amplification by PCR. Briefly, four DNA extractions were performed for each soil sample. Each DNA sample was barcoded during PCR. After pooling the 88 PCR products, the DNA library was sequenced with SMRT sequencing in two cells. The output from the sequencing center was two FASTQ files with a sum of 723,031 sequences. After filtering for quality and demultiplexing, 484, 694 sequences remained, ranging from 1956 to 7806 with a median of 5551 sequences per sample. These sequences were clustered into operational taxonomic units (OTUs) at 97% level, which typically correspond to the species level (Schloss and Handelsman, 2005). The end product from the bioinformatic analysis is an OTU table containing 88 rows (22 soil samples $\times$ 4 replicates) and 881 columns (OTU).

5.2 Rarefaction curves

To investigate whether the sampling effort was sufficient to characterize the diversity of the fungal community, rarefaction curves were plotted for each sample with `vegan::rarecurve()`. This is a plot of the expected number of species (here OTUs) as a function of number of individuals (here sequences). At regular steps along the x-axis, a sub-sample of the data is drawn and the number of OTUs (richness) is estimated with Hurlbert's equation (Borcard *et al.*, 2018):

$$E(q') = \sum_{i=1}^{q} \left[1 - \frac{\binom{n-n_i}{n'}}{\binom{n}{n'}} \right],$$

where $n' \leq (n - n_1)$, n_1 is the number of individuals in the most abundant species. This equation estimates the number q' of species in a sampling unit of n' individuals based on the real sampling unit containing q species, n individuals and n_i individuals belonging to species i. For Figure 5.1, a step of 200 (i) and a standardized sampling unit of 1000 (n') were chosen because they produced a smooth line and did not take too much computing time. The rarefaction curves are starting to level off around 4000 sequences; only six samples have less sequences. This indicates that the fungal community is well characterized with this sequencing effort. However, a few samples are found in the increasing part of the rarefaction curve, which indicates that the observed richness will be lower than the true richness.

5.3 Alpha diversity

Next, we investigated alpha diversity, which is the diversity within a sample (Whittaker, 1972). Species richness (q) can be computed with colSums(otuTable>0). Shannon diversity (H) can be calculated with `vegan::diversity()`, from which Hill's number was computed with Equation (4.1). Finally, evenness was computed with Sheldon's Equation (4.2).

The three indices differ across the different fields (Figure A.1). Species richness and Shannon diversity were lowest in Field 17 and highest in Field 10, whereas Sheldon evenness was lowest in Field 12 and highest in Field 15. Both Shannon diversity and Sheldon evenness plots indicate that one replicate for Field 14 is an outlier. This replicate is very

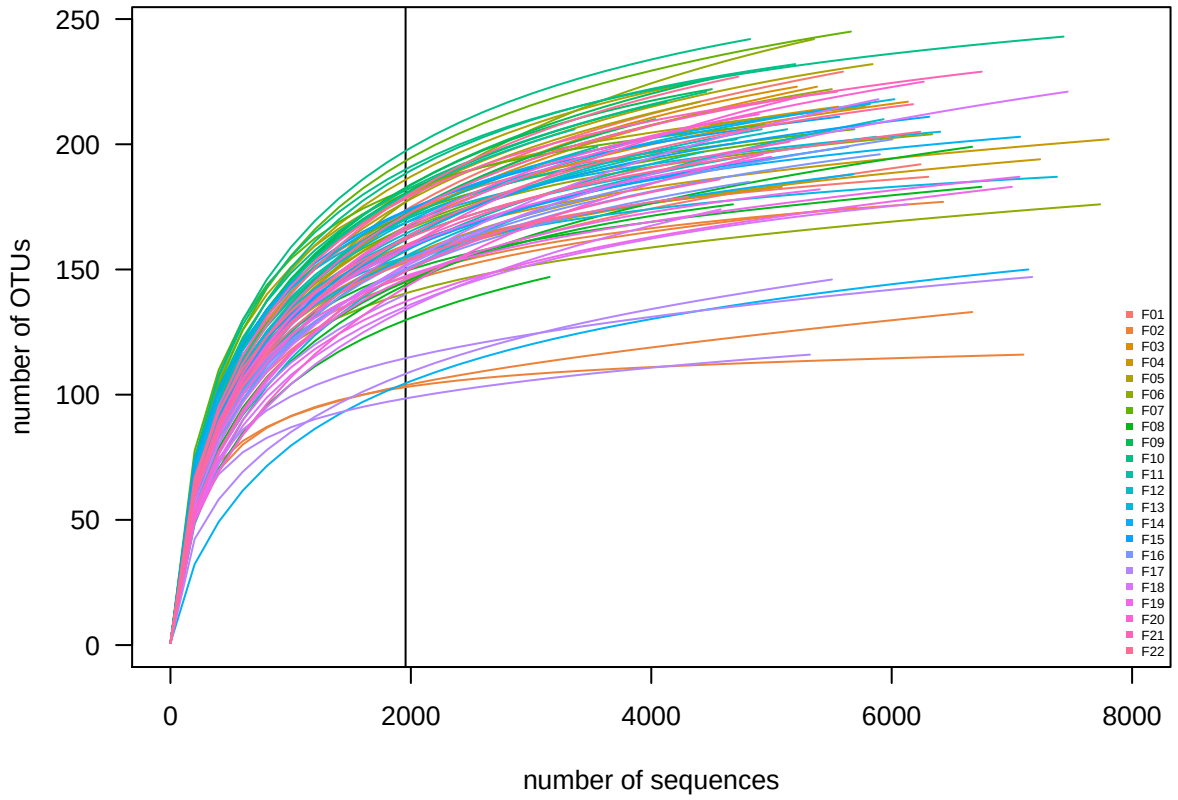

Figure 5.1: Rarefaction curves. The vertical line marks the sample with the smallest library size across all samples.

uneven, meaning that it is dominated by one OTU, OTU_10, which accounts for 4959 sequences out of 7137 sequences in that sample.

5.4 Filtering and transforming

Most OTUs are present in only a few samples, as is visible on the Figure A.2. This figure presents 9 panels for each of the phyla present in this data set as well as the OTUs which were not assigned any taxonomy at the phylum level. There are only a few OTUs for Entorrhizomycota, Glomeromycota, Mucoromycota and Olpidiomycota. Most OTUs are only present in a few samples, as can be seen from the heavy tail on the left of each plot.

Rare species can be truly rare, for example because they require specific environmental conditions to thrive. On the other hand, rare species can be an artifact from the sampling process. The first step in any amplicon library sequencing project is DNA extraction which is known to impact both diversity and composition of soil microbial communities (Martin-Laurent *et al.*, 2001). The second step is PCR which introduces many sources

Table 5.1: Summary of the data after filtering

	number of OTUs	number of sequences
before filtering	881	484694
after filtering	171	412311
percent kept	19	85

of bias: for example primers and polymerase that have different preferences for DNA sequences (Gohl *et al.*, 2016). In addition, rare species have too much influence on results from correspondence analysis (Legendre and Gallagher, 2001). For all these reasons, it is recommended to remove rare OTUs for the analysis of amplicon sequencing data. However, we do not filter the data for alpha diversity, following recommendations by `phyloseq` developers (McMurdie and Holmes, 2019b).

After experimenting with different filtering parameters, we chose to use the same filtering thresholds used by McMurdie and Holmes (2013) in the paper presenting `phyloseq`: "OTUs were trimmed that were not observed at least 3 times in 20% of samples". There is a trade-off between the number of OTUs which are retained and the number of sequences. The aim of the filtering step is to remove the long tail of rare OTUs while retaining the greatest number of sequences. After filtering, the number of OTUs was reduced to 171 while retaining 85% of the sequences as shown in Table 5.1.

Next the OTU table was transformed. The most common transformation is relative abundance or proportion, whereby the species counts are divided by the total number of sequences in each sample. Alternative transformations include $\log(1 + x)$ (Callahan *et al.*, 2016), where x is the species count, and Hellinger transformation, Equation (4.4), for PCA and RDA (Legendre and Gallagher, 2001).

5.5 Community composition

Figure 5.2 shows the composition of the community at the phylum level. After averaging the relative abundance of the four replicates for each field, the OTUs with the same taxonomy at the phylum level were merged with `phyloseq::tax_glom()`. Mortierellomycota and Ascomycota are the two major phyla followed by Basidiomycota and Chytridiomycota. In Field 12, Olpidiomycota are also quite abundant. In most fields, Mortierellomycota sequences are more abundant than Basidiomycota, except in Field 2, where the opposite pattern was observed.

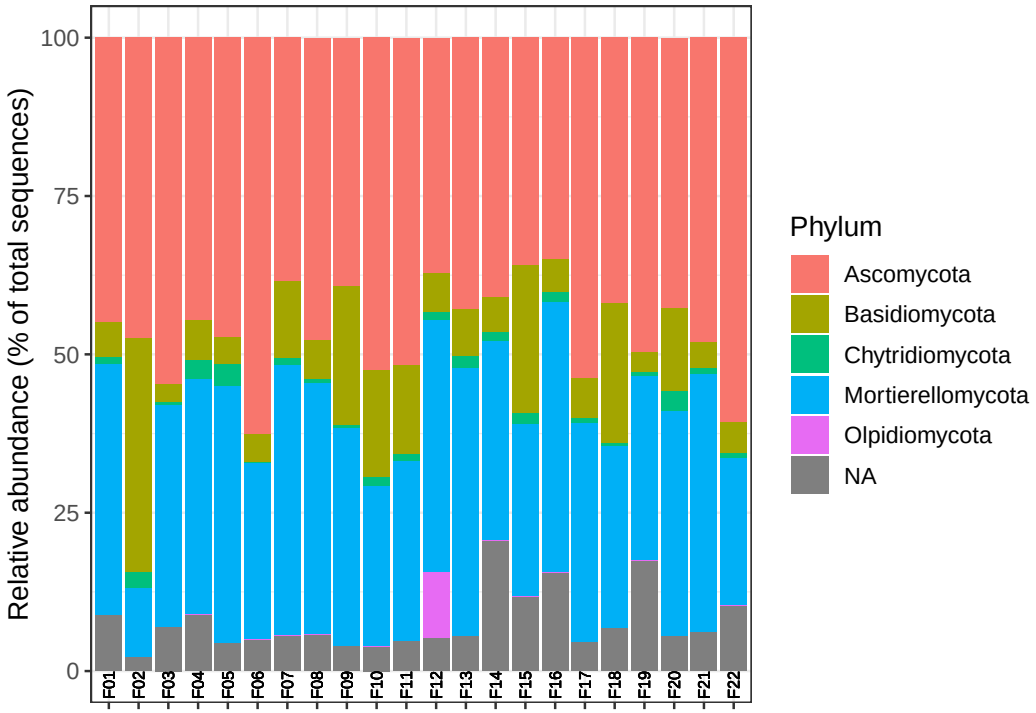

Figure 5.2: Community composition at the phylum level. Mean of four replicates was calculated before merging OTUs with the same phylum.

5.6 Unconstrained ordination

The aim of ordination is to represent the data with less dimensions, constructed so that the dimensions represent the main trends in the data. In this section, we present the results of three methods, namely PCA, CA and PCoA.

5.6.1 Principal component analysis

PCA was computed with `vegan::rda()` with the filtered and Hellinger transformed data set. The first component accounts for 14.2% of the variance while the second component accounts for 10.4% of the variance (Figure 5.3). The first component separates well most of the fields. The second component separates Field 2 from the others. Samples from Field 7 and Field 9 are close to each other, on the left of the plot, indicating that they have a similar species composition; similarly for samples from Field 12 and Field 19 at the top of the plot. Most of the OTUs are centered at the origin, meaning that only a few OTUs drive the differences between samples. Importantly, the four replicates of each soil generally cluster together.

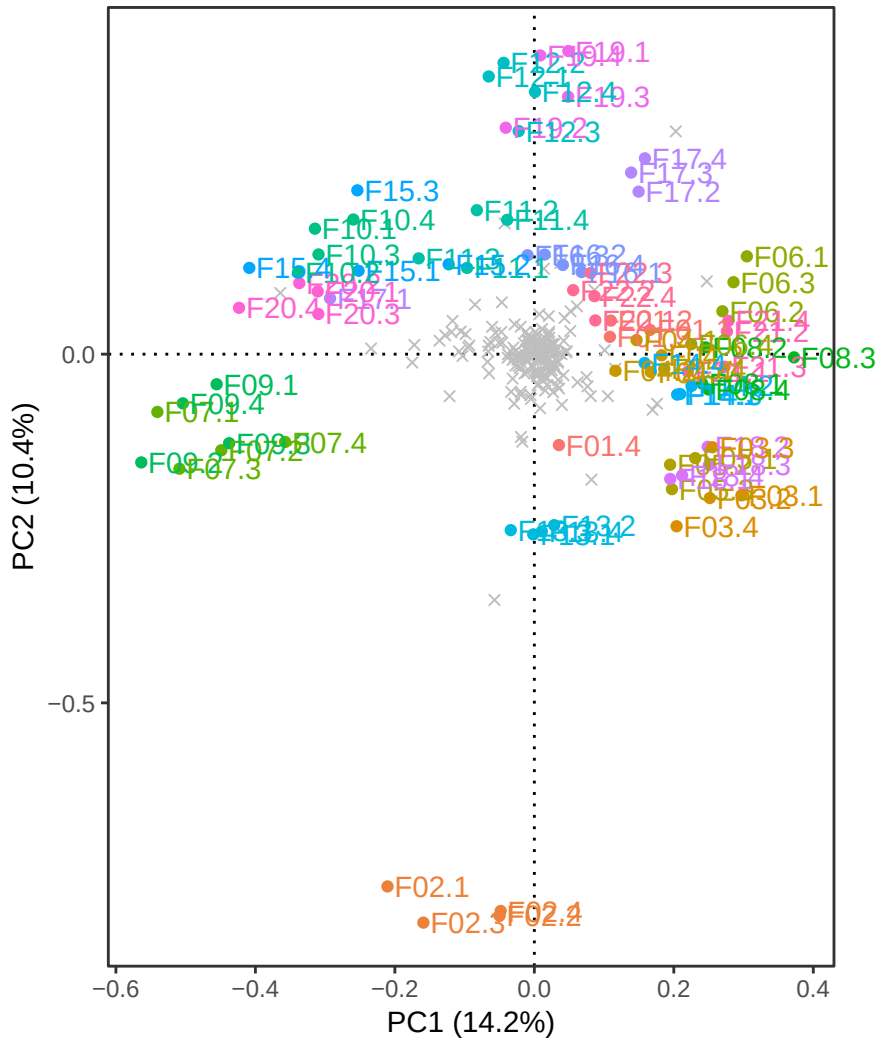

Figure 5.3: Biplot of PCA computed with filtered and Hellinger transformed data. The fields are shown as colored dots and the OTUs as grey crosses.

