Variation in Soil Microbial Communities

Chengzhu 'Charlie' Zhang
Chris Burgess
Joel Gorder

3/16/2017

Introduction

Soil microbes are critically important because they provide the functional backbone for many soil ecosystem services such as carbon and nitrogen cycles. The structure and composition of microbial communities have a large impact on the rates at which these ecosystem services are performed. However, discerning ecological information from soil microbial communities can be very difficult due to the fact that microbial communities are extremely diverse (Fierer et al. 2007).

One way of understanding the potential relationships between microbial communities and composition is to identify the underlying relationship between the environment and its microbial community composition. Several studies have explored different aspects of microbial community composition and its relationship with environment parameters.

There is an field-wide consensus that soil pH is one of the dominant predictors of microbial diversity (Fierer and Jackson 2006, Rousk et al 2010, Griffiths et al 2011); however, pH is not the only indicator for bacterial species composition. In a recent study along a 3700km transect along the Mongolian grasslands, Wang et al. (2015) found a strong correlation between community composition and aridity. Other studies have shown that land use influences microbial composition (Hartman et al 2015). These studies illustrate both the complex relationships microbial communities have with their environment and that the relationships governing their distribution generally cannot be explained by one universal factor.

In this study, we aimed to explore soil microbial communities and their correlation with environmental factors at seven forested locations across a regional latitudinal transect in Oregon. We were interested in discovering and defining several important factors influencing microbial community composition along this transect.

Methods

Study Site and Soil Measurements

The samples were collected from seven undisturbed forested locations across the Pacific Northwest along a latitudinal transect during the summer of 2011 (Figure 1). There were seven forested sites sampled with 3 biological replicates taken at each site bring a total of 21 samples. These forested sites have been studied since the 1980s and have been used to explore topics ranging from questions about primary productivity rates using remote sensing (Running et al. 1986) to exploring the relationship between forested soils and their productivity (Myrold et al. 1989). The eastern-most site is Cascade Head, which is in the Coastal Range. The primary plant cover at Cascade Head is mainly western hemlock, while western juniper dominates the vegetation cover at the most eastern site, Horse Ridge. The east-west transect also encompasses several different environmental gradients including precipitation (lush to arid), elevation (sea-level to mountain passes), and temperature (cold to hot). Along with this environmental gradient, there are several different biologically mediated gradients such as plant fauna and nitrogen availability.

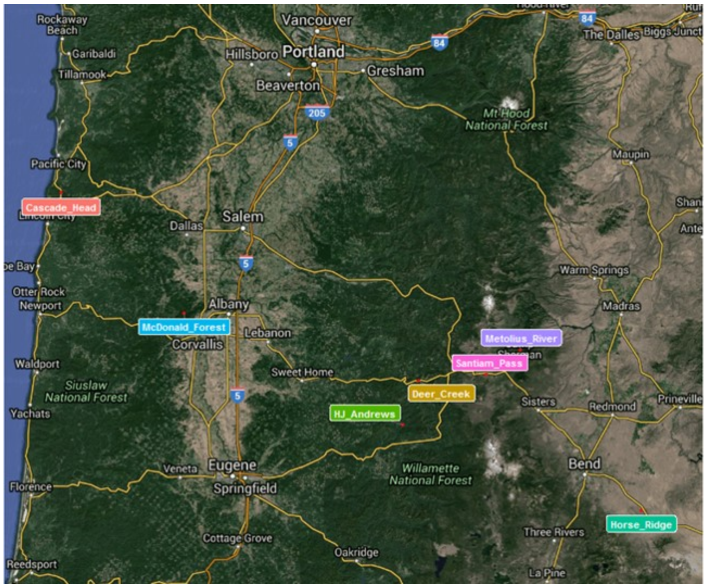


Figure 1: Oregon Sites

DNA and Informatic

DNA was extracted from each sample using an MoBio Kit. Amplicon sequencing of the 16s V4 region using 454 pyrosequencing at Argonne Lab. The sequences were then filtered and clustered into operational taxonomic units (OTUs) at a 97% threshold using QIIME and VSEARCH, an open source version of USEARCH (Rognes et al 2016). The data is compiled into a .BIOM file which had 3 components: an OTU table, phylogeny data, and metadata. The OTU table is a frequency or abundance table of all the OTUs in each sample. The phylogeny data is a table linking each OTU to its phylogenetic lineage. Lastly, the metadata is a table of covariates of environmental and soil characteristics for each sample.

