Statistical Analysis Report

Biodiversity (based on amplicon data)

Biomcare ApS

09/06/2024

Customer	Innovation Centre for Organic Farming, Tove Mariegaard Pedersen
Customer ID	DA00204
Project	The microbial community of the field (2021, 2022 and 2023).
Sample Type	Soil
Number of samples	102 samples
Type of data	Marker gene sequencing of bacteria (16S) and fungi (ITS)

Association of biodiversity with variables of interest

With the marker gene (amplicon) sequencing data that was generated for two markers (ITS for fungi and 16S rRNA for bacteria) across fields, we calculate measures of biodiversity. The technical term in microbiome bioinformatic context for the measures is ‘alpha diversity’ and we calculate the two measures observed (or richness) and Shannon. The measures are introduced in **Report 2**.

We use the amplicon data specifically for biodiversity analyses as it has some advantage over shotgun sequencing data when it comes to diversity estimations. Amplicon data is more sensitive to less abundant taxa because of the targeted nature of the method (we use DNA-primers to fish for the DNA from the type of organism we want to detect (here fungi or bacteria)), and because we estimate the diversity based on all detected organisms independent of their presence in a reference database, we can include many more less known organisms in the measures resulting in a measures that better reflect the true diversity and not just the “diversity of well known organisms”.

In the report all evaluations will be performed for both measures, for both bacteria and fungi.

In the statistical analyses, we adjust for year when deemed sensible. That is in most cases but not for variables that are closely entangled with year such as the 3 measures of climate.

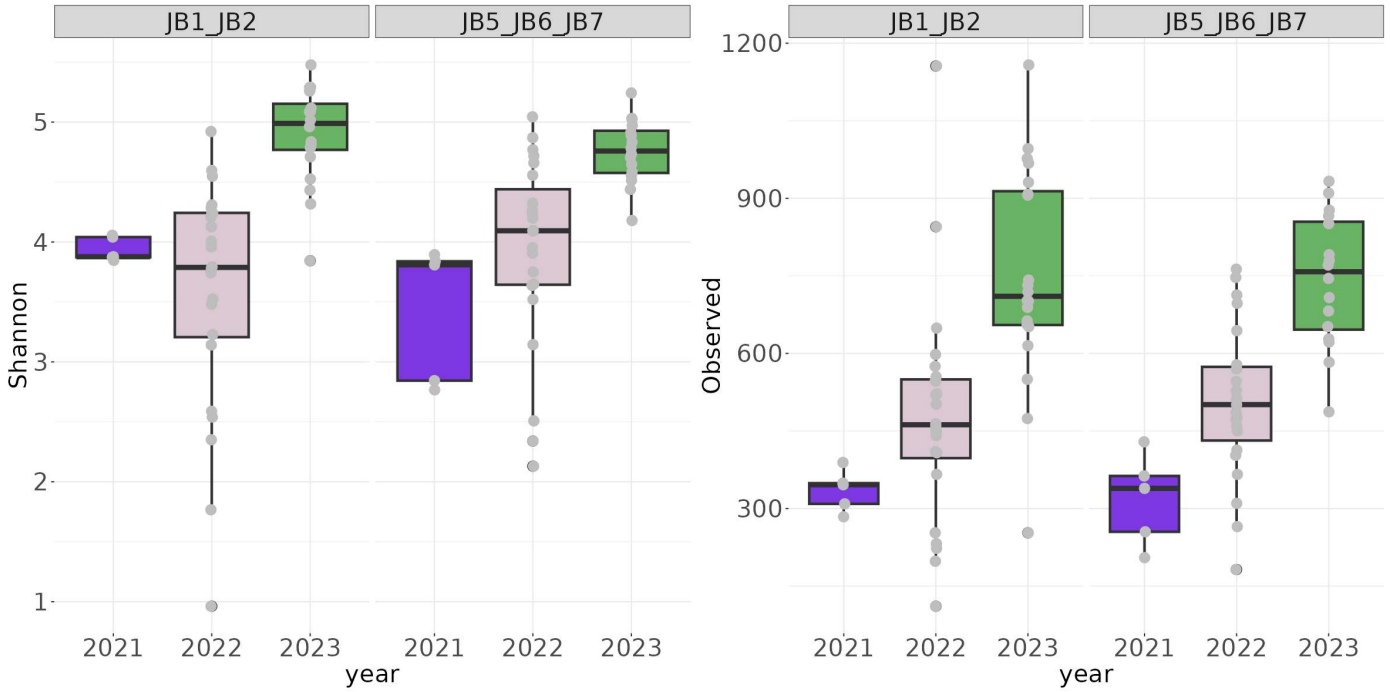
Alpha-diversity in relation to each environmental variable

In the following plots the measures of biodiversity is evaluated in relation to each key environmental variable. Following the plots is tables with the statistical analysis of the relations.