5.6.2 Correspondence analysis

CA was computed with `vegan::cca()` within `phyloseq` with the filtered and normalized data set (relative abundance). The first axis accounts for 13.7% of the inertia and the second axis accounts for 10.2% of the inertia. Most of the sites are grouped together, except two field sites: the first axis separates the four replicates from Field 2 from the others and the second axis separates two replicates of Field 12 from the others (Figure 5.4). This is mainly driven by the abundance of two OTUs: OTU52 which is very abundant in all 4 replicates of Field 2 (33.3% mean relative abundance) and which was assigned to Basidiomycota, and OTU1420 which is very abundant in two replicates from site Field 12 (34.8 and 6.4% relative abundance) and which was assigned to Olpidiomycota.

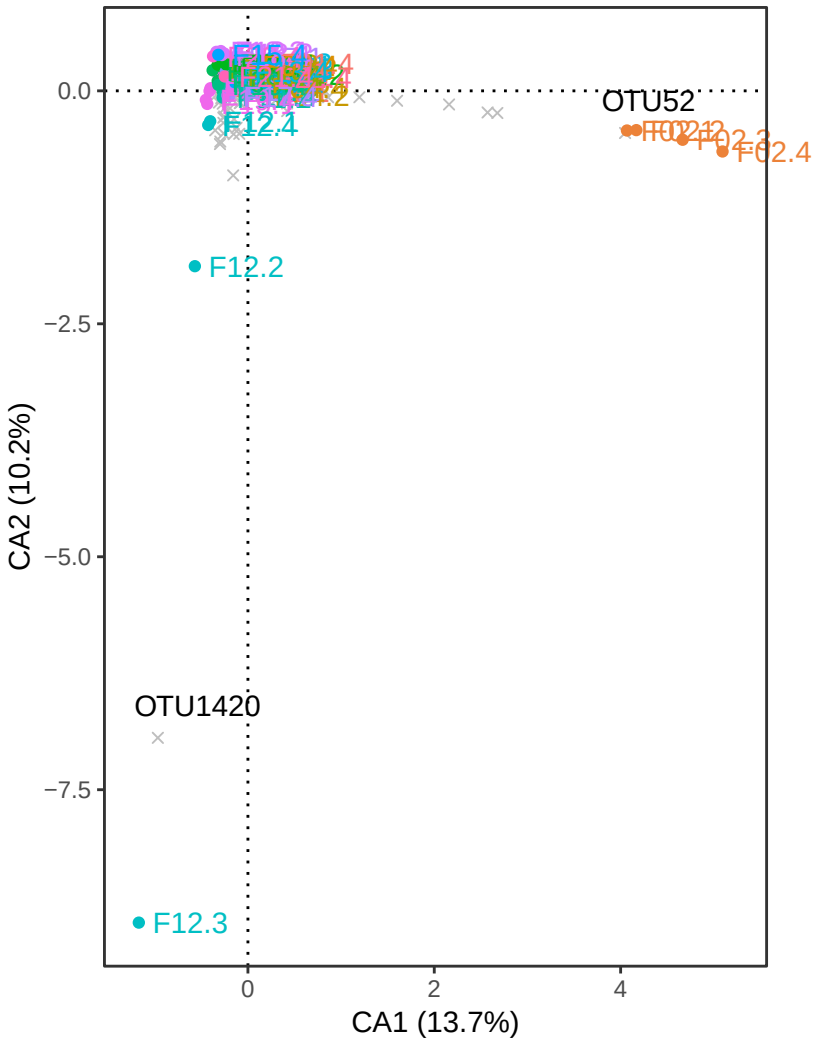

Figure 5.4: Biplot of CA computed with filtered and normalized data. The fields are shown as colored dots and the OTUs as grey crosses. The two most extreme OTUs are labeled with their name.

5.6.3 Principal coordinate analysis

First, the matrix of Bray-Curtis dissimilarity was computed with `vegan::vegdist()`; it is the the default `method`. Next, PCoA was performed with `cmdscale()`. The first axis accounts for 18.1% of the inertia and the second axis accounts for 12.6% of the inertia. Similarly to PCA, the four replicates generally cluster together (Figure 5.5), indicating that they are close in community composition. Field 7 and Field 9 are close together on the right of the plot, likewise for Field 12 and Field 19 at the bottom of the plot, which indicates that they share abundant species. The species score for the biplot need to be calculated as weighted averages, for details see Legendre and Gallagher (2001). The species on the PCoA are more spread out compared to PCA (Figure 5.3)

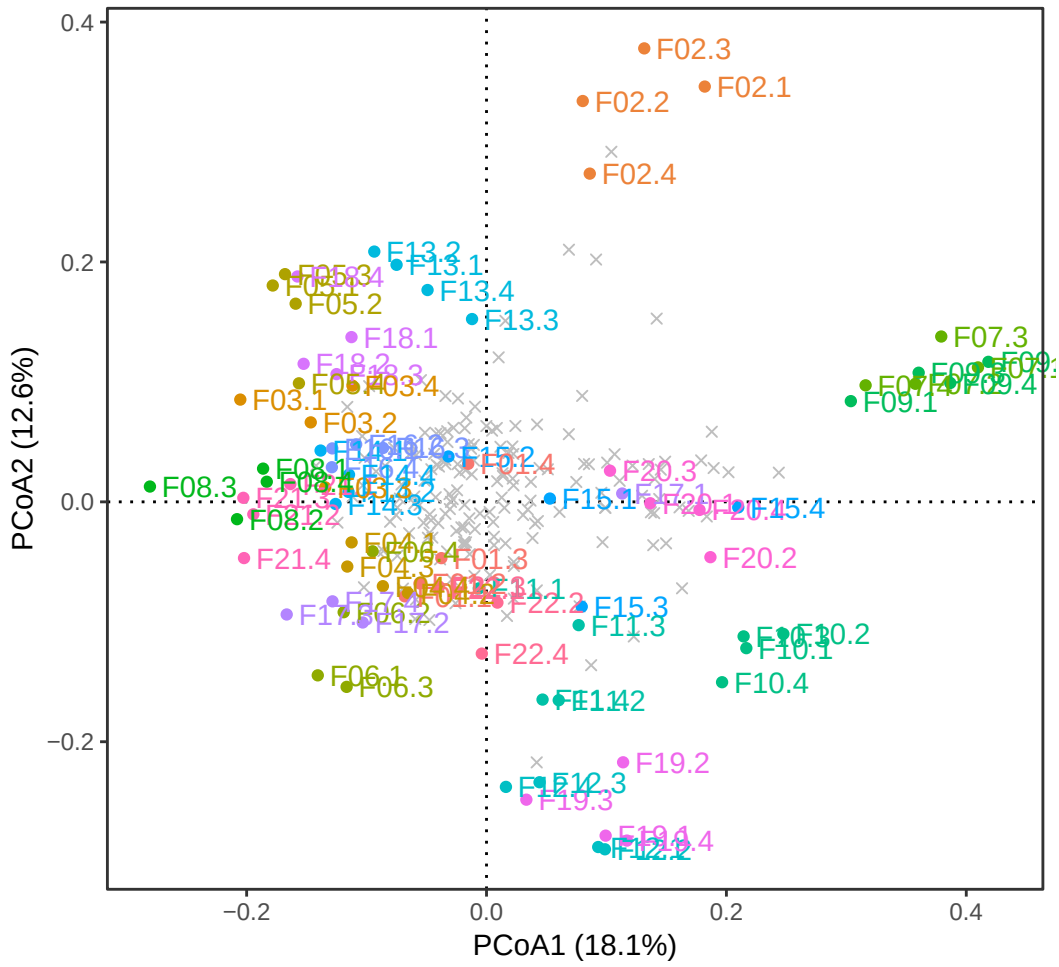

Figure 5.5: Biplot of PCoA of Bray-Curtis dissimilarity matrix computed with filtered and normalized data. The fields are shown as colored dots and the OTUs as grey crosses.

5.7 Summary

In this chapter, we explored different ways to characterize community diversity, both within sample (alpha diversity) and between sample (beta diversity). Three different methods of unconstrained ordination were compared. There is variation across the fields sampled in this thesis. However, the four replicates generally cluster well together.

Chapter 6

Combining soil and fungal community data

In this chapter, we analyze together the two sets of data, the soil environmental data (physical, chemical and biological parameters) and the fungal community data. The fungal community data is the response data ($\boldsymbol{Y}$ matrix) and the soil data contains the explanatory variables ($\boldsymbol{X}$ matrix). Possible methods include indirect comparison and direct comparison. In indirect comparison, correlation of the ordination vectors is performed *a posteriori* with the environmental variables. The matrix $\boldsymbol{X}$ does not participate in the ordination of matrix $\boldsymbol{Y}$. In direct comparison, matrix $\boldsymbol{X}$ participates in the ordination, and forces the ordination vectors to be related to combinations of the variables of $\boldsymbol{X}$ (Borcard *et al.*, 2018).

For the fungal community data, four DNA extractions and PCR were performed for each soil sample but the soil parameters were only measured once. Because the methods from this chapter need to have the same number of samples for both sets of data, the mean of the four technical replicates per field was calculated. The four replicates cluster close together with PCA and PCoA, justifying our decision. Averaging was done with the filtered and normalized data (relative abundance). For redundancy analysis, the data was further transformed with the Hellinger transformation (square root of the relative abundance).

6.1 Indirect comparison

First, we examined whether any of the environmental variables showed a gradient across the ordination with Bray-Curtis dissimilarities. Principal coordinate analysis was performed with `ape::pcoa()` within `phyloseq`. Figure 6.1, made with the mean of 4 replicates, is quite similar to Figure 5.5, where all the replicates are analyzed together, except that they are mirrored. For example, on both plots, Field 2 is at the top, Field 7 and Field 9 are very close to each other, indicating that they share a lot of species, and Field 12 and Field 19 are close to each other as well.

Each of the 9 variables selected in Chapter 3.2 was inspected by coloring the sites from the PCoA with the variable as a gradient. There was a convincing gradient only with `pH` (Figure 6.1): from the bottom-left corner (fields with lower pH) to the top-right corner (fields with highest pH). Field 2 is the field with highest pH and is alone, away from the others on the y axis, suggesting that: 1) this fields does not share a lot of species with the other fields, and 2) there is another factor than pH explaining the changes in community composition for that site.

Second, `vegan::envfit()` was tested. This function fits environmental vectors onto an ordination. Furthermore, the function does a permutation test to assess significance of the fitted vectors (by default, 999 permutations). The function works with any ordination object from `vegan` and others as well. For example, `vegan::envfit()` was used with `cmdscale()` which performed PCoA with the Bray-Curtis dissimilarities (Figure 6.1). On this figure, only vectors with p-values < 0.1 are plotted. The evidence for a correlation of the vector of `pH` with the ordination was moderate (p-value $= 0.002$). The evidence was very weak for `Nmin` (p-value $= 0.081$) and `P` (p-value $= 0.093$). The arrow with pH points to the top-right corner, confirming what was previously observed with the pH gradient on Figure 6.1. The two plots were obtained with different functions, `ape::pcoa()` for top and `cmdscale()` for bottom, but the results of the ordinations look identical.