Principal Coordinate Analysis (PCoA)

Principal Coordinates Analysis – a form of multidimensional scaling – is a method used to explore and to visualize similarities or dissimilarities of data. It begins by defining a similarity matrix or dissimilarity (or distance) matrix and assigns each item a location in a low-dimensional space.

PCoA tries to find the main axes through a matrix manipulations . It is uses singular value decomposition to calculate the eigenvalues and eigenvectors of a dissimilarity matrix. Each eigenvalue is associated an eigenvector, and there are as many eigenvectors and eigenvalues as there are rows in the initial matrix.

A dissimilarity matrix shows the distance between every possible pair of objects. PCoA is a set of data analysis techniques that displays the structure of (complex) distance-like data by mapping a high dimensional space onto a lower dimensional space while preseverving much of the variance or information. The goal of PCoA is to faithfully represent these distances with the lowest possible dimensional space. The result is a rotation of the data matrix along its principal coordinates: it does not change the positions of points relative to each other but maps them onto a new coordinate system.

Eigenvalue Analysis of PCoA

$$D^2 = -2B, \quad B = UV^TU, \quad \sqrt{\frac{1}{2}}D = \sqrt{B} = UV^{1/2}$$

.

Where D is the dissimilarity matrix, $d_{i,j}$ represents the distance or dissimilarity between observation i and observation j. B is the doubling centering on D^2, which is symmetric. When we want to find out the r dimensional configurations of the data in the space generated by the r principal components, we only used the first leading r spectrums and r columns of U to generate X. The following matrix we call is Configuration Matrix:

$$X = U_r\sqrt{V_r}$$

.

where V_r is the $r \times r$ sub-matrix of V and U_r is a matrix of size $\mathrm{N} \times r$.

Bray-Curtis Dissimilarity

In ecology and biology, the Bray-Curtis dissimilarity – named for J. Roger Bray and John T. Curtis – is a statistic used to quantify the compositional dissimilarity between two different sites, based on counts at each site. As defined by Bray & Curtis, the index of dissimilarity is:

$$BC_{i,j} = 1 - \left(\frac{2C_{i,j}}{S_i + S_j}\right)$$

where $C_{i,j}$ is the sum of the lesser values for only those species in common between both sites. S_i and S_j are the total number of specimens counted at both sites. The index reduces to $1 - 2C/2 = 1 - C$ where abundances at each site are expressed as a percentage.

The Bray-Curtis dissimilarity is bounded between 0 and 1, where 0 means the two sites have the same composition (i.e. they share all the species), and 1 means the two sites do not share any species.

The Bray-Curtis dissimilarity is often erroneously called a distance. It is not a distance since it does not satisfy triangle inequality, and should always be called a dissimilarity to avoid confusion. As a result, using PCoA is better than PCA which is only applicable to data for which the distance is appropriate.

Principal Coordinates regression

Our project was primarily concernted with the exploration of 14 environmental covariates and their ability to explain the diversity of species identified at 7 distince sites in Oregon. Naturally, this led us to consider a regression model built form the principle coordinates we obtained in the ordination discussed above. We chose to build models for the first three principal coordinates, which had captured a large portion of the variation in the dissimilarity matrix.