Grouping variables

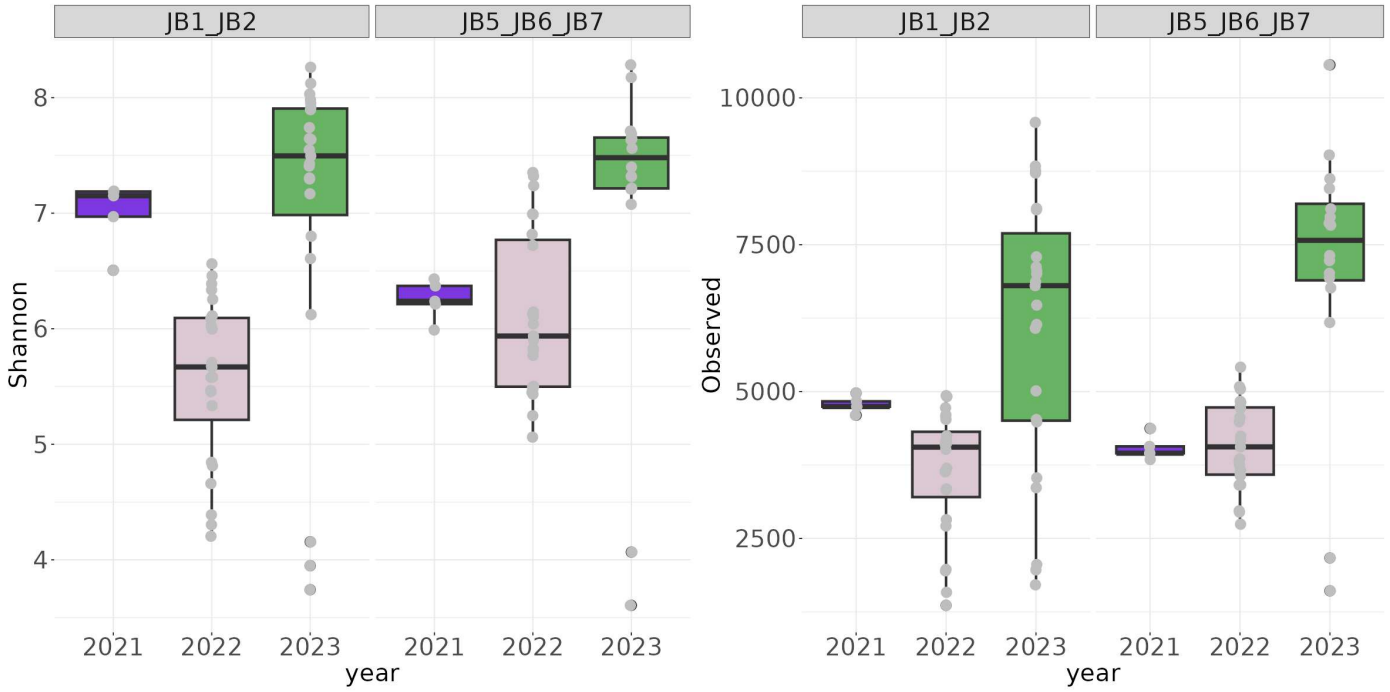
Year | JB groups | Geographic_location_letters | Earthworm_status | Cold soil | Compact soil | Field well drained | Mulching_of_straw | Clovergrass within 3 years | No plough | Conservation Agriculture | Years since plowing | Organic farm | Years since turning organic | Livestock

Livestock manure | Commercial fertilizer | Vinasse | Cast | Degassed fertilizer | Chalked | Haveparkaffald



)

Figure 1: Illustration of the fungal diversity levels across the levels or values of the environmental variable.