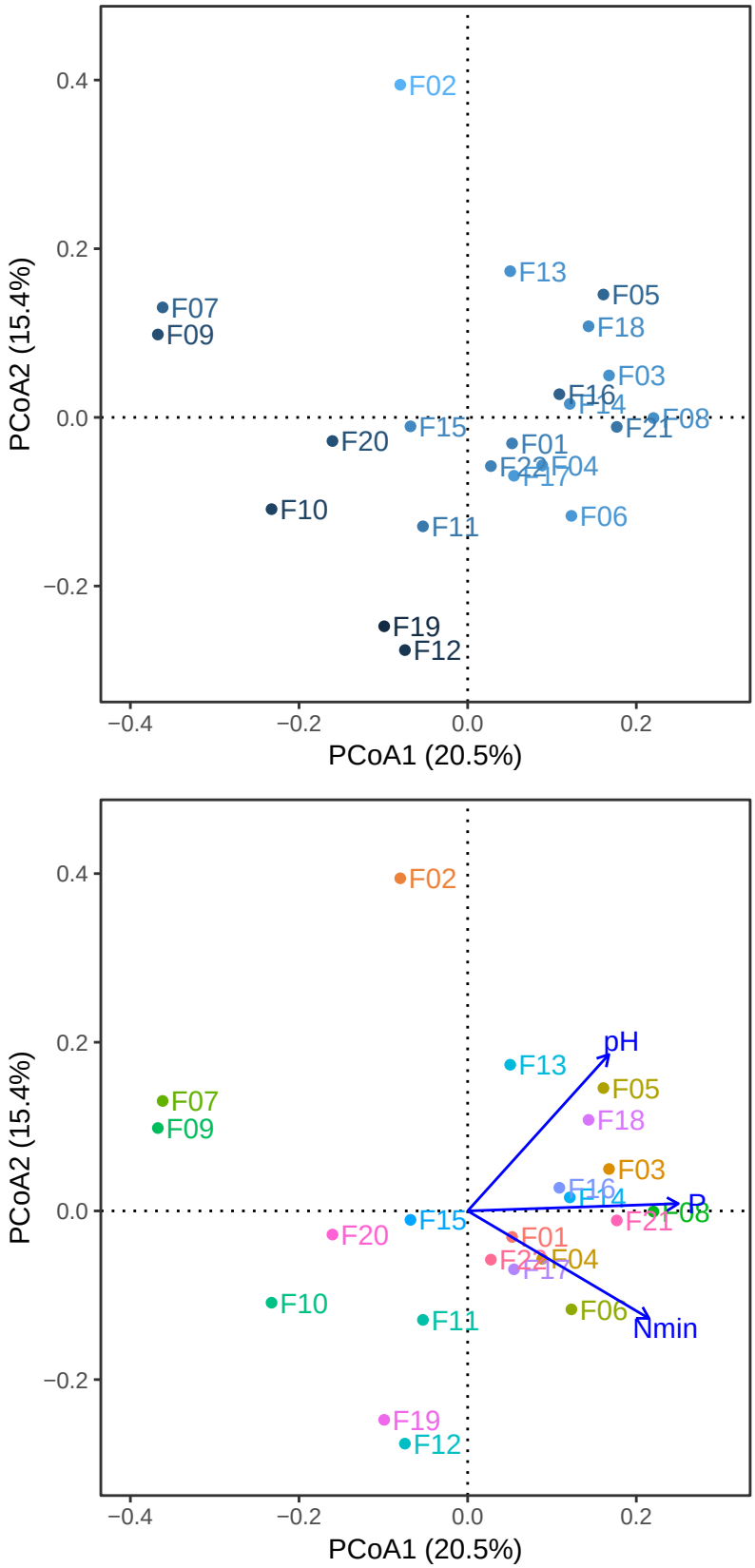

Figure 6.1: PCoA of filtered and normalized data (mean 4 replicates) with Bray-Curtis dissimilarity. Top: Sites are colored by *pH*. Bottom: Environmental variables are fitted to the ordination with *vegan::envfit()*. Arrows show variables with p-values < 0.1.

6.2 Constrained analysis

In constrained analysis, the environmental data participates in the ordination of the community data.

6.2.1 Redundancy analysis

For RDA analysis, the community data needs to be transformed with the Hellinger transformation for reasons described above (Chapter 4.3) and in Legendre and Gallagher (2001). Afterwards, a global test of the RDA result was performed with `vegan::anova.cca()`, which tests the significance of the relationship of the community data ($\boldsymbol{Y}$) with the environmental data $\boldsymbol{X}$. There is moderate evidence for the global test of RDA (p-value = 0.001). The same function can be used to test for the significance of each axis. The evidence for the first axis is moderate (p-value = 0.001) and weak for the second axis (p-value = 0.044).

Next, the variance explained by the model was assessed. In RDA, the variance has two components: constrained and unconstrained. The sum of the eigenvalues for all the constrained axes is the constrained variance. In our case, the constrained variance is 13.285 and the total variance is 24.994, so the proportion of variance explained by the constraints is 0.532, which is the R^2. As noted above (Section 4.4.5), this R^2 is biased (Borcard *et al.*, 2018). The adjusted R^2, computed with `vegan::RsquareAdj`, was lower (R^2_{adj} = 0.18).

Finally, a triplot can be drawn to show the results of RDA (Figure 6.2). By default with no argument, `vegan::plot.cca()` plots the species, `'bp'` (the biplot arrows), and `'wa'` (the weighted sums of site scores). However, plotting the `'lc'`, the linear constraints, is preferred instead of plotting the `'wa'` because the `'lc'` represent the values constrained by the model ($\hat{\boldsymbol{Y}}\boldsymbol{U}$), whereas `'wa'` include the environmental noise ($\boldsymbol{Y}\boldsymbol{U}$) (Borcard *et al.*, 2018).

The interpretation of a triplot is similar to a biplot for species and sites; however, two scalings are available, where either the species or sites scores are scaled by eigenvalues. In this thesis, to simplify, only the scaling of 2 (or `'species'`) was represented. For the interpretation of scaling of 1 (or `'site'`), see (Borcard *et al.*, 2018). The scaling of 2 produces a "correlation" triplot. Here, the angle between the explanatory variables and

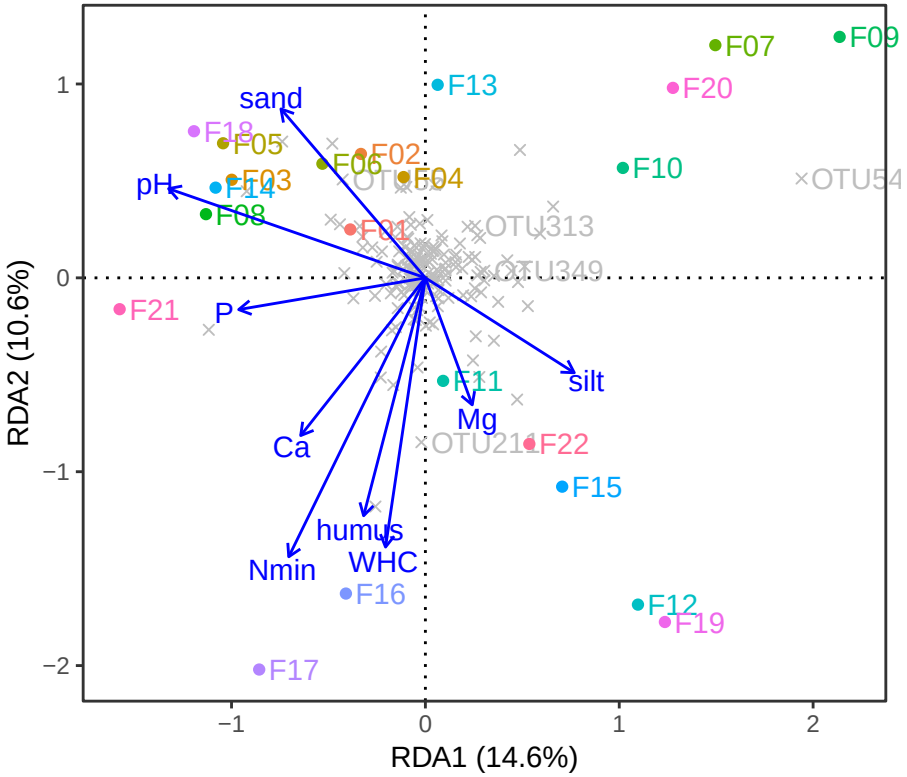

Figure 6.2: Triplot of RDA analysis. Field are presented by colored points and OTUs by grey crosses, environmental variables represented by blue arrows. OTUs with cumulative goodness-of-fit larger than 0.5 (arbitrary) are labeled by their name.

the species or between the explanatory variables indicates their correlation (the smaller the angle, the stronger their correlation). Moreover, the projection of a site at at right angle on a species vector or a quantitative explanatory vector approximates the value of that site on those variables.

The largest arrow on Figure 6.2 is for *Nmin* while the second largest is the arrow for *pH*, which suggest that both variables structures the community most strongly. By contrast, the arrow for *Mg* is the smallest, indicating that this variable is not very important. The *Nmin* arrow points towards the bottom-left corner, in the direction of Field 17, which indicates the the value of *Nmin* is high in that field, as can be seen in Table 3.2. The arrow for *P* points in the direction of Field 21, indicating the relatively high value of *P* in the field. Close to that arrow are the arrows for *humus* and *WHC*, which are pointing in the direction of Field 16. The angle between the arrows for *humus* and *WHC* is small,

indicating their strong correlation with each other, as was also observed in Figure 3.4. This is expected because the more soil organic matter in a soil, the more water it can retain.

All the OTUs of the filtered data are plotted in grey Figure 6.2. They are mostly clustered in the middle of the plot. A few OTUs were selected based on goodness-of-fit with `vegan::goodness.cca()`. This function calculates the cumulative proportion of variance accounted by each species on each axis. According to Borcard *et al.* (2018), "the higher the goodness-of-fit, the better the species is fitted on the corresponding axis". Five OTU with cumulative goodness-of-fit larger than 0.5 (arbitrary) were identified: OTU 349, OTU 313, OTU 211, OTU 55, and OTU 54). OTU 54 is the furthest away from the cluster of OTUs, it is close to the x-axis and close to Field 7, Field 9 and Field 20, which indicates that it is more abundant in those fields.

6.2.2 Canonical correspondence analysis

Canonical correspondence analysis (CCA) was run with `vegan::cca()`. For this analysis, the filtered and normalized (relative abundance) data was used. The evidence for the global test of the constraint was weak (p-value = 0.039). The constrained explained inertia is 0.882 and the total inertia is 1.824, so the adjusted explained inertia is 0.483 (R^2). In this case, Ezekiel's formula cannot be used to compute the adjusted R^2 so a bootstrap procedure, with 1000 permutation by default, is used by `vegan::RsquareAdj`. The adjusted R^2 was much lower ($R^2_{adj} = 0.0957$).

Similarly to RDA, a triplot can also be drawn to show the results of CCA (Figure 6.3). Similar to RDA, the CCA triplot displays how the community is organized in relation to the environmental constraints. An advantage of CCA is that species are "ordered along the canonical axes following their ecological optima". The two scalings are also available with CCA. For interpretation of scaling 1 (by `'sites'`) and further details about CCA triplot, see Borcard *et al.* (2018). With scaling 2 (by `'species'`), the projection of a species at right angle on the environmental variable displays the optimum of a species. The largest arrow on the CCA is pH. The OTUs on the CCA are more spread out than the OTUs on the RDA (Figure 6.3). There were only three OTUs with a goodness-of-fit >0.5: OTU 349, OTU 1 and OTU 54.

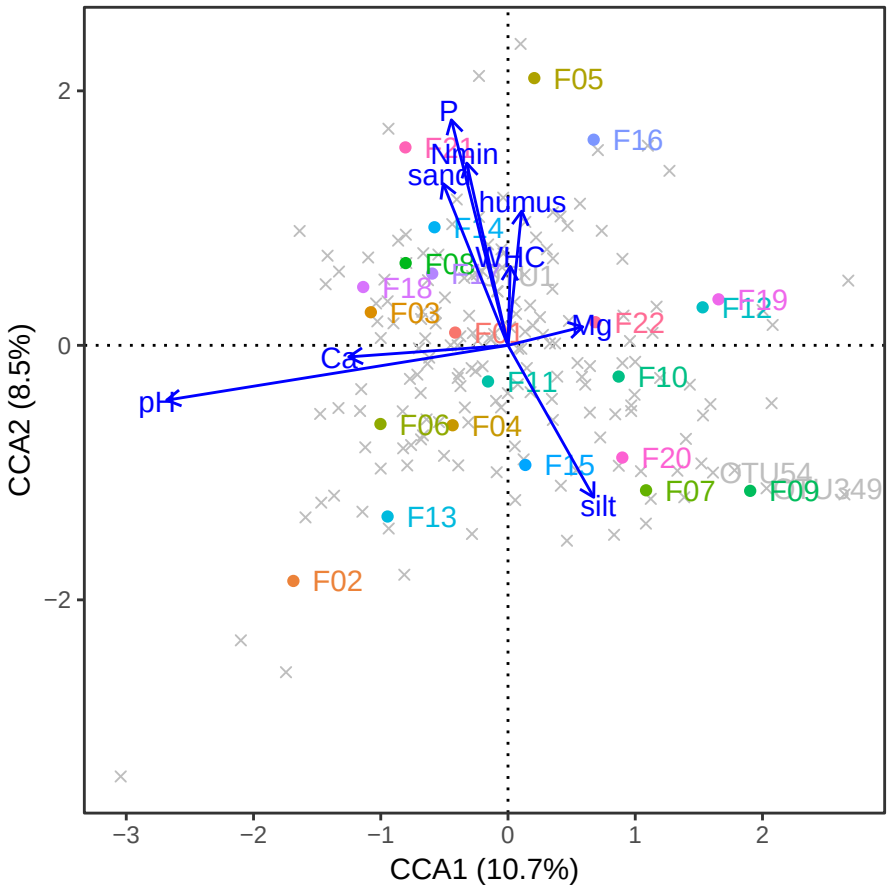

Figure 6.3: Triplot of CCA analysis. Field are presented by colored points and OTUs by grey crosses, environmental variables represented by blue arrows. Only OTUs with cumulative goodness-of-fit larger than 0.5 (arbitrary) are labeled with their name.