Because we wished to build well-fitting models that explained as much variability as possible, but remained simple in their biological interpretation, our criteria were *adjusted r-squared* (R_a^2) and parsimony. As an ancillary criterion, and an indicator of collinearity, we also reported the *variance inflation factor* (VIF). Mathematical formulations of these criteria are listed below.

$$R_a^2 = 1 - \left(\frac{n-1}{n-p}\right)\left(\frac{RSS}{SST}\right), \quad VIF = \left(\frac{1}{1 - R_j^2}\right)$$

Because the data were measured on differing scales, we scaled the data before model selection by subtracting the mean and dividing by the standard deviation for each variable. This created consistency among the resulting coefficients, and was easier to interpret than if we had left the variables in their original scales. We relied on the `regsubets()` function in the `leaps` package to automate the process of model selection. This function allowed us to set certain criteria and evaluate resulting models numerically and visually. This expedited model selection process, and allowed us to focus on selecting models based on biological interpretability.

Results

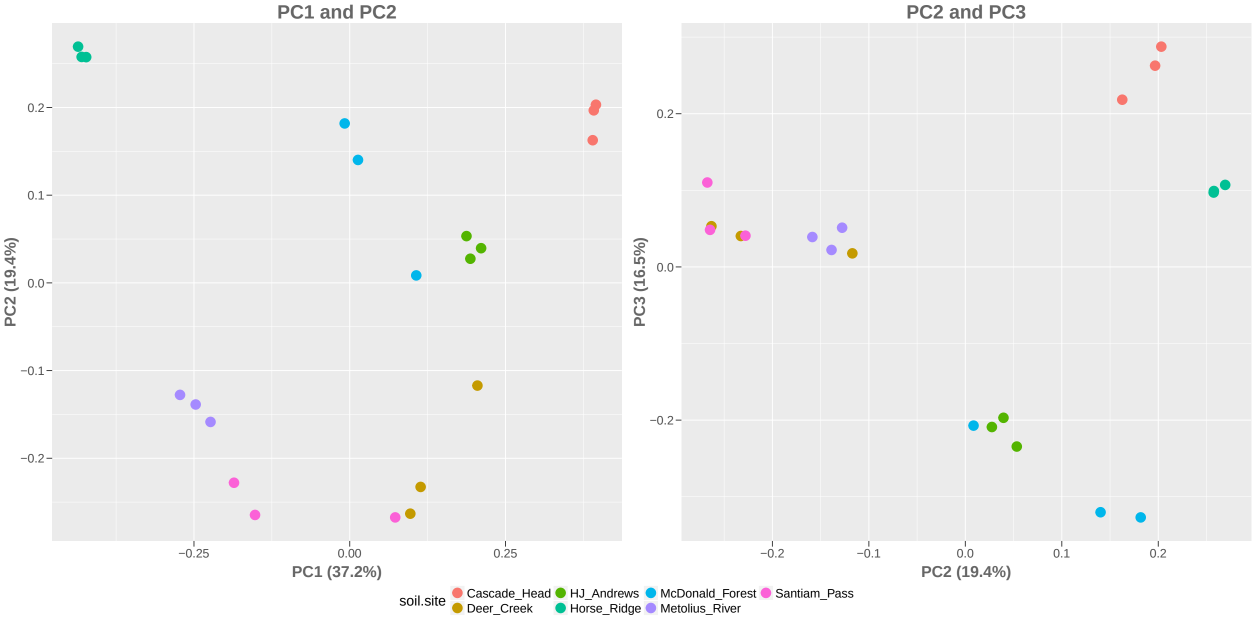
Eigenvalues/Variance

Eigenvalues	Var Explained	Cumulative Variance
1.4574198	0.3724227	0.3724227
0.7609199	0.1944422	0.5668649
0.6454430	0.1649337	0.7317986