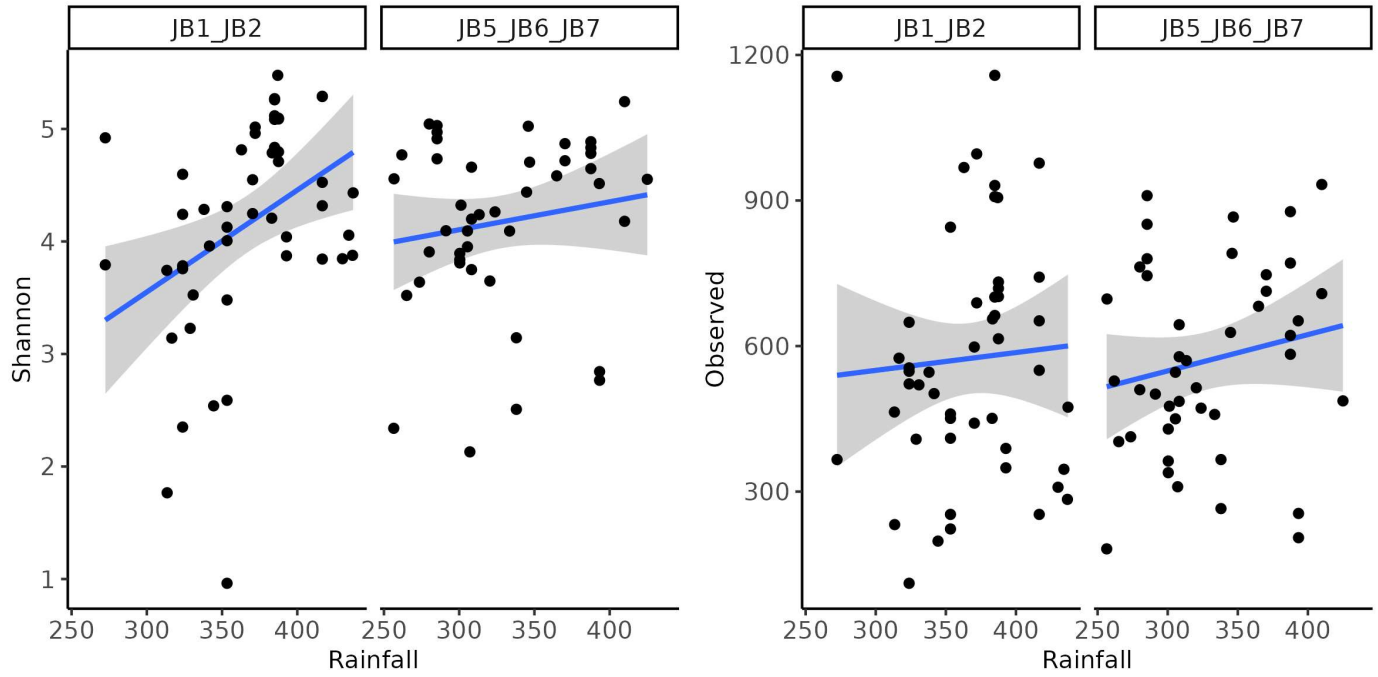


)

Figure 2: Illustration of the bacterial diversity levels across the levels or values of the environmental variable.

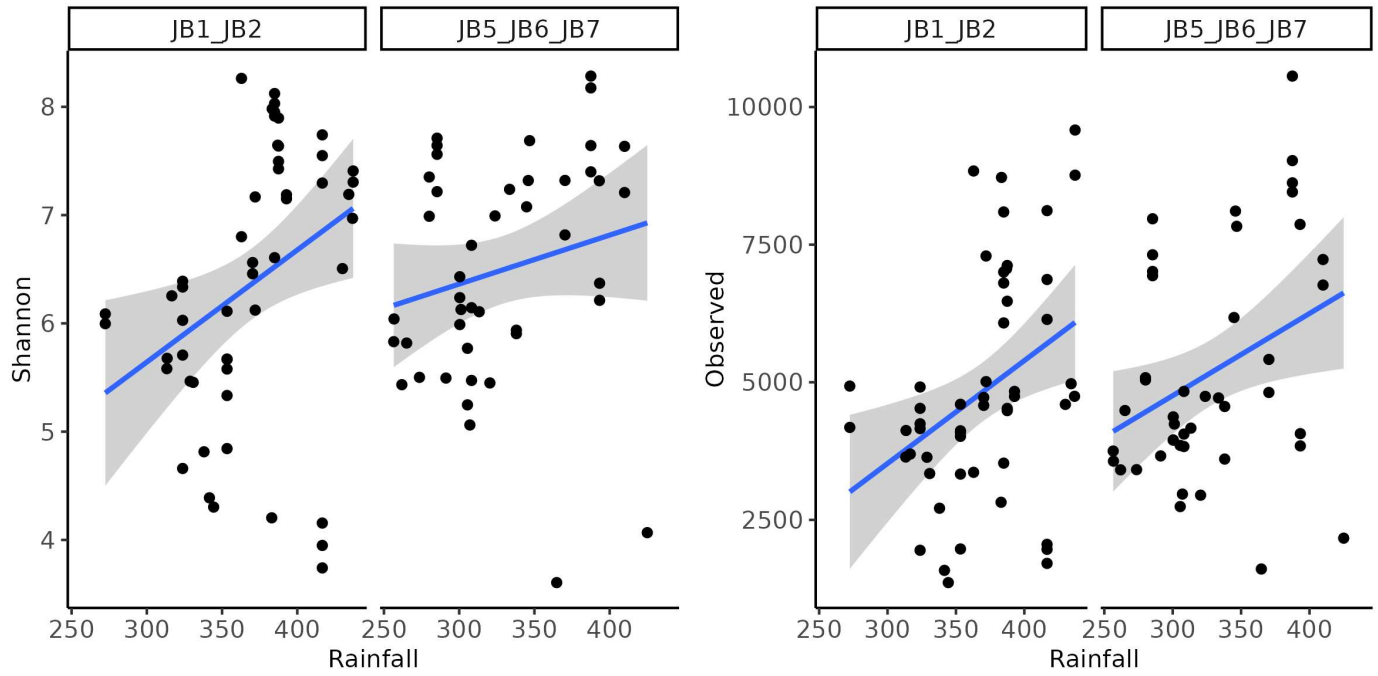
Continous variables

Rainfall | Average drought index | Average temp. | Rt | Phosphorus | Potassium | Magnesium | Cobber | Organic material (perc) | Clay (perc) | Nitrogen (perc)



)

Figure 45: Illustration of the fungal diversity levels across the levels or values of the environmental variable.