6.2.3 Distance-based redundancy analysis

Distance-based redundancy analysis (dbRDA) was run with `vegan::capscale`. Contrary to what the name suggest, this function does not perform canonical analysis of principal coordinates (CAP), initially described by Anderson and Willis (2003); nevertheless, the axes from the output are called CAP1 and CAP2. dbRDA was performed with the filtered and normalized (relative abundance) data. There is only moderate evidence for the global test of RDA (p-value = 0.005), and only the first axis is significant (p-value = 0.004). The R^2 was 0.557, while the adjusted R^2 was 0.225.

Finally, a triplot can be drawn to show the results of the dbRDA (Figure 6.4). The species are plotted as weighted average species score. `vegan::goodness.cca()` was not implemented for `capscale`. We noticed similarities between the RDA triplot and the dbRDA triplot, but the signs of both coordinates needed to be inverted, which is allowed

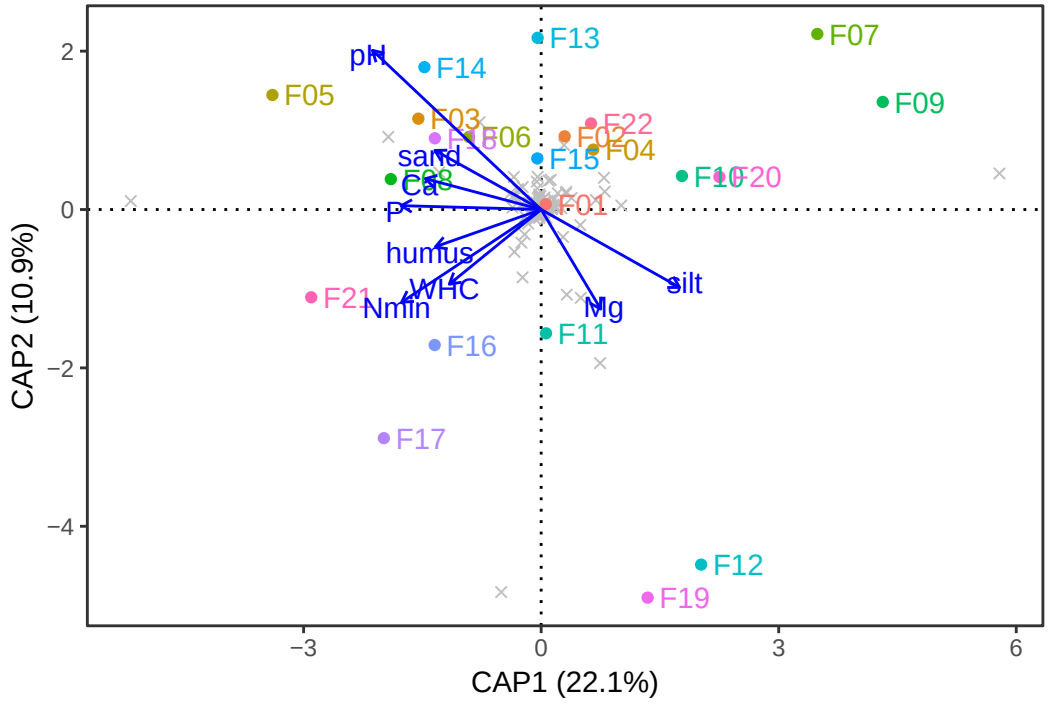

Figure 6.4: Triplot of dbRDA analysis. Field are presented by colored points and OTUs by grey crosses, environmental variables represented by blue arrows.

since RDA which is sign-invariant. After this transformation, Figure 6.4 shares many similarities with Figure 6.2: 1) arrows for pH and sand point to the top-left corner; 3) arrows for humus, WHC, and Nmin point to the bottom-left corner where the same two sites are found (Field 16, Field 17); 2) arrows for Mg and silt point to the bottom-left corner, where the same two sites are found (Field 12, Field 19); 4) there are no arrows pointing to the top-left corner and the same two sites are present there (Field 7, Field 9).

6.3 Summary

In this chapter, we combined the soil and the fungal data. We first explored two methods of indirect comparison. Next, three methods of constrained ordination were compared. We found that RDA and dbRDA produced similar results.

Chapter 7

Predicting community composition

In this chapter, we investigate the prediction tool from `vegan`. Leave-one-out validation is used to compare results from RDA, CCA and dbRDA. Only constrained ordination is explored because the goal of this thesis is to predict community composition from environmental data. Nevertheless, prediction is also possible with unconstrained ordinations.

7.1 Description of `vegan` functions

Recall from Section 4.4.1, the first step of RDA is a regression of each column of $\boldsymbol{Y}$, the matrix of centered response data (community data) on $\boldsymbol{X}$, the matrix of centered explanatory variables (environmental data). The result is $\hat{\boldsymbol{Y}}$, the matrix of fitted values. Next, PCA is applied to $\hat{\boldsymbol{Y}}$, which yield eigenvalues and a matrix $\boldsymbol{U}$ of canonical eigenvectors. This matrix is used to compute $\hat{\boldsymbol{Y}}\boldsymbol{U}$, the linear combinations.

The function `vegan:fitted.cca()` provides an approximation of the original data matrix from the constrained ordination; the fitted results can be either scaled and centered (`type='working'`), which is $\hat{\boldsymbol{Y}}$, or in the original scale of the response (`type='response'`). On the other hand, the function `vegan::residuals.cca()` provides an approximation of the original data from the unconstrained ordination. The function `vegan::scores()` is convenient to extract the scores if the user would like to plot by hand instead of using `vegan::plot.cca()`.

The function `vegan::predict.cca()` acts differently depending on the `type` and `newdata`. First, it can predict sites scores from species (`type = 'wa'`). For this version, `newdata` contains new site(s) but all species from the original data must also be

present in the new site(s). Second, `predict.cca()` can also predict species scores from site constraints (`type='species'`). For example, RDA is run with the subset of most abundant species. Then `predict.cca()` can be used to predict the species scores for the subset of rare species using the RDA object obtained with the abundant species. For that version, `newdata` may contain new species but all sites from the original data must be present. Importantly, if the ordination is computed on transformed species abundance data (for example, in the case of PCA or RDA, Hellinger transformation is typically used), the new species should be transformed using the same transformation (square root of relative abundance). Third, with `type = 'lc'`, the function can predict the linear combination scores ($\hat{\boldsymbol{Y}}\boldsymbol{U}$) from environmental data. In that version, `newdata` must contain all the environmental variables of the original model. Lastly, the function can predict the response data with `type = 'response'` or the scaled and centered community data with `type = 'working'`. For these two versions, `newdata` must either contain environmental variables ($\boldsymbol{X}$) in order to predict the constrained component or community data matrix ($\boldsymbol{Y}$) in order to predict the unconstrained component. According to the help for `vegan::predict.cca()`: with these two versions, the function first uses `newdata` to find new `'lc'` (constrained) or `'wa'` scores (unconstrained), and from these, the function computes the `response` or the `working` data.

In this thesis we will explore two versions of `predict.cca()`, both of them with `newdata` that contains environmental data: 1) with `type = 'lc'`, and 2) with `type = 'response'`.

7.2 Predict linear combinations

Leave-one-out validation is used to compare the results from the prediction of the three constrained ordination methods seen in Section 6.2. For leave-one-out validation, each field is removed from both the community data matrix ($\boldsymbol{Y}$) and the environmental data matrix ($\boldsymbol{X}$). A new constrained ordination is run with the two smaller data sets. First the linear combination scores are predicted from environmental data, so the third method from predict was used with `type = 'lc'` and `newdata` being the environmental data of the removed field. It is important to chose the same `scaling` for both prediction and ordination of the complete data set (either 1 for site or 2 for species). This process

was repeated for each 22 fields in this data set and the result was a matrix with 22 rows (the fields) and 9 columns (the canonical axes). The same process was performed with RDA, CCA and dbRDA. To compare the predicted linear combination with the observed, the site constraints of the ordination with the complete data set were obtained with `vegan::scores()`. As before, Hellinger transformed data is used for RDA while the filtered data is used for CCA and dbRDA.

Figure 7.1 shows the comparison of leave-one-out prediction and observed ordination with RDA and dbRDA methods. `ggplot2` was used to plot the data with `coord_equal( ratio =1)` which fixes the coordinate system to a ratio of 1 in order to achieve the same scale for both axes. The observed data are presented in red while the predicted data is represented in blue. Similar to Figure 6.4, the signs of both coordinates of the dbRDA were inverted in order to make the ordination more similar to RDA. In order to spread out the points and to maximize space use, the two canonical axes were inverted for both ordinations of Figure 7.1: the first canonical axis was plotted on the y-axis while the second canonical axis was plotted on the x-axis. The observed ordination in red is the same as Figure 6.2, respectively Figure 6.4, except for the rotation of the axes.

Figure 7.2 shows the comparison of leave-one-out prediction and observed ordination with CCA, with the customary order of the axes. The observed ordination in red is exactly the same as Figure 6.3.

In some cases, the predicted and observed coordinates are relatively close to each other; for example, Field 9 and Field 21 in all three types of ordination. By contrast, predicted and observed coordinates for Field 5 are far apart from each other for all three ordinations.

7.3 Predict species

Leave-one-out validation was also performed for the last method of predict (`type='response'`). Similarly to above, each field was removed, RDA was performed with the smaller data set and a response vector was predicted from the environmental data of the removed field. The response is in the same scale as the original data; moreover, scaling does not matter with this method. Prediction is repeated for each field, which yields a matrix with 22 rows (the fields) and 171 columns (the OTUs).

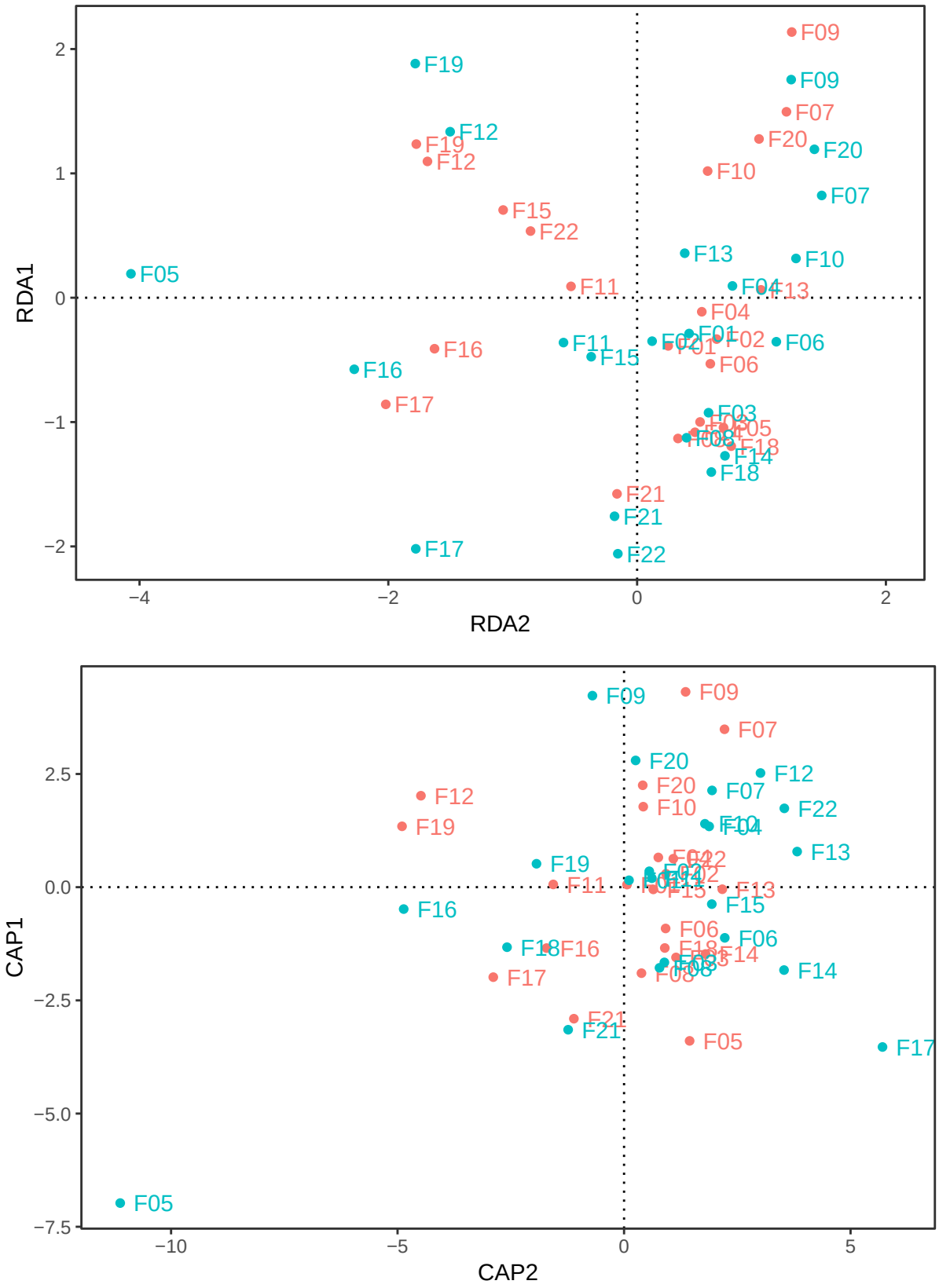

Figure 7.1: Leave-one-out prediction of linear combination scores. Red: observed, blue: predicted. Top: RDA. Bottom: dbRDA.