Sites and Principal Coordinates

Sites	PC1	PC2	PC3
Cascade Head	0.3917409	0.1967276	0.2627136
Cascade Head	0.3952381	0.2031589	0.2875555
Cascade Head	0.3900419	0.1626820	0.2182895
Deer Creek	0.1138079	-0.2326888	0.0403550
Deer Creek	0.0972553	-0.2631860	0.0531385
Deer Creek	0.2050002	-0.1171128	0.0177164
HJ Andrews	0.1936658	0.0275291	-0.2089856
HJ Andrews	0.1874368	0.0533320	-0.2344749
HJ Andrews	0.2109052	0.0395541	-0.1969496
Horse Ridge	-0.4230368	0.2574331	0.0970832
Horse Ridge	-0.4360096	0.2693736	0.1070719
Horse Ridge	-0.4303004	0.2576540	0.0990871
MacDonald Forest	0.1068059	0.0085133	-0.2071417
MacDonald Forest	0.0132854	0.1401890	-0.3202395
MacDonald Forest	-0.0080956	0.1818021	-0.3269875
Metolious River	-0.2472694	-0.1386825	0.0221560
Metolious River	-0.2234232	-0.1585659	0.0390546
Metolious River	-0.2723869	-0.1277255	0.0511874
Santiam Pass	-0.1857308	-0.2279188	0.0407616
Santiam Pass	0.0730490	-0.2674424	0.1102203
Santiam Pass	-0.1519796	-0.2646262	0.0483882

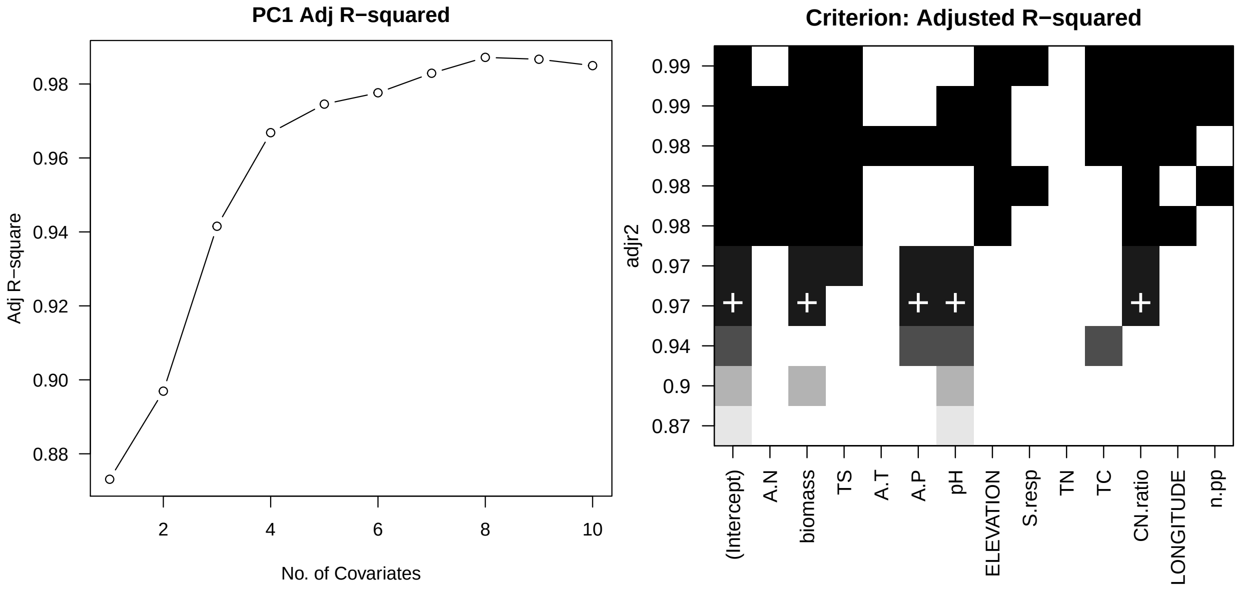
Ordination Results



Regression Results

Principal Coordinate 1

The plots below are generated using output from the `regsubsets()` function. The left plot is the R_a^2 by the number of covariates, and the right plot a visual representation of which covariates should be included to achieve this level of explanation.