)

Figure 46: Illustration of the bacterial diversity levels across the levels or values of the environmental variable.

Statistical assessment

An analysis of Variance Model (ANOVA) was used to evaluate if the mean diversity differed significantly between the levels of grouping variables, and a robust linear regression was used to assess the relationship for continuous variables.

Bacterial biodiversity | Fungal biodiversity

Bacterial diversity in JB 1-2 (group variables)

	Observed	Shannon

	Observed				Shannon			
Variable	Df	Sum.Sq	F.value	P	Df	Sum.Sq	F.value	P
year	2	69594264.879	11.696	7.04e-05	2	28.233	13.688	1.89e-05
Geographic_location_letters	3	11948246.45	1.369	2.64e-01	3	9.591	3.592	2.05e-02
Earthworm_status	1	1379862.027	0.459	5.01e-01	1	1.296	1.264	2.67e-01
Cold_soil	1	906259.454	0.3	5.86e-01	1	0.57	0.548	4.63e-01
Compact_soil	1	1861081.782	0.621	4.35e-01	1	0	0	9.95e-01
field_well_drained	1	6169739.726	2.121	1.52e-01	1	4.399	4.577	3.75e-02
Mulching_of_straw	1	3523989.593	1.189	2.81e-01	1	3.239	3.287	7.61e-02
Clovergrass_within_3_years	1	8425307.8	2.944	9.26e-02	1	3.339	3.396	7.15e-02
No_plough	1	7815834.81	2.719	1.06e-01	1	4.81	5.05	2.93e-02
ConservationAgriculture	1	664444.458	0.22	6.41e-01	1	0.495	0.475	4.94e-01
Years_since_plowing	1	1083006.655	0.359	5.52e-01	1	1.285	1.252	2.69e-01
Organic_farm	1	13038759.157	4.715	3.49e-02	1	4.709	4.932	3.11e-02
Years_since_turning_organic	1	9346864.523	3.083	9.14e-02	1	2.64	2.701	1.13e-01
Livestock	1	3853268.625	1.303	2.59e-01	1	0.036	0.034	8.54e-01
Livestock_manure	1	4692716.897	1.597	2.12e-01	1	2.559	2.561	1.16e-01
Commercial.fertilizer	1	13038759.157	4.715	3.49e-02	1	4.709	4.932	3.11e-02
Vinasse	1	1236101.057	0.41	5.25e-01	1	0.11	0.105	7.47e-01
Cast	1	827080.958	0.274	6.03e-01	1	0.619	0.595	4.44e-01

	Observed				Shannon			
Degassed.fertilizer	1	1795690.751	0.599	4.43e-01	1	4.266	4.426	4.07e-02
Haveparkaffald	1	4895546.113	1.668	2.03e-01	1	2.922	2.945	9.26e-02
Chalked	1	4088855.909	1.385	2.45e-01	1	2.358	2.349	1.32e-01

Table 1: Results from ANOVA analysis across JB1 and JB2 fields. The table shows results from ANOVA analyses including samples from all fields. The table shows the obtained statistical values for each of the environmental variables (rows) and the three microbiome features (columns).

Bacterial diversity in JB 5-7 (group variables)

	Observed				Shannon			
Variable	**Df**	**Sum.Sq**	**F.value**	**P**	**Df**	**Sum.Sq**	**F.value**	**P**
year	2	93421590.188	21.125	4.95e-07	2	9.322	5.158	1.00e-02
Geographic_location_letters	5	31387893.889	3.813	7.12e-03	5	10.642	2.902	2.66e-02
Earthworm_status	1	850492.556	0.379	5.42e-01	1	0.163	0.177	6.76e-01
Cold_soil	1	444019.226	0.197	6.60e-01	1	0.575	0.631	4.32e-01
Compact_soil	1	1088345.624	0.486	4.90e-01	1	0.784	0.865	3.58e-01
field_well_drained	1	418499.415	0.186	6.69e-01	1	1.305	1.46	2.34e-01
Mulching_of_straw	1	62457.823	0.028	8.69e-01	1	0.055	0.059	8.09e-01
Clovergrass_within_3_years	1	1871569.094	0.843	3.64e-01	1	1.204	1.344	2.53e-01
No_plough	1	503309.387	0.223	6.39e-01	1	0.299	0.326	5.71e-01
ConservationAgriculture	1	4.07	0	9.99e-01	1	0.003	0.003	9.56e-01
Years_since_plowing	1	2075077.409	0.937	3.39e-01	1	1.691	1.913	1.74e-01
Organic_farm	1	35602.949	0.016	9.01e-01	1	0.027	0.029	8.66e-01
Years_since_turning_organic	1	323532.937	0.142	7.10e-01	1	0.422	0.455	5.07e-01