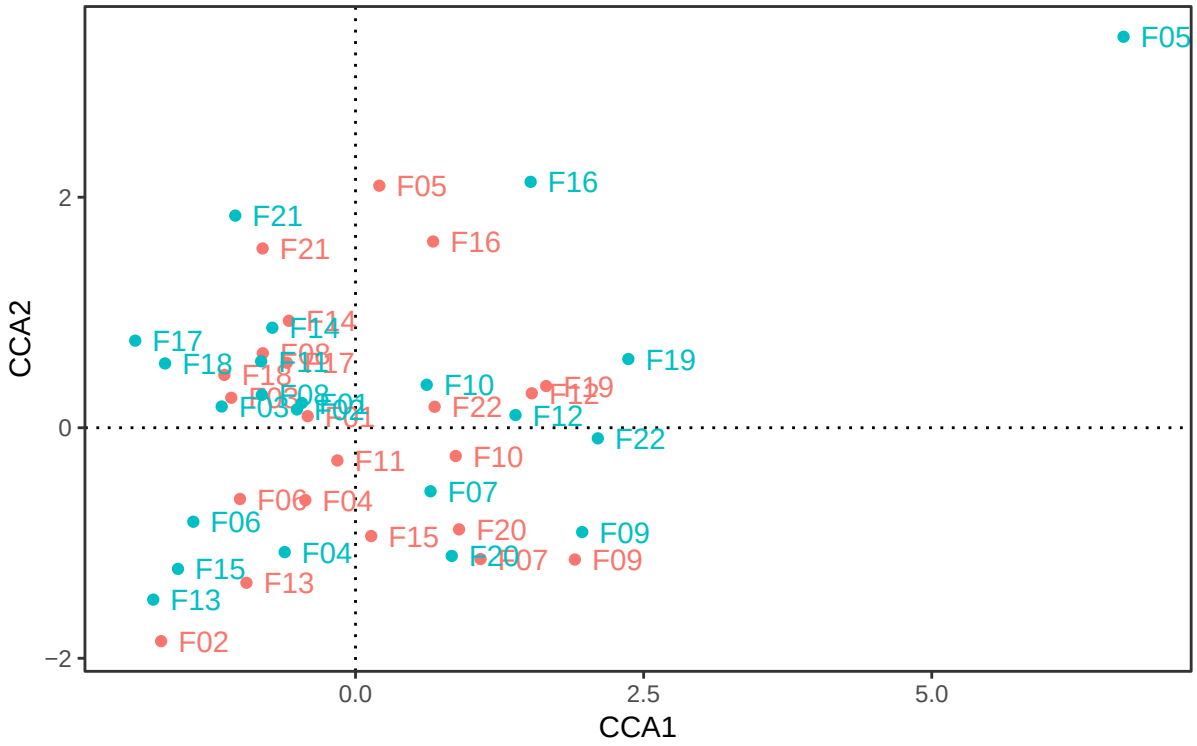

Figure 7.2: Leave-one-out prediction of linear combination scores. Red: observed CCA, blue: predicted CCA.

Figure 7.3 shows the predicted response compared to the $\boldsymbol{Y}$, in this case Hellinger transformed data. Four sites are presented in Figure 7.3: on the top, the two fields with lowest R^2, and on the bottom, the two fields with the highest R^2. Surprisingly, the predicted response may be negative, even though abundance cannot be negative. Field 5 was the site which was most different in the prediction of linear combinations (Figure 7.1), and it is also one of the field with lowest R^2 for the prediction of the response (Figure 7.3). Field 21 was one of the field for which the predicted linear coordinates were more similar to the observed (Figure 7.1) and it is also the field with highest R^2 for prediction of response (Figure 7.3).

7.4 Summary

In this chapter, the predictions of linear combinations for the three methods of constrained ordination were compared. Results were similar, with some fields having predicted coordinates close to the observed and other fields having very different predictions. The

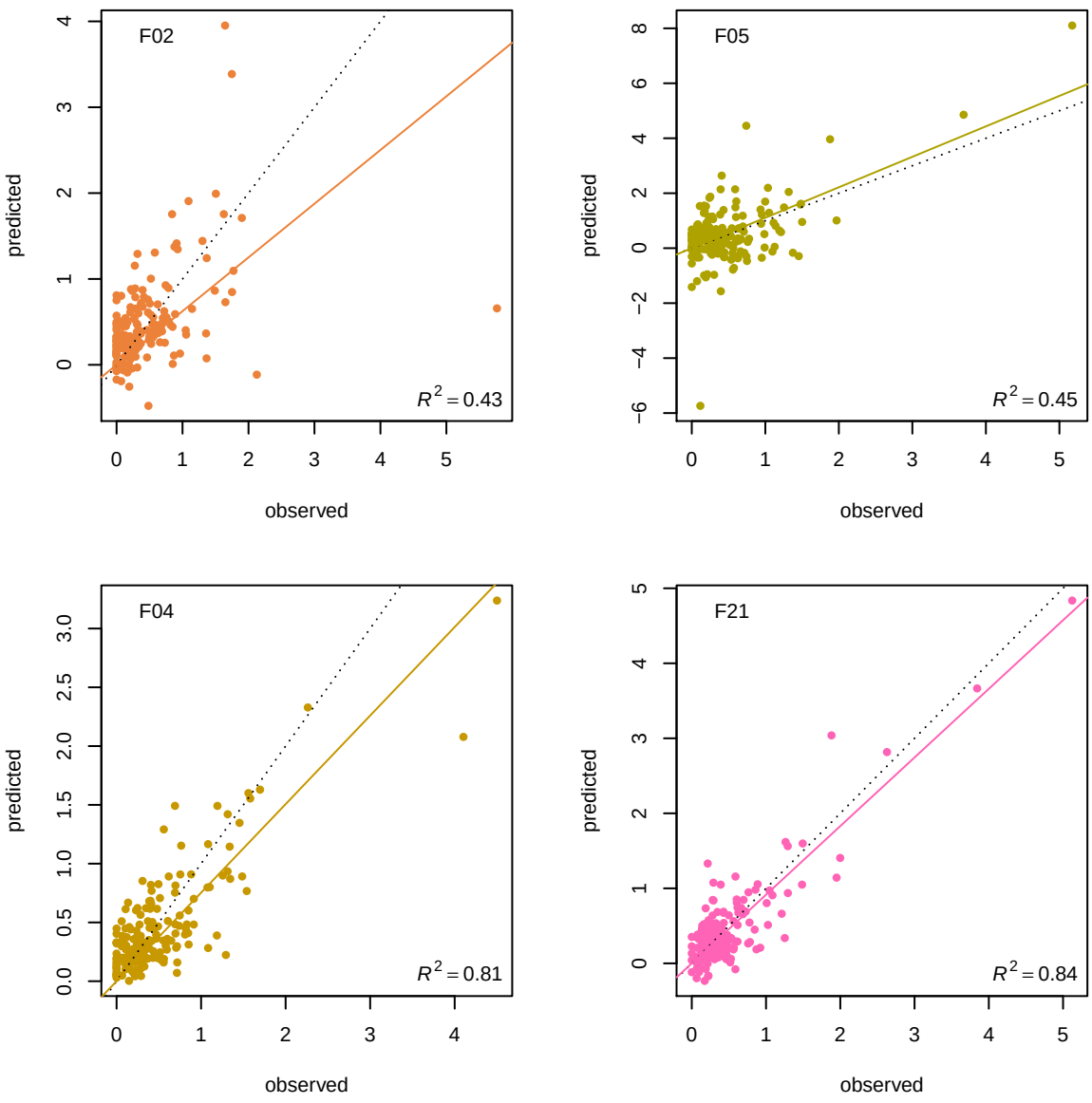

Figure 7.3: Leave-one-out prediction of response. In each panel, the name of the field is indicated in the top-left corner and the R^2 is presented in the lower-right corner. Further, the dotted line shows the identity function and the solid line shows the estimate from the linear model with zero intercept. The color are the same as throughout this thesis.

prediction tool with `type='response'` has only been implemented with RDA. Similarly, the predicted response was similar to the observed data for some fields, but not all.

Chapter 8

Discussion, Outlook, and Conclusion

8.1 Absence of Glomeromycota

The original goal of this thesis was to predict arbuscular mycorrhizal fungi (AMF) community composition; however, we obtained very little sequences for the AMF, which belong to the phylum Glomeromycota. After removing rare OTUs, no Glomeromycota sequences remain (Figure 5.2). Only three OTUs belonging to the phylum Glomeromycota were observed in the unfiltered data (Figure A.2). This suggests that AMF are not very abundant in these soils. Alternatively, their DNA is not well extracted by the DNA extraction method used in this project and/or their DNA is not well amplified with this PCR reaction.

DNA extraction is known to be biased (Martin-Laurent *et al.*, 2001). The DNA of certain species might not be extracted because, for example, some microbes have a very tough cell wall which prevents efficient lysis. Another problem with DNA extraction is that the amount of soil used for DNA extraction is generally quite small. For instance, the kit used in this thesis allows from 250 mg up to 500 mg of soil. By contrast, classical soil analysis are performed with larger amounts (for example 10 g for Nmin, or 20 g for microbial biomass analysis). The small amount of soil used for DNA extraction might not give an accurate representation of the full expanse of the soil microbial community, especially for fungi which are usually larger than bacteria. Other DNA kits allow the extraction of larger amounts of soil; for example, Frey *et al.* (2016) used the "Ultra Clean Soil DNA Mega Prep kit" from MO-BIO, which can extract DNA from up to 10 g of soil. This particular product does not exist anymore since MO-BIO, the company producing

it, was since then bought by QIAGEN. Because some people prefer not to depend on commercial products and because they would like to control the ingredients in the extraction buffers, they extract DNA with homemade buffers, for example with a method called CTAB. The first step of CTAB includes bead beating with a lysis buffer containing 0.2% hexadecyltrimethylammonium bromide (CTAB), followed by phenol-chloroform extraction (Bürgmann *et al.*, 2001). In addition, another advantage of homemade DNA extraction protocol is that the volume of buffers can be adapted to the amount of soil; for example, Brierley *et al.* (2009) extracted DNA from 60 g of soil using the CTAB method. On the other hand, one disadvantage of homemade DNA extraction is that it requires more skill to work with these solutions. In addition, most projects, including this one, include several hundreds of samples, so the current trend is towards high-throughput methods where DNA is extracted in parallel in 96-well plates, which is difficult with the homemade extraction protocol and impossible with the commercial "mega" kit. In spring 2019, our technician compared different homemade extraction protocols but was not able to increase the amount of AMF sequences with the primer pair used in this study; therefore, we decided to try another avenue.

PCR is also biased. Firstly, primers amplify some species preferably. For example, Bellemain *et al.* (2010) have shown with *in silico* analysis that primer ITS1-F, used in this study, amplifies preferably Basidiomycetes whereas primer ITS4, also used in this study, amplifies preferably Ascomycetes. Secondly, polymerase might show a preference for certain secondary structure or guanine cytosine (GC) content (Nichols *et al.*, 2018); GC content is the percentage of guanine cytosine nucleotides in a DNA molecule, this affects the melting temperature. To enrich for Glomeromycota, we could use previously published primers that specifically amplify AMF sequences (Schlaeppi *et al.*, 2016). However, our technician tested these primers with samples from this thesis, but high yield and reproducible amplification was not achieved with the original PCR conditions from the paper. We are currently testing this primer pair with alternative polymerases and different PCR conditions to improve the yield of the PCR.

8.2 Variable selection

We measured soil parameters in two different laboratories and obtained 50 variables for 22 fields, so there are more than twice more variables than samples. The rule of thumb for linear regression, is that there should be 5 times more samples than coefficients in the model. In particular, the goal is to avoid collinearity, caused by correlated explanatory variables, which leads to inaccurate estimation of the coefficients. For constrained ordination, there is no rule of thumb but note that in the classical example of Borcard *et al.* (2018), the data has 29 sites and 10 environmental variables. Therefore, we attempted to reduce the number of variables.

In Chapter 2, we reduced the number of variables to nine using different methods. First, the data was sorted into three categories: chemical, physical, and biological properties. Some variables were redundant because they were measured by the two laboratories (e.g. soil texture, nitrogen, humus, soil respiration); we kept the variables measured by Agroscope because the methods used by this laboratory are more reliable. Most of the macronutrients were extracted with different buffers; after checking for correlation of the different extractions with one another, only one measurement was kept for each macronutrient. Next, the six most relevant variables were chosen based on the literature. Finally, `varrank` was used to select three more variables.

Alternatively, we could have performed variable selection directly with constrained ordination. Several functions in `vegan` and `adespation` are available for this purpose. For example, `vegan::vif.cca()` computes the variance inflation factor (VIF). Variables with a VIF above 10 should be avoided, because they indicate correlation between variables. `vegan::ordistep()` can be used for forward and backward selection and it accepts also factor explanatory variables; however, according to Borcard *et al.* (2018), this function may be too liberal. Therefore, `adespation::forward.sel()` is preferred. On the other hand, this function does not accept factor explanatory variables directly, these need to be re-coded with dummy variables. A third option is `adespation::ordiR2step()`, which does accept factor variable; however, this function can only do forward variable selection and only works with RDA and dbRDA. In the future, we plan to refine our models using variable selection with one of those procedures.