From this output, it can be seen that the R_a^2 levels off after four covariates. The second plot indicates that the model should include: biomass, annual precipitation, pH, and the carbon/nitrogen ratio.

$$PC1 = \beta_0 + \beta_1 * biomass + \beta_2 * A.P + \beta_3 * pH + \beta_4 * CN.ratio$$

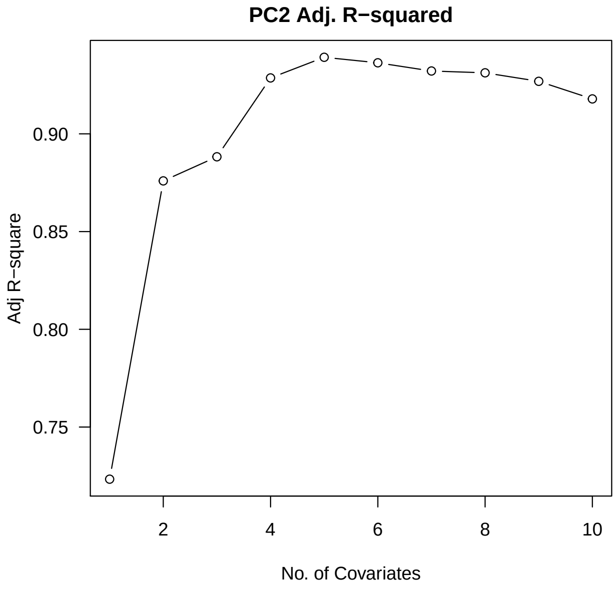
The table below gives the coefficients, standard errors, p values, and VIFs.

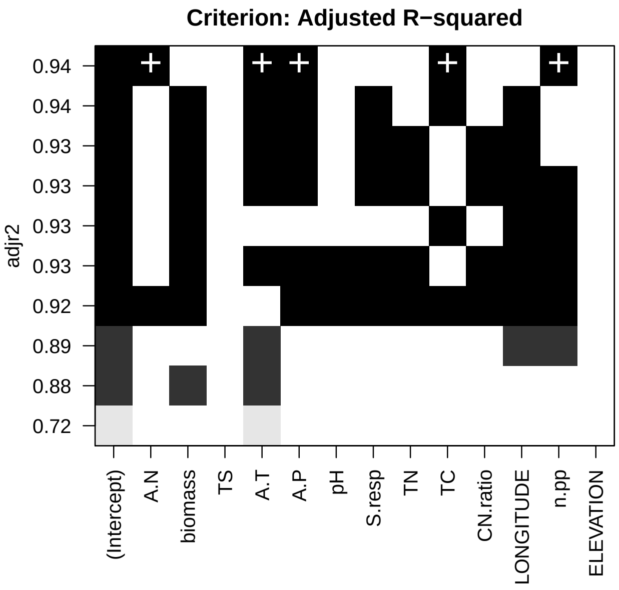
PC1 Regression Coefficients

	Estimate	Std. Error	t-stat.	p value	VIF
Intercept	0.0196643	0.0109272	1.799570	0.0951699	NA
biomass	-0.0653454	0.0153476	-4.257691	0.0009338	2.531463
Ann. Precipitation	0.0938434	0.0163737	5.731357	0.0000692	2.335517
pH	-0.2145230	0.0197341	-10.870693	0.0000001	3.454922
Carbon/Nitrogen ratio	-0.0561432	0.0149342	-3.759356	0.0023846	1.444483

Principal Coordinate 2

From the plots below, we find that covariate regression with principal coordinate 2 achieves a peak R_a^2 at 5 covariates.





According to our selection criteria, this model should include nitrogen, annual temperature, annual precipitation, total carbon, and net primary productivity.

$$PC2 = \beta_0 + \beta_1 * A.N + \beta_2 * A.T + \beta_3 * A.P + \beta_4 * TC + \beta_5 * N.pp$$

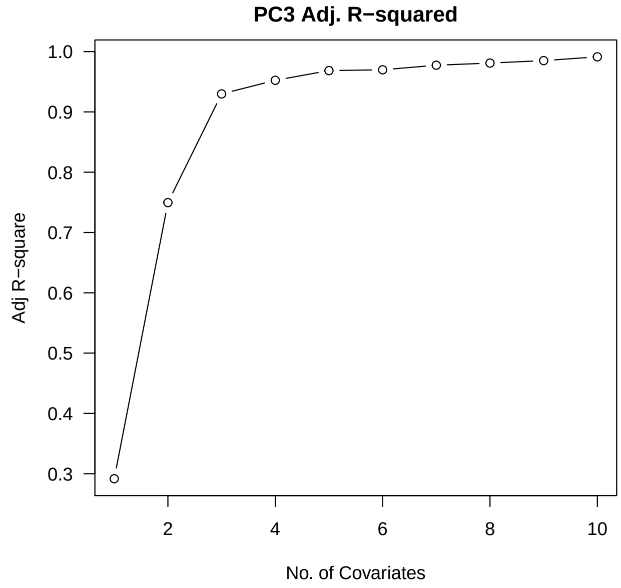
The table below gives the coefficient values, standard errors, p values, and VIFs.