	Observed				Shannon			
Livestock	1	3026613.511	1.381	2.47e-01	1	1.047	1.164	2.87e-01
Livestock_manure	1	580407.997	0.258	6.14e-01	1	0.642	0.705	4.06e-01
Commercial.fertilizer	1	850749.949	0.379	5.42e-01	1	0.048	0.051	8.22e-01
Vinasse	1	4645371.412	2.16	1.49e-01	1	0.586	0.643	4.27e-01
Cast	1	960874.506	0.428	5.16e-01	1	0.789	0.87	3.57e-01
Degassed.fertilizer	1	48100.834	0.021	8.85e-01	1	0.002	0.003	9.59e-01
Haveparkaffald	1	1128285.917	0.504	4.82e-01	1	0.234	0.254	6.17e-01
Chalked	1	2890418.499	1.317	2.58e-01	1	1.954	2.227	1.43e-01

Table 2: Results from ANOVA analysis across JB5, JB6 and JB7 fields. The table shows results from ANOVA analyses including samples from all fields. The table shows the obtained statistical values for each of the environmental variables (rows) and the three microbiome features (columns).

Bacterial diversity in JB 1-2 (continous variables)

	Observed				Shannon			
Variable	**Estimate**	**SE**	**t.value**	**P**	**Estimate**	**SE**	**t.value**	**P**
Rainfall	19.97	9.867	2.024	4.83e-02	0.016	0.004	4.605	2.87e-05
Average_drought_index	-925.764	148.673	-6.227	9.71e-08	-0.566	0.075	-7.504	9.79e-10
Average_temp.	1760.46	1070.396	1.645	1.06e-01	0.495	0.564	0.877	3.85e-01
Rt	67.809	602.178	0.113	9.11e-01	-0.147	0.264	-0.556	5.81e-01
Phosphorus	322.922	244.104	1.323	1.92e-01	0.137	0.071	1.935	5.89e-02
Potassium	-37.779	118.157	-0.32	7.51e-01	-0.034	0.047	-0.716	4.78e-01
Magnesium	73.21	197.046	0.372	7.12e-01	0.001	0.136	0.006	9.95e-01
Cobber	4.475	359.886	0.012	9.90e-01	0.045	0.128	0.353	7.26e-01
Organic_material_perc	-126.286	339.911	-0.372	7.12e-01	-0.007	0.075	-0.088	9.30e-01
Clay_perc	-35.982	247.215	-0.146	8.85e-01	-0.039	0.068	-0.565	5.74e-01
Nitrogen_perc	-5476.694	9886.249	-0.554	5.82e-01	-0.922	1.847	-0.499	6.20e-01

Table 3: Results from robust linear regression analysis across JB1 and JB2 fields. The table shows results from the robust regression analyses including samples from all fields. The table shows the obtained statistical values for each of the environmental variables (rows) and the three microbiome features (columns).

Bacterial diversity in JB 5-7 (continous variables)

	Observed				Shannon			
Variable	**Estimate**	**SE**	**t.value**	**P**	**Estimate**	**SE**	**t.value**	**P**
Rainfall	19.928	19.665	1.013	3.17e-01	0.009	0.003	3.188	2.71e-03
Average_drought_index	-1066.277	118.627	-8.988	2.45e-11	-0.403	0.067	-5.976	4.29e-07
Average_temp.	958.517	1345.807	0.712	4.80e-01	-0.155	0.36	-0.432	6.68e-01
Rt	-317.831	288.481	-1.102	2.77e-01	-0.351	0.426	-0.825	4.15e-01
Phosphorus	105.956	98.452	1.076	2.88e-01	0.129	0.07	1.837	7.36e-02
Potassium	19.113	50.441	0.379	7.07e-01	0.017	0.032	0.524	6.03e-01
Magnesium	63.145	45.568	1.386	1.74e-01	0.049	0.068	0.713	4.80e-01
Cobber	195.205	238.496	0.818	4.18e-01	0.174	0.135	1.287	2.06e-01
Organic_material_perc	40.327	180.644	0.223	8.24e-01	-0.002	0.228	-0.009	9.93e-01
Clay_perc	18.919	61.988	0.305	7.62e-01	0.002	0.071	0.027	9.79e-01
Nitrogen_perc	-4416.239	4158.094	-1.062	2.95e-01	-2.18	6.142	-0.355	7.25e-01

Table 4: Results from robust linear regression analysis across JB5, JB6 and JB7 fields. The table shows results from the robust regression analyses including samples from all fields. The table shows the obtained statistical values for each of the environmental variables (rows) and the three microbiome features (columns).