8.3 Unconstrained ordination

In this thesis, three unconstrained ordination methods were compared (Table 8.1). Different transformations are recommended depending on the ordination. For PCA and RDA, appropriate transformations need to be used to avoid the double-zero problem, where two samples sharing no species appear to be similar to each other even though the reasons for not having that species are different. To avoid this problem, we used the Hellinger transformation for PCA and RDA. However, in some cases, a double zero should be taken into account, for example when analyzing data from synthetic community experiments, where the starting community is the same for all the samples. In this case, PCA without transformation would be suitable.

With unconstrained ordination, the community data is analyzed on its own; this category of methods allows to see which sites are similar to each other and which species occurs at which site. The results of PCA and PCoA were comparable once the sign for x-axis and y-axis were inverted (Figure 5.3 and Figure 5.5). Field 2 is isolated from the other fields, suggesting that it does not share many species with those fields. On the other hand, Field 7 and Field 9, as well as Field 12 and Field 19, are close to each other, indicating that these two pairs of fields share a common community composition. Although the two plots were produced with different ordinations, distances and transformations, they are similar. Similarly, Legendre and Gallagher (2001) also found that PCoA of Bray-Curtis distance matrix is more similar to PCA with Hellinger transformation than CA.

By contrast the ordination of the species scores are very different between the ordination methods. With PCA, most of species are clustered around the origin, while with PCoA, they are more spread out. Recall that the species scores for PCoA need to be computed as weighted averages (Legendre and Gallagher, 2001).

The results of CA are very different from PCA and PCoA (Figure 5.4). Most of the

Table 8.1: Ordination methods in this thesis.

unconstrained	constrained	distance	transformation	R function
PCA	RDA	Euclidean	Hellinger	`vegan::rda()`
CA	CCA	χ^2	relative abundance	`vegan::cca()`
PCoA	dbRDA	any	relative abundance	`cmdscale()`, `ape::pcoa()`, `vegan::capscale()`

samples are clustered together around the origin, except for Field 2, which is separated from the others on the first axis, and Field 12, on the second axis. Interestingly, these were also the two fields most different from others on the barplot (Figure 5.2). Two OTUs are driving these differences. Indeed, Field 12 was one of the field with lowest evenness (Figure A.1), which indicates that this field is dominated by one ore more OTUs. CA appears to be sensitive to extreme outliers. Other transformation such as log or square root might help to improve the separation of the samples on the ordination.

Finally, we observed that that the four replicates are close to each other with PCA and PCoA (Figure 5.3 and Figure 5.5). For each soil sample, four DNA extractions were performed and one PCR reaction was prepared with each DNA template to prepare the libary for sequencing, so there are four replicates for the fungal community. On the other hand, the soil properties were only measured once for each field. Because the community composition of the four replicates is similar, we can average the four replicates before doing constrained ordination, in order to analyze the fungal data together with the soil data.

8.4 Constrained ordination

We also compared the results of three constrained ordination methods (Table 8.1). In constrained ordination, the environmental data participates in the ordination of the community data. The fraction of the variance of the community data which is explained by the environmental data is called R^2; however, the adjusted R^2 is preferred because R^2 is biased. The adjusted R^2 was highest for dbRDA, followed by RDA, finally CCA. Ecologists generally prefer to use dbRDA because this method works with any dissimilarity matrix, including beta diversity indices and in particular the popular Bray-Curtis dissimilarity. Triplots of RDA and dbRDA produced similar results for our data set (Figure 6.2 and Figure 6.4), after inverting the signs of the coordinates of dbRDA. On both plots, the arrows for `Nmin`, `WHC` and `humus` are pointing towards the same direction, indicating the strong correlation of those variables with each other, and these variables are close to Field 16 and Field 17, suggesting they have high values in these fields. By contrast, the results of RDA and dbRDA were different with the data set used in Borcard *et al.* (2018).

pH was identified as one of the most important factor shaping the fungal community of the fields sampled in this thesis. *pH* was the only variable which showed a visible gradient on the PCoA (Figure 6.1, top), as well as one of the variables selected by `vegan::envfit()` to fit the PCoA ordination (Figure 6.1, bottom). Moreover, *pH* was one of the largest arrow on the RDA (Figure 6.2), indicating that this variable has a large variance. In previous studies, pH has been shown to be a major predictor of microbial communities. For example, Lauber *et al.* (2009) have shown that pH shapes bacterial communities at the continental scale. Fungal communities appear to be less affected by pH compared to bacterial communities (Rousk *et al.*, 2010). However, we did not assess bacterial community composition in this project because the larger goal is to predict AMF community composition in order to improve inoculation success. AMF communities are also structured by pH (Van Geel *et al.*, 2018), but we were not able to obtain enough AMF sequences to confirm this result.

In the fields analyzed in this thesis, the two macronutrients N and P were identified to be important factors shaping the fungal community. *Nmin* and *P* were both selected by `vegan::envfit()` (Figure 6.1) and *Nmin* was the largest arrow on the RDA (Figure 6.2). Nitrogen is a well-known factor shaping soil bacterial communities, for example, Fierer *et al.* (2012). Moreover, application of phosphorus and nitrogen induced community shifts in soil bacterial and fungal communities (Leff *et al.*, 2015). Macronutrient levels are expected to affect community composition because high macronutrient levels favor copiotrophic microbes, which are organisms that prefer nutrient-rich environment, whereas low macronutrient levels favor oligotrophic microbes, which are organisms that live in environments with low nutrients.

8.5 Missing variables

With all three constrained methods shown in this thesis, a relatively small proportion of the variance was explained by the environmental constraints (for example, for RDA, $R^2_{adj} = 0.18$). By comparison, the proportion of variance explained by the environmental variables was much higher for the data set used in Borcard *et al.* (2018), which are fish communities along the Doubs river in France (for RDA, $R^2_{adj} = 0.522$). In this thesis, we considered fungi; probably less is known about the factors that affect microbial commu-

nities. On the other hand, we might have failed to measure important soil properties.

In fact, not all the classical soil properties were measured in this thesis. In particular, we did not measure bulk density which is known to be important for soil microbes. For example, Hartmann *et al.* (2014) showed that forest microbes, and in particular fungi, were affected by soil compaction. Similarly, Gattinger *et al.* (2002) showed that microbes in agricultural fields are impacted by tractor driving. Our reasoning for not including bulk density in the list of properties measured was that we expected soil compaction not to vary much across the different fields because all fields in this study were tilled and tractor driving during tillage is known to affect compaction. Moreover, we did not measure bulk density because it is more difficult to measure and it is not part of the standard soil analysis done by farmers. Nevertheless, bulk density might have been important in structuring the fungal communities.

In addition, we did not take into account the environment (altitude, temperature, micro-climate). In particular the amount of precipitation is known to affect microbial community composition, for example, Naylor *et al.* (2017). Precipitation data is available for over 260 automatic weather stations all over Switzerland from the Federal Office of Meteorology and Climatology MeteoSwiss (2019). We could obtain precipitation data from the closest weather-station for each field, but this might not be accurate because the network of weather stations is not extensive

Finally, we did not consider farmer management of the fields. Management factors expected to influence microbial communities include: pre-crop, cover crop, fertilizer type and application, and compost application (Bender *et al.*, 2016). Questionnaires were sent to farmers to ask them how they had managed the fields sampled in this study in previous years; unfortunately, only 90% of the farmers returned the completed form. In the future, we will include information from the farmer questionnaires to further refine our model.

8.6 Prediction of communities

The main goal of this thesis was to predict community composition. Prediction with the three constrained ordination methods was compared to observed data using leave-one-out validation. All three constrained ordination methods presented in this thesis allow the prediction of linear combinations, but only RDA can predict species abundance.

Predicted and observed linear combinations were farthest apart for Field 5 for all the ordination methods presented in this thesis (Figure 7.1 and Figure 7.2). Field 5 has a soil texture classified as sandy clay loam (Figure 3.3). On the other hand, Field 18 has the same soil texture but predicted and observed coordinates were much closer to each other. Field 5 is not an outlier in any other parameters measured, either soil parameters (Figure 3.4) or alpha diversity (Figure A.1). This indicates that other factor(s), not measured here, make prediction difficult for that field.

Predicted species abundances were relatively close to the observed values for most of the fields, with the notable exception of Field 2 and Field 5. Field 2 was very far from others on both PCA (Figure 5.3) and PCoA (Figure 5.5), indicating that it does not share a lot of species with the other sites. Field 2 was also very different in composition at the phylum level, with more abundant Basidiomycota compared to all the other sites (Figure 5.2). Field 2 was the field with highest pH (Table 3.2), but it was not very different from the other fields for the other environmental variables (Figure 3.4). Interestingly, the prediction of species abundance for Field 2 was different from observed even though prediction of linear combinations was similar to observed.

Prediction of both linear combinations and species abundance were surprisingly successful for the other fields. One explanation is that fields sampled in this thesis are relatively similar to each other: they are suitable for maize crop, they have mostly similar texture (Figure 3.3), they were all regularly tilled. We excluded fields which were not tilled or from organic farmers because we expected these fields to have different microbial community composition (Banerjee *et al.*, 2019). All in all, selecting only fields run under conventional management (as opposed to organic or no-till) led to good prediction of fungal communities based on soil properties.

8.7 Outlook

8.7.1 Improving the ordination

RDA was limited to the environmental variables measured in this thesis. Results could be improved by including more variables (Section 8.5); in particular, answers of questionnaires about farm management will be taken into account. As noted above, the number of fields was smaller than the number of soil variables measured. However, the number of

sites will increase in the next years. This thesis is embedded in a three year project. In the first year, 2018, 22 fields were sampled. In the spring of 2019, 25 fields were sampled; data for these fields will become available in the fall. In 2020, 10 more fields will be sampled. Altogether, there will be more than 60 sites. The new sites will be included in the analysis presented in this thesis. Increasing the number of fields will improve the prediction, although we will need to take into account `year` in the model as a random factor. This is possible with `vegan::rda()` and is called "partial" RDA. The information about the random factors is supplied as conditioning matrix `Z`. Partial RDA removes the effect of the random variable(s).

8.7.2 Other transformations and ordinations

In this thesis, we used Hellinger transformation for PCA and RDA and relative abundance for the other ordination methods. We did not try log-transformation of the data, which was recommended for example by Callahan *et al.* (2016). Microbiome data typically contains a lot of zero so a pseudo-count need to be added before log-transformation; because the choice of pseudo-count is arbitrary, it can affect the ordination (Hawinkel *et al.*, 2019a).

Gloor and Reid (2016) have argued that microbiome data is compositional data, which they defined as "a data set in which the parts in each sample have an arbitrary or non-informative sum". In the case of microbiome data, the sum is the number of sequences. The problem with analyzing compositional data with relative abundance, is that if the true abundance of one OTU in one sample increase, the relative abundance of that OTU will increase as a result but the relative abundances of the other OTUs will appear to decrease, because the number of sequences provided by the sequencing instrument is fixed. This phenomenon is called "negative correlation bias" (Gloor and Reid, 2016). Aitchison (1982) and others have developed different tools to deal with compositional data, this has been called compositional data analysis (CoDA). The first step in CoDA is to convert the relative abundance to ratios between all parts, for example using the centered log-ratio transformation. Standard multivariate analysis techniques can then be applied to the transformed data set.

We restricted the choice of methods to three unconstrained and three constrained ordinations (Table 8.1). Canonical analysis of principal coordinates (CAP), initially described

by Anderson and Willis (2003), is similar to dbRDA because both methods first compute a distance matrix. CAP analysis is available with the popular program *PRIMER-e*, which was developed for multivariate analysis for ecology (Clarke and Gorley, 2015). However, this program is not command-line and thus does not allow reproducible research. According to its authors, *BiodiversityR::CAPdiscrim()*'s implementation is closer to the method described by Anderson and Willis (2003); while the method implemented by *vegan::capscale()* is closer to distance-based Redundancy Analysis (Legendre and Anderson, 1999). However, *BiodiversityR* provides a Graphical User Interface (GUI) for *vegan* and is thus also not command-line. There is no corresponding *vegan* function for CAP.

Recently, Hawinkel *et al.* (2019a) introduced a new R package *RCM* (Hawinkel *et al.*, 2019b), which stands for Row-Column interaction model of dimension M; M is the number of dimensions of the ordination, typically 2 or 3. The model RC(M) combines ideas of dispersion estimation from sequencing data with flexible response function estimation. Further, similar to RDA, it allows for conditioning of confounding variables as well as constrained ordination. Finally, it provides diagnostic tools to check assumptions of the model. We chose to only use methods available in the *vegan* package for this thesis, but alternative transformations and ordinations should be tested in the future.

8.8 Conclusion

Unfortunately, we did not obtain enough Glomeromycota sequences to predict AMF community composition from soil data, the original goal of this thesis. We used the available data about soil fungi to explore different methods of studying microbial communities. We compared three unconstrained and three constrained ordination methods and we predicted species abundance from soil properties. In the future, we hope to obtain a better data set with improved molecular methods and eventually to be able to use the methods evaluated in this thesis to reach our initial goal.

Appendix A

Appendix

This Appendix has two sections: the first section collects information about R version and packages loaded in this thesis, the second section presents supplementary table and figures.