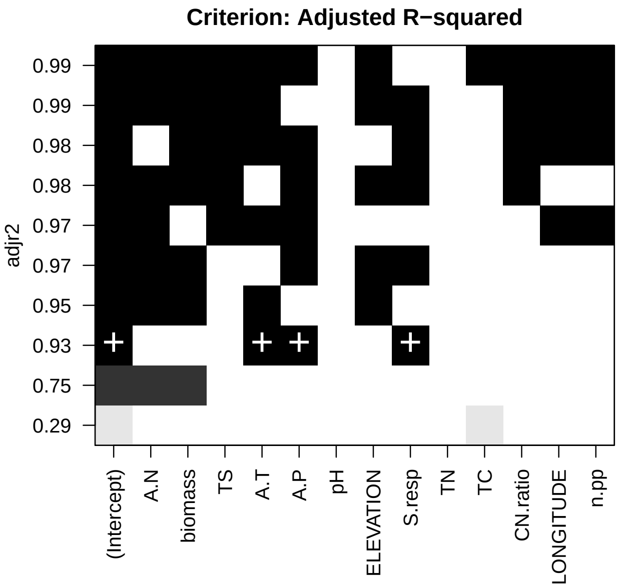
PC2 Regression Coefficients

	Estimate	Std. Error	t-stat.	p value	VIF
Intercept	-0.0466004	0.0129684	-3.593365	0.0036906	NA
Available Nitrogen	0.1468116	0.0305401	4.807176	0.0004283	8.494739
Ann. Temp.	0.1819893	0.0185619	9.804470	0.0000004	3.144050
Ann. Precip.	0.1329513	0.0262195	5.070697	0.0002749	5.075288
Total Carbon	0.1045463	0.0249532	4.189694	0.0012547	5.947764
Net primary Productivity	-0.2962229	0.0597299	-4.959376	0.0003311	25.477761

Principal Coordinate 3

Using the third principal coordinate as a regressand, we see the R_a^2 levelling off after 3 covariates.





From the right plot above, we can see that this model should include annual temperature, annual precipitation, and soil respiration.

$$PC3 = \beta_0 + \beta_1 * A.T + \beta_2 * A.P + \beta_3 S.resp$$

The table below shows the coeffecients, standard errors, p values, and VIFs.

PC3 Regression Coefficients

	Estimate	Std. Error	t-stat.	p value	VIF
Intercept	0.0000000	0.0113273	0.000000	1	NA
Ann. Temperature	-0.2360680	0.0169214	-13.950867	0	2.125341
Ann. Precipitation	-0.2309886	0.0249595	-9.254539	0	4.624111
Soil Respiration	0.3581184	0.0283121	12.648950	0	5.949782

Discussion

Ordination

In the Sites and Principal Coordinates table above, the row names here represent the site names. On a high level, we can see that for the most part that the PCoA axis values are consistent within each site—they share the same sign. The two exceptions to this are samples found in McDonald Forest and Santiam pass; this is not completely surprising given the heterogeneous nature of soils and an ecosystem in general. At Santiam Pass, there was substantial tree death due to pine beetle infestation. Though we tried to sample under trees that were unaffected the magnitude of the alterations to the ecosystem could have easily increase the variability in the soils in the area increasing the spatial heterogeneity. However, for the most part the variation we see in our OTU table does not seem to be coming from within biological replicates—which can easily be seen in the two PCoA figures. In both figures we see that for the most part the points cluster together by their sampling site. Therefore, there is likely some other underlying relationships contributing to the spatial variation we see in our OTU table.