Detecting most imporant variables usign mashine learning

The rapid advancement of machine learning technologies has provided new tools for understanding complex agricultural systems. One such method, the Random Forest algorithm, is particularly effective in identifying key variables that influence continuous outcome variables. In this section, we apply the Random Forest technique to identify the variables most important for shaping soil diversity.

Random Forest is an ensemble learning method that constructs multiple decision trees during training and merges their outputs to improve accuracy and control overfitting. Each tree in the forest considers a random subset of the variables, allowing the algorithm to identify which factors consistently play a crucial role in determining the outcome. This capability makes Random Forest especially useful in scenarios with a large number of potential influencing factors and complex, non-linear relationships, such as those found in agricultural ecosystems.

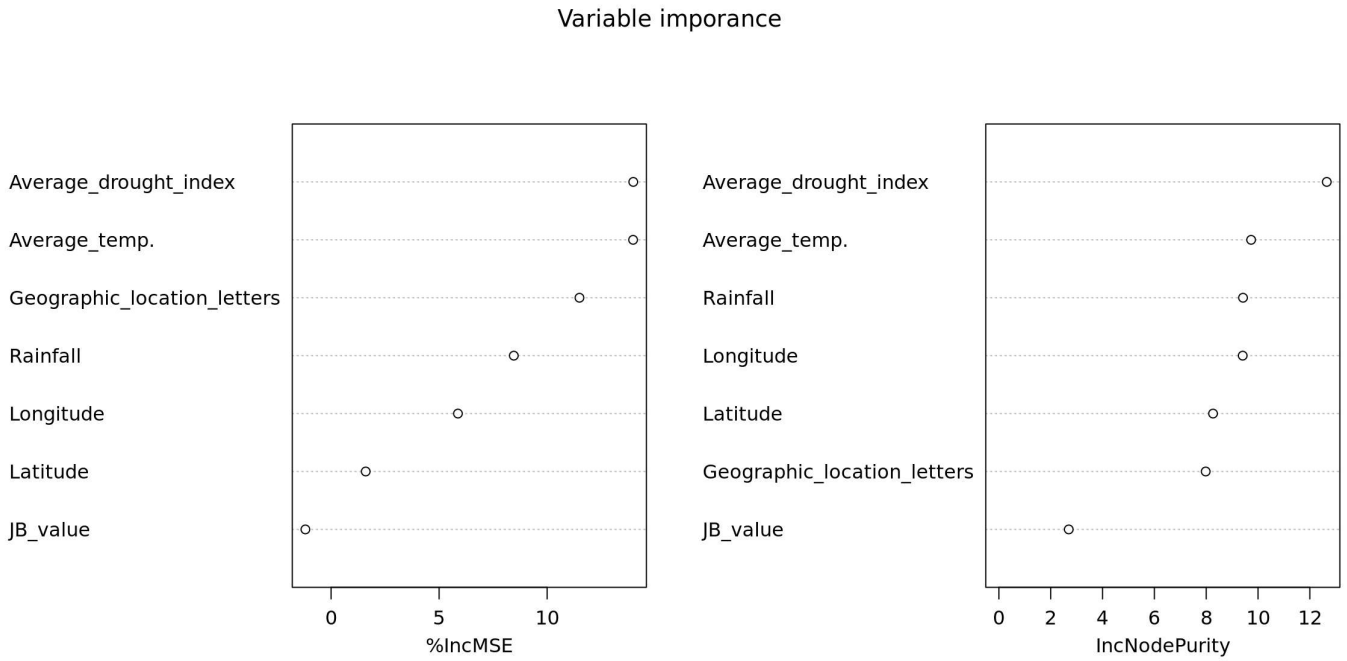
By utilizing this method, we can pinpoint the most significant factors that affect soil microbiome diversity.

Fungal Shannon diversity

We perform three RF tests for each diversity measure to access the imporatance of the predictors.

1. Using main variables on climate and location.
2. Using all variables (excluding 2 with many missing values)
3. Using all but removing the effect of year.

1) Using main variables on climate and location.



Higher the value of mean decrease accuracy or mean decrease gini score , higher the importance of the variable in the model.