A.1 sessionInfo()

```
## R version 3.6.1 (2019-07-05)
## Platform: x86_64-apple-darwin15.6.0 (64-bit)
## Running under: macOS Mojave 10.14.6
##
## Matrix products: default
## BLAS:   /Library/Frameworks/R.framework/Versions/3.6/Resources/lib/libRblas.0.dylib
## LAPACK: /Library/Frameworks/R.framework/Versions/3.6/Resources/lib/libRlapack.dylib
##
## Random number generation:
##  RNG:     Mersenne-Twister
##  Normal:  Inversion
##  Sample:  Rounding
##
## locale:
## [1] en_US.UTF-8/en_US.UTF-8/en_US.UTF-8/C/en_US.UTF-8/en_US.UTF-8
##
```

```
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## other attached packages:
##  [1] kableExtra_1.1.0  xtable_1.8-4      vegan_2.5-5
##  [4] lattice_0.20-38   permute_0.9-5     varrank_0.2
##  [7] plyr_1.8.4        phyloseq_1.28.0   forcats_0.4.0
## [10] stringr_1.4.0     dplyr_0.8.3       purrr_0.3.2
## [13] readr_1.3.1       tidyr_0.8.3       tibble_2.1.3
## [16] ggplot2_3.2.1     tidyverse_1.2.1   soiltexture_1.5.1
## [19] knitr_1.24
##
## loaded via a namespace (and not attached):
##  [1] nlme_3.1-141        lubridate_1.7.4     webshot_0.5.1
##  [4] httr_1.4.1          tools_3.6.1         backports_1.1.4
##  [7] R6_2.4.0            lazyeval_0.2.2      BiocGenerics_0.30.0
## [10] mgcv_1.8-28         colorspace_1.4-1    ade4_1.7-13
## [13] withr_2.1.2         sp_1.3-1            tidyselect_0.2.5
## [16] compiler_3.6.1      cli_1.1.0           rvest_0.3.4
## [19] Biobase_2.44.0      xml2_1.2.2          scales_1.0.0
## [22] digest_0.6.20       rmarkdown_1.15      XVector_0.24.0
## [25] pkgconfig_2.0.2     htmltools_0.3.6     highr_0.8
## [28] rlang_0.4.0         readxl_1.3.1        rstudioapi_0.10
## [31] FNN_1.1.3           generics_0.0.2      jsonlite_1.6
## [34] magrittr_1.5        biomformat_1.12.0   Matrix_1.2-17
## [37] Rcpp_1.0.2          munsell_0.5.0       S4Vectors_0.22.0
## [40] Rhdf5lib_1.6.0      ape_5.3             stringi_1.4.3
## [43] MASS_7.3-51.4       zlibbioc_1.30.0     rhdf5_2.28.0
## [46] grid_3.6.1          parallel_3.6.1      crayon_1.3.4
## [49] Biostrings_2.52.0   haven_2.1.1         splines_3.6.1
## [52] multtest_2.40.0     hms_0.5.0           zeallot_0.1.0
## [55] pillar_1.4.2        tcltk_3.6.1         igraph_1.2.4.1
```

```
## [58] reshape2_1.4.3      codetools_0.2-16    stats4_3.6.1
## [61] glue_1.3.1          evaluate_0.14       data.table_1.12.2
## [64] modelr_0.1.5        vctrs_0.2.0         foreach_1.4.7
## [67] cellranger_1.1.0    gtable_0.3.0        assertthat_0.2.1
## [70] xfun_0.9            broom_0.5.2         survival_2.44-1.1
## [73] viridisLite_0.3.0   iterators_1.0.12    IRanges_2.18.1
## [76] cluster_2.1.0
```

A.2 Supplementary table and figures

Table A.1: 50 variables measured but only 17 retained at the end of the 1st step of variable reduction. The dropped variables are represented in grey. The retained variables are represented in black.

variable	type	variable name
clay (%)	physical	*clay_agro*, *clay_lbu*
sand (%)	physical	*sand_agro*
silt (%)	physical	*silt_agro*, *silt_lbu*
aggregation	physical	*vast_lbu*
water holding capacity	physical	*WHC_agro*
organic matter	chemical	*humus_agro*, *Corg_agro*, *humus_lbu*
pH	chemical	*ph_lbu*, *hydrogen_agro*
nitrogen	chemical	*Nmin_agro*, *ammonium_agro*, *ammonium_H2O_lbu*, *nitrate_agro*, *nitrate_H2O_lbu*, *slan_lbu*
phosphorus	chemical	*phosphorus_CO2_lbu*, *P_tot_agro*, *P_olsen_lbu*, *phosphorus_EDTA_lbu*, *phosphorus_H2O_lbu*
potassium	chemical	*potassium_CO2_lbu*, *potassium_agro*, *potassium_EDTA_lbu*, *potassium_H2O_lbu*
magnesium	chemical	*magnesium_CC_lbu*, *magnesium_agro*, *magnesium_EDTA_lbu*, *magnesium_H2O_lbu*
calcium	chemical	*calcium_H2O_lbu*, *calcium_EDTA_lbu*, *calcium_agro*
sodium	chemical	*sodium_H2O_lbu*, *sodium_agro*
micronutrients	chemical	*manganese_EDTA_lbu*, *iron_EDTA_lbu*, *iron_H2O_lbu*, *copper_EDTA_lbu*, *zinc_EDTA_lbu*, *boron_EDTA_lbu*, *boron_H2O_lbu*
cation exchange capacity	chemical	*CEC_agro, BS_agro*
microbial biomass	biological	*cMIC_agro*, *nMIC_agro*
substrate induced respiration	biological	*respiration_agro*, *respiration_lbu*

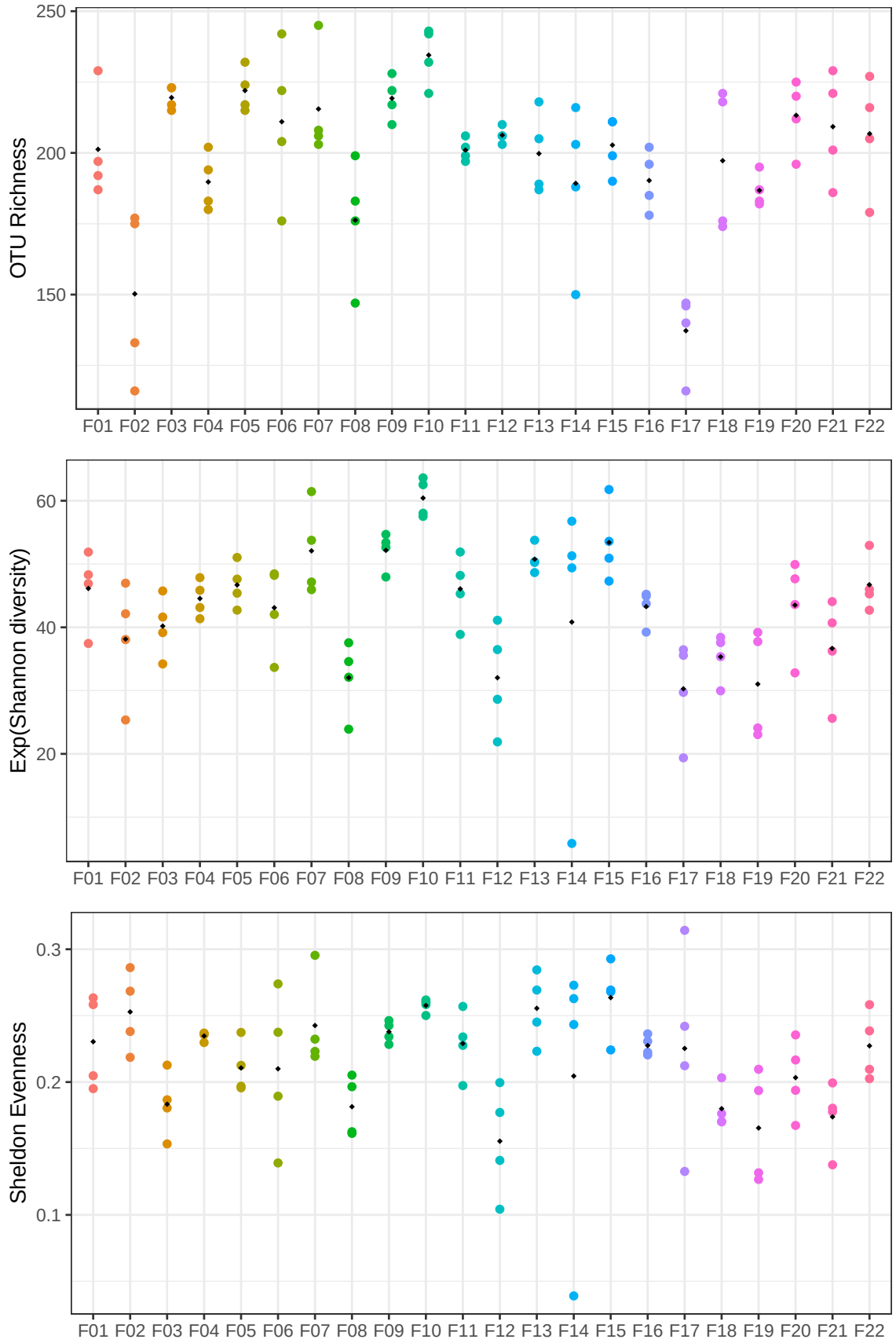

Figure A.1: Three indices of alpha diversity. Top: OTU Richness (N_0). Middle: Exponential of Shannon diversity ($N_1 = \exp(H)$). Bottom: Sheldon evenness (N_1/N_0). Black points represent the mean for each field.

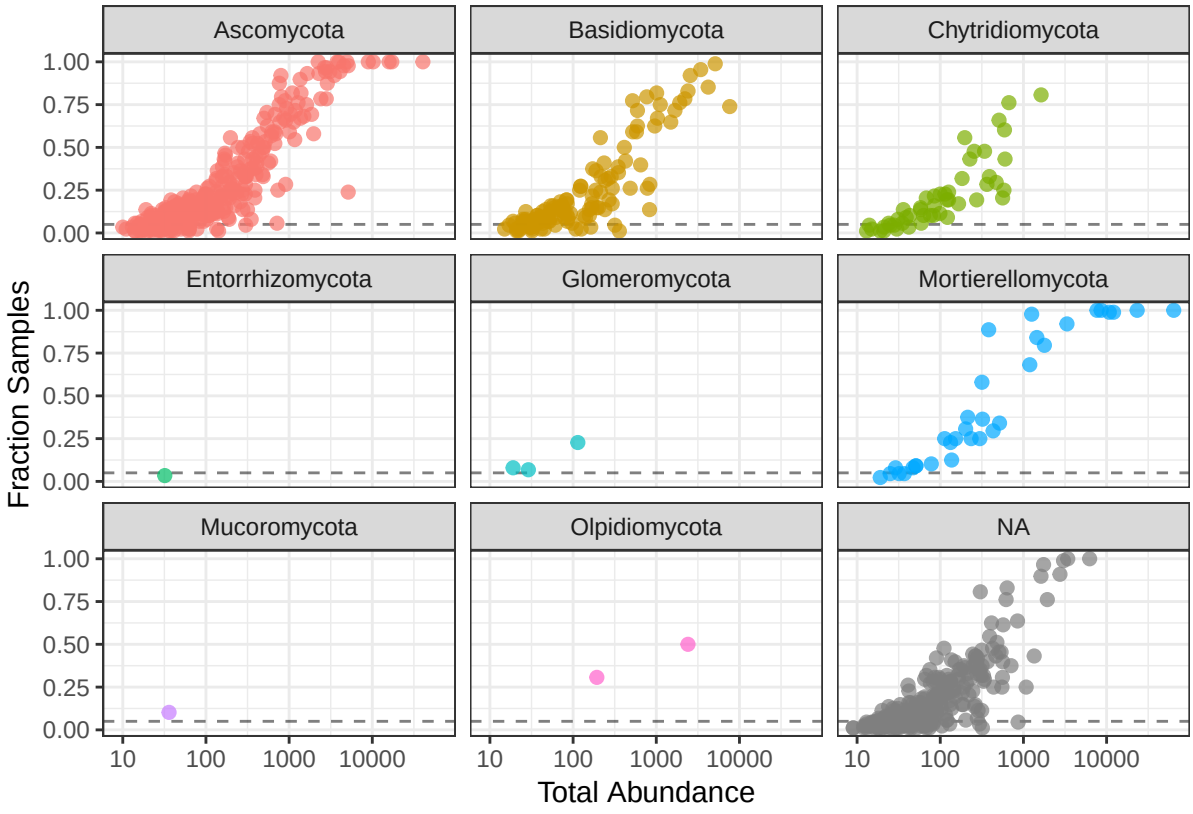

Figure A.2: Prevalence plot (made with R code from Callahan *et al.*, 2016). On the x-axis is the sum of the counts for each OTU and on the y-axis is the proportion of samples where the OTU was found.