In the Eigenvalues/Variance table listed above, we see that the cumulative variance explained by the first three principal coordinates is > 73%. Specifically, axis 1 explains 37.2%; axis 2 explains 19.4%; and axis 3 explains 16.5% of the variance in the dissimilarity table (i.e. species abundance).

Regression

PCoA1

In our first axis, we found a very strong relationship between microbial biomass Carbon, soil pH, annual precipitation and Carbon:Nitrogen ratio ($R_a^2 = 0.97$). From the regression results listed above, we see that the coefficients are significant ($p < 0.05$) and the VIFs are low, indicating no collinearity. These covariates are somewhat related, and, taken as a whole, offer a description of ecosystem differences. We believe they explain the major differences observed at each site from an ecosystem level. In broad brush strokes these characteristics describe underlying differences from one ecosystem to the next. It is worth noting that several of these variables are highly correlated in ecology. Soil pH and annual precipitation are inversely related. The carbon to nitrogen ratio tends to be an index of how accessible the nutrients are in the system, which, in turn, tends to identify the type community in the system.

PCoA2

The second model achieves an $R_a^2 = 0.94$. It is interesting to note that we found similar biological explanatory variables on our second axis: total carbon and nitrogen, mean annual precipitation and mean annual temperature. From the regression results listed above, we see that the coefficients are signficant, and the VIFs aren't very concerning. These parameters taken together explain the general productivity and energy going into the soil itself. Generally, the higher net primary productivity of a system, the more energy that is flowing into a system, which in turn leads to communities with different survival strategies. High energy and productive system generally select for copiotrophic organisms while on the other side of the spectrum there is a higher tendency for oligotrophic survival strategies. However, soil systems are extremely complex which is why these factors do not completely explain the microbial community structure.

PCoA3

Along the third principal coordinate, we found that annual temperature, annual precipitation, and soil respiration explain much of the variation found in our community data. This model achieves an $R_a^2 = 0.92$. This is not a peak, but in the interest of parsimony, we chose to cut off the model at 3 covariates. From the regression results listed above, we see that the coefficients are significant, and the VIFs aren't very concerning. These parameters generally indicate the overall activity level of the microbial community: different groups of organisms have varying activity levels. Mean annual temperature and precipitation are both environmental characteristics that shape the type of ecosystem in the context of the first axis; however, here with the inclusion of soil respiration they indicate an explanation for the variation in soil microbial community directly influenced by energy and water dynamics; whereas, in the first model, these variables indicate overall ecosystem variation.

Conclusions

Soil microbial communities are extremely complex, hence it is challenging to discern ecological trends from samples like these. Moreover, it can be difficult to determine covariate relationship with high dimensional data. Here we showed that using a few axes in an ordination can help describe the relationships between soil microbial community and their environment.

We were pleased to discover that soil pH, which is thought to be the master variable in explaining the variation in our microbial community, was only related to one of the PCoA axes. Ecologist and soil scientists often oversimplify relationships between community and environmental characteristics, and our results show that such a simplification would fail to accurately categorize most of the variation in our community data. However, we feel that we did find reasonably simple models with ecologically tractable interpretations. More to the point, our models, taken in context, suggest different mechanisms at behind the variation in microbial communities at different sites in Oregon.

Naturally, more study will be necessary to develop a clearer picture of the determinants and complex relationship between soil attributes and the microbial communities ensconceed therein.

References

Caporaso, J.G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F.D., Costello, E.K., Fierer, N., Peña, A.G., Goodrich, J.K., Gordon, J.I., Huttley, G.A., Kelley, S.T., Knights, D., Koenig, J.E., Ley, R.E., Lozupone, C.A., McDonald, D., Muegge, B.D., Pirrung, M., Reeder, J., Sevinsky, J.R., Turnbaugh, P.J., Walters, W.A., Widmann, J., Yatsunenko, T., Zaneveld, J., Knight, R., 2010. QIIME allows analysis of high-throughput community sequencing data. Nat Meth 