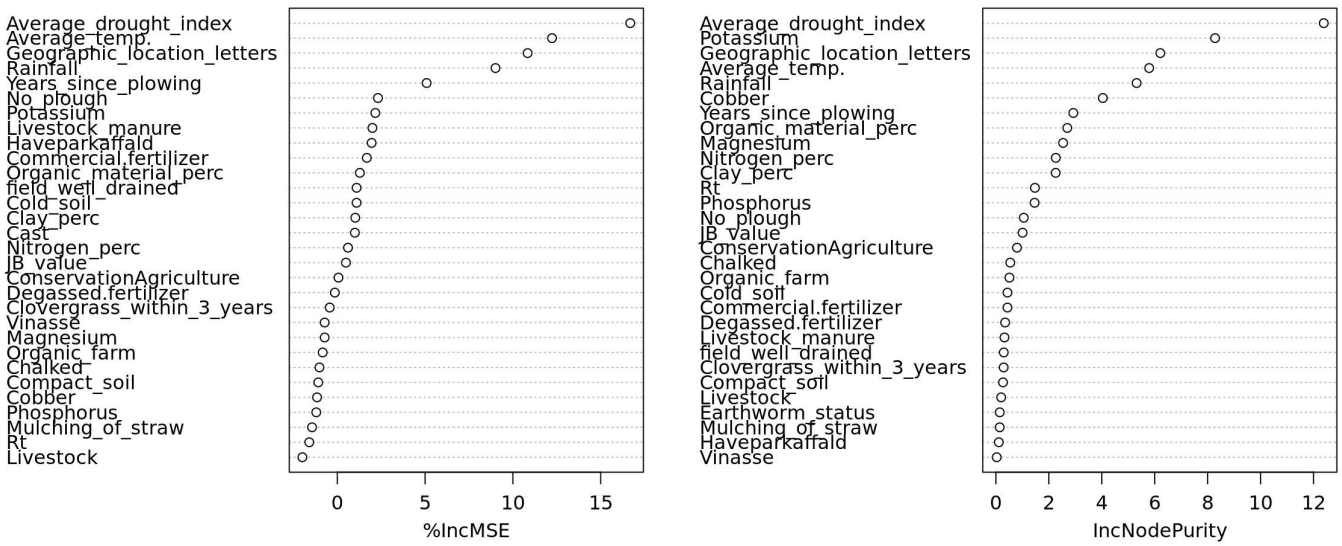
- Mean Decrease Accuracy - How much the model accuracy decreases if we drop that variable.
- Mean Decrease Gini - Measure of variable importance based on the Gini impurity index used for the calculation of splits in trees.

Prediction of diversity

Now we look if the variables can predict the value of the outcome measure, calculated as a % of variance explained by the tested variables.

```
##
## Call:
##  randomForest(formula = Shannon ~ ., data = meta_asv_count_sub,      ntree = 500, importance = TRUE)
##                Type of random forest: regression
##                      Number of trees: 500
## No. of variables tried at each split: 2
##
##           Mean of squared residuals: 0.5673058
##                     % Var explained: 22.92
```

2) Using all variables (excluding 2 with many missing values).



Higher the value of mean decrease accuracy or mean decrease gini score , higher the importance of the variable in the model.

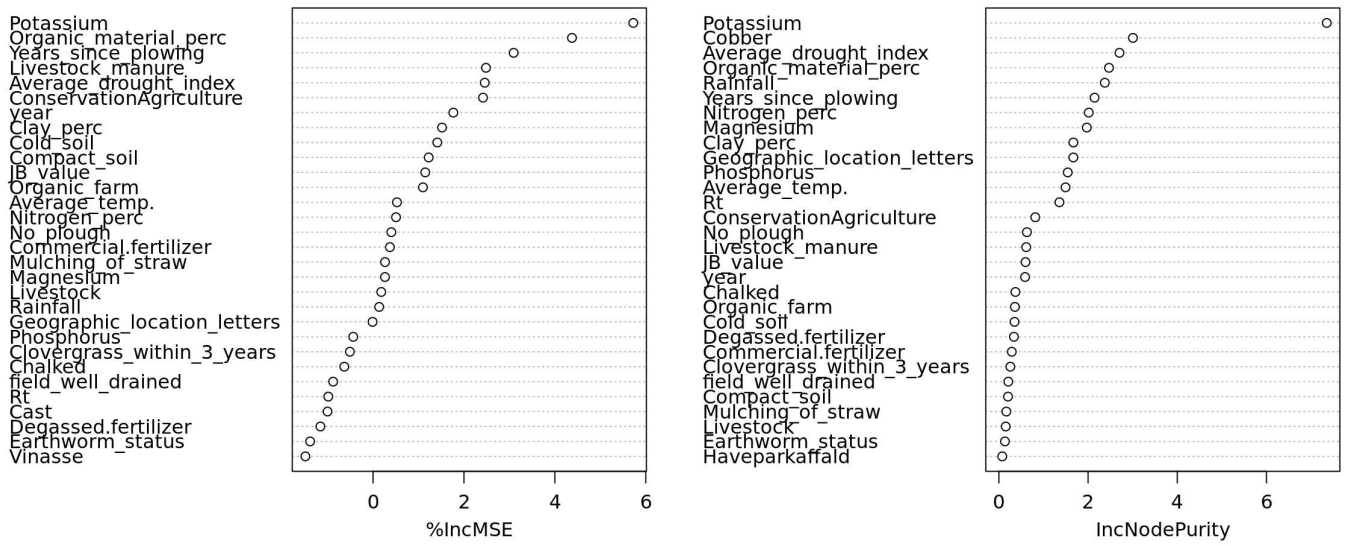
- Mean Decrease Accuracy - How much the model accuracy decreases if we drop that variable.
- Mean Decrease Gini - Measure of variable importance based on the Gini impurity index used for the calculation of splits in trees.