Bibliography

Abarenkov, K., Nilsson, R. H., Larsson, K. H., Alexander, I. J., Eberhardt, U., Erland, S., Høiland, K., Kjøller, R., Larsson, E., Pennanen, T., Sen, R., Taylor, A. F. S., Tedersoo, L., Ursing, B. M., Vr*r*alstad, T., Liimatainen, K., Peintner, U., and Kõljalg, U. (2010). The UNITE database for molecular identification of fungi – recent updates and future perspectives. *New Phytologist*, **186**, 281–285. 9

Aitchison, J. (1982). The statistical analysis of compositional data. *Journal of the Royal Statistical Society: Series B (Methodological)*, **44**, 139–160. 61

Anderson, M. J. and Willis, T. J. (2003). Canonical analysis of principal coordinates: a useful method of constrained ordination for ecology. *Ecology*, **84**, 511–525. 28, 45, 62

Banerjee, S., Walder, F., Büchi, L., Meyer, M., Held, A. Y., Gattinger, A., Keller, T., Charles, R., and Van Der Heijden, M. G. (2019). Agricultural intensification reduces microbial network complexity and the abundance of keystone taxa in roots. *The ISME Journal*, **13**, 1722. 60

Bellemain, E., Carlsen, T., Brochmann, C., Coissac, E., Taberlet, P., and Kauserud, H. (2010). ITS as an environmental DNA barcode for fungi: an *in silico* approach reveals potential PCR biases. *BMC Microbiology*, **10**, 189. 54

Bender, S. F., Schlaeppi, K., Held, A., and van der Heijden, M. G. (2019). Establishment success and crop growth effects of an arbuscular mycorrhizal fungus inoculated into Swiss corn fields. *Agriculture, Ecosystems & Environment*, **273**, 13–24. 2, 17

Bender, S. F., Wagg, C., and van der Heijden, M. G. (2016). An underground revolution: biodiversity and soil ecological engineering for agricultural sustainability. *Trends in Ecology & Evolution*, **31**, 440–452. 59

Blume, H.-P., Brümmer, G. W., Fleige, H., Horn, R., Kandeler, E., Kögel-Knabner, I., Kretzschmar, R., Stahr, K., and Wilke, B.-M. (2015). *Scheffer/Schachtschabel Soil Science*. Springer, Berlin, Heidelberg. 6, 7, 12, 13, 14, 15, 16

Borcard, D., Gillet, F., and Legendre, P. (2018). *Numerical Ecology with R*. Springer. 22, 23, 24, 26, 27, 28, 29, 32, 39, 42, 44, 55, 57, 58

Breuillin, F., Schramm, J., Hajirezaei, M., Ahkami, A., Favre, P., Druege, U., Hause, B., Bucher, M., Kretzschmar, T., Bossolini, E., Kuhlemeier, C., Martinoia, E., Franken, P., Scholz, U., and Reinhardt, D. (2010). Phosphate systemically inhibits development of arbuscular mycorrhiza in *Petunia hybrida* and represses genes involved in mycorrhizal functioning. *The Plant Journal*, **64**, 1002–1017. 17

Brierley, J. L., Stewart, J. A., and Lees, A. K. (2009). Quantifying potato pathogen DNA in soil. *Applied Soil Ecology*, **41**, 234–238. 54

Bürgmann, H., Pesaro, M., Widmer, F., and Zeyer, J. (2001). A strategy for optimizing quality and quantity of DNA extracted from soil. *Journal of Microbiological Methods*, **45**, 7–20. 54

Callahan, B. J., Sankaran, K., Fukuyama, J. A., McMurdie, P. J., and Holmes, S. P. (2016). Bioconductor workflow for microbiome data analysis: from raw reads to community analyses. *F1000Research*, **5**, 1492. 34, 61, 68

Clarke, K. and Gorley, R. (2015). Getting started with PRIMER v7. 62

Cordell, D., Drangert, J.-O., and White, S. (2009). The story of phosphorus: global food security and food for thought. *Global Environmental Change*, **19**, 292–305. 1

Edgar, R. C. (2013). UPARSE: highly accurate OTU sequences from microbial amplicon reads. *Nature Methods*, **10**, 996–998. 9

Edgar, R. C. (2016). *SINTAX: a simple non-Bayesian taxonomy classifier for 16S and ITS sequences*. https://www.biorxiv.org/content/10.1101/074161v1. 9

Everitt, B. and Hothorn, T. (2011). *An Introduction to Applied Multivariate Analysis with R*. Springer, New York, NY. 18, 24

Federal Office of Meteorology and Climatology MeteoSwiss (2019). Measurement values. https://www.meteoswiss.admin.ch/home/measurement-values.html?param=messnetz-automatisch, accessed: 23.07.2019. 59

Fierer, N., Lauber, C. L., Ramirez, K. S., Zaneveld, J., Bradford, M. A., and Knight, R. (2012). Comparative metagenomic, phylogenetic and physiological analyses of soil microbial communities across nitrogen gradients. *The ISME Journal*, **6**, 1007. 58

Food and Agriculture Organization of the United Nations (2019). Conservation agriculture. http://www.fao.org/conservation-agriculture/en/, accessed: 31.07.2019. 1

Frey, B., Rime, T., Phillips, M., Stierli, B., Hajdas, I., Widmer, F., and Hartmann, M. (2016). Microbial diversity in European alpine permafrost and active layers. *FEMS Microbiology Ecology*, **92**, fiw018. 53

Gardes, M. and Bruns, T. D. (1993). ITS primers with enhanced specificity for basidiomycetes-application to the identification of mycorrhizae and rusts. *Molecular ecology*, **2**, 113–118. 8

Gattinger, A., Ruser, R., Schloter, M., and Munch, J. C. (2002). Microbial community structure varies in different soil zones of a potato field. *Journal of Plant Nutrition and Soil Science*, **165**, 421–428. 59

Gloor, G. B. and Reid, G. (2016). Compositional analysis: a valid approach to analyze microbiome high-throughput sequencing data. *Canadian Journal of Microbiology*, **62**, 692–703. 61

Gohl, D. M., Vangay, P., Garbe, J., MacLean, A., Hauge, A., Becker, A., Gould, T. J., Clayton, J. B., Johnson, T. J., Hunter, R.,*et al.* (2016). Systematic improvement of amplicon marker gene methods for increased accuracy in microbiome studies. *Nature Biotechnology*, **34**, 942. 34

Hartmann, M., Niklaus, P. A., Zimmermann, S., Schmutz, S., Kremer, J., Abarenkov, K., Lüscher, P., Widmer, F., and Frey, B. (2014). Resistance and resilience of the forest soil microbiome to logging-associated compaction. *The ISME Journal*, **8**, 226–244. 59

Hawinkel, S., Kerckhof, F.-M., Bijnens, L., and Thas, O. (2019a). A unified framework for unconstrained and constrained ordination of microbiome read count data. *PLOS ONE*, **14**, e0205474. 61, 62

Hawinkel, S., Kerckhof, F.-M., Bijnens, L., Thas, O., and R Core Team (2019b). *RCM: a unified approach to unconstrained and constrained visualization of microbiome read count data*. R package version 1.0.0. 62

Herbold, C. W., Pelikan, C., Kuzyk, O., Hausmann, B., Angel, R., Berry, D., and Loy, A. (2015). A flexible and economical barcoding approach for highly multiplexed amplicon sequencing of diverse target genes. *Frontiers in Microbiology*, **6**, 8966. 8

Hoeksema, J. D., Chaudhary, V. B., Gehring, C. A., Johnson, N. C., Karst, J., Koide, R. T., Pringle, A., Zabinski, C., Bever, J. D., Moore, J. C., Wilson, G. W. T., Klironomos, J. N., and Umbanhowar, J. (2010). A meta-analysis of context-dependency in plant response to inoculation with mycorrhizal fungi. *Ecology Letters*, **13**, 394–407. 2, 17

Jost, L. (2007). Partitioning diversity into independent alpha and beta components. *Ecology*, **88**, 2427–2439. 22

Koleff, P., Gaston, K. J., and Lennon, J. J. (2003). Measuring beta diversity for presence–absence data. *Journal of Animal Ecology*, **72**, 367–382. 23

Kratzer, G. and Furrer, R. (2018). *varrank: an R package for variable ranking based on mutual information with applications to observed systemic datasets*. https://arxiv.org/abs/1804.07134, R package version 0.2. 15, 18

Lauber, C. L., Hamady, M., Knight, R., and Fierer, N. (2009). Pyrosequencing-based assessment of soil ph as a predictor of soil bacterial community structure at the continental scale. *Applied and Environmental Microbiology*, **75**, 5111–5120. 16, 58

Leff, J. W., Jones, S. E., Prober, S. M., Barberán, A., Borer, E. T., Firn, J. L., Harpole, W. S., Hobbie, S. E., Hofmockel, K. S., Knops, J. M. H., McCulley, R. L., La Pierre, K., Risch, A. C., Seabloom, E. W., Schütz, M., Steenbock, C., Stevens, C. J., and Fierer, N. (2015). Consistent responses of soil microbial communities to elevated nutrient inputs

in grasslands across the globe. *Proceedings of the National Academy of Sciences of the United States of America*, **112**, 10967–10972. 58

Legendre, P. and Anderson, M. J. (1999). Distance-based redundancy analysis: testing multispecies responses in multifactorial ecological experiments. *Ecological Monographs*, **69**, 1–24. 27, 28, 62

Legendre, P. and Gallagher, E. D. (2001). Ecologically meaningful transformations for ordination of species data. *Oecologia*, **129**, 271–280. 24, 25, 26, 34, 37, 42, 56

Lindahl, B. D., Nilsson, R. H., Tedersoo, L., Abarenkov, K., Carlsen, T., Kjøller, R., Kõljalg, U., Pennanen, T., Rosendahl, S., Stenlid, J., and Kauserud, H. (2013). Fungal community analysis by high-throughput sequencing of amplified markers – a user's guide. *New Phytologist*, **199**, 288–299. 8

Martin-Laurent, F., Philippot, L., Hallet, S., Chaussod, R., Germon, J., Soulas, G., and Catroux, G. (2001). DNA extraction from soils: old bias for new microbial diversity analysis methods. *Applied and Environmental Microbiology*, **67**, 2354–2359. 33, 53

McMurdie, P. J. and Holmes, S. (2013). phyloseq: an R package for reproducible interactive analysis and graphics of microbiome census data. *PLOS ONE*, **8**, e61217. 34

McMurdie, P. J. and Holmes, S. (2019a). *phyloseq: An R package for reproducible interactive analysis and graphics of microbiome census data*. R package version 1.28.0. 10

McMurdie, P. J. and Holmes, S. (2019b). Phyloseq Tutorial. https://joey711.github.io/phyloseq/import-data.html, accessed: 13.05.2019. 34

Moeys, J. (2018). *soiltexture: Functions for Soil Texture Plot, Classification and Transformation*. R package version 1.5.1. 16

Naylor, D., DeGraaf, S., Purdom, E., and Coleman-Derr, D. (2017). Drought and host selection influence bacterial community dynamics in the grass root microbiome. *The ISME journal*, **11**, 2691. 59

Nichols, R. V., Vollmers, C., Newsom, L. A., Wang, Y., Heintzman, P. D., Leighton, M., Green, R. E., and Shapiro, B. (2018). Minimizing polymerase biases in metabarcoding. *Molecular Ecology Resources*, **18**, 927–939. 54

Oksanen, J., Blanchet, F. G., Friendly, M., Kindt, R., Legendre, P., McGlinn, D., Minchin, P. R., O'Hara, R. B., Simpson, G. L., Solymos, P., Stevens, M. H. H., Szoecs, E., and Wagner, H. (2019). *vegan: Community Ecology Package*. R package version 2.5-4. 29

Ramette, A. (2007). Multivariate analyses in microbial ecology. *FEMS Microbiology Ecology*, **62**, 142–160. 25

Richner, W., Sinaj, S., Carlen, C., Flisch, R., Gilli, C., Huguenin-Elie, O., Kuster, T., Latsch, A., Mayer, J., Neuweiler, R.,*et al.* (2017). Grundlagen für die Düngung landwirtschaftlicher Kulturen in der Schweiz (GRUD 2017). *Agrarforschung Schweiz*, **8**, 47–66. 13, 19

Rousk, J., B*r*a*r*ath, E., Brookes, P. C., Lauber, C. L., Lozupone, C., Caporaso, J. G., Knight, R., and Fierer, N. (2010). Soil bacterial and fungal communities across a ph gradient in an arable soil. *The ISME Journal*, **4**, 1340. 58

Schlaeppi, K., Bender, F. S., Mascher, F., Russo, G., Patrignani, A., Camenzind, T., Hempel, S., Rillig, M. C., and Heijden, M. G. A. (2016). High-resolution community profiling of arbuscular mycorrhizal fungi. *New Phytologist*, **212**, 780–791. 54

Schloss, P. D. and Handelsman, J. (2005). Introducing DOTUR, a computer program for defining operational taxonomic units and estimating species richness. *Applied and Environmental Microbiology*, **71**, 1501–1506. 9, 31

Sheldon, A. L. (1969). Equitability indices: dependence on the species count. *Ecology*, **50**, 466–467. 23

Ter Braak, C. J. (1986). Canonical correspondence analysis: a new eigenvector technique for multivariate direct gradient analysis. *Ecology*, **67**, 1167–1179. 26, 27

Van Geel, M., Jacquemyn, H., Plue, J., Saar, L., Kasari, L., Peeters, G., van Acker, K., Honnay, O., and Ceulemans, T. (2018). Abiotic rather than biotic filtering shapes the arbuscular mycorrhizal fungal communities of European seminatural grasslands. *New Phytologist*, **220**, 1262–1272. 16, 58

White, T. J., Bruns, T., Lee, S., and Taylor, J. W. (1990). *PCR Protocols: a Guide to Methods and Applications*, chapter Amplification and direct sequencing of fungal ribosomal RNA genes for phylogenetics, 315–322. Academic Press San Diego, CA. 8

Whittaker, R. H. (1972). Evolution and measurement of species diversity. *Taxon*, **21**, 213–251. 21, 32

Zhang, S., Lehmann, A., Zheng, W., You, Z., and Rillig, M. C. (2019). Arbuscular mycorrhizal fungi increase grain yields: a meta-analysis. *New Phytologist*, **222**, 543–555. 2, 13

Zhang, Y. and Thas, O. (2016). Constrained ordination analysis with enrichment of bell-shaped response functions. *PLOS ONE*, **11**, e0154079. 27