Prediction of diversity

Now we look if the variables can predict the value of the outcome measure, calculated as a % of variance explained by the tested variables.

```
##
## Call:
##  randomForest(formula = Shannon ~ ., data = meta_asv_count_sub_clean,      ntree = 500, importance =
TRUE)
##                Type of random forest: regression
##                      Number of trees: 500
## No. of variables tried at each split: 10
##
##           Mean of squared residuals: 0.4836676
##                     % Var explained: 34.28
```

3) Using all but removing the effect of year.



Higher the value of mean decrease accuracy or mean decrease gini score , higher the importance of the variable in the model.

- Mean Decrease Accuracy - How much the model accuracy decreases if we drop that variable.
- Mean Decrease Gini - Measure of variable importance based on the Gini impurity index used for the calculation of splits in trees.

Prediction of diversity

Now we look if the variables can predict the value of the outcome measure, calculated as a % of variance explained by the tested variables.

```
##
## Call:
##  randomForest(formula = residuals ~ ., data = meta_asv_count_sub_clean,      ntree = 500, importance
= TRUE)
##                Type of random forest: regression
##                      Number of trees: 500
## No. of variables tried at each split: 10
##
##           Mean of squared residuals: 0.4806235
##                     % Var explained: -6.02
```

Version information

Table 9: List of used software including the used R-programming environment packages.

Package	Version	Package	Version
OS	Ubuntu 20.04.4 LTS	**glue**	1.7.0
R	4.3.3	**nlme**	3.1-164
rstudioapi	0.16.0	**rhdf5filters**	1.14.1
magrittr	2.0.3	**Rtsne**	0.17
TH.data	1.1-2	**cluster**	2.1.6
estimability	1.5	**reshape2**	1.4.4
nloptr	2.0.3	**ade4**	1.7-22

Package	Version	Package	Version
rmarkdown	2.26	**generics**	0.1.3
zlibbioc	1.48.2	**gtable**	0.3.5
vctrs	0.6.5	**tzdb**	0.4.0
multtest	2.58.0	**hms**	1.1.3
minqa	1.2.6	**xml2**	1.3.6
RCurl	1.98-1.14	**utf8**	1.2.4
htmltools	0.5.8.1	**foreach**	1.5.2
S4Arrays	1.2.1	**pillar**	1.9.0
curl	5.2.1	**robustbase**	0.99-2
cellranger	1.1.0	**splines**	4.3.3
Rhdf5lib	1.24.2	**survival**	3.5-8
SparseArray	1.2.4	**deldir**	2.0-4
rhdf5	2.46.1	**tidyselect**	1.2.1
sass	0.4.9	**V8**	4.4.2
bslib	0.7.0	**svglite**	2.1.3
plyr	1.8.9	**xfun**	0.43
sandwich	3.1-0	**DEoptimR**	1.1-3
zoo	1.8-12	**stringi**	1.8.4
cachem	1.0.8	**yaml**	2.3.8
igraph	2.0.3	**boot**	1.3-30
lifecycle	1.0.4	**evaluate**	0.23
iterators	1.0.14	**codetools**	0.2-20
pkgconfig	2.0.3	**interp**	1.1-6
R6	2.5.1	**cli**	3.6.2
fastmap	1.1.1	**RcppParallel**	5.1.7
GenomeInfoDbData	1.2.11	**systemfonts**	1.0.6
digest	0.6.35	**xtable**	1.8-4
colorspace	2.1-0	**munsell**	0.5.1
hwriter	1.3.2.1	**jquerylib**	0.1.4
fansi	1.0.6	**coda**	0.19-4.1
timechange	0.3.0	**png**	0.1-8
abind	1.4-5	**parallel**	4.3.3
mgcv	1.9-1	**latticeExtra**	0.6-30
compiler	4.3.3	**jpeg**	0.1-10
withr	3.0.0	**bitops**	1.0-7
highr	0.10	**viridisLite**	0.4.2

Package	Version	Package	Version
MASS	7.3-60.0.1	**mvtnorm**	1.2-4
DelayedArray	0.28.0	**pcaPP**	2.0-4
biomformat	1.30.0	**crayon**	1.5.2
tools	4.3.3	**rlang**	1.1.3
rrcov	1.7-5	**mnormt**	2.1.1
ape	5.8	**multcomp**	1.4-25
zip	2.3.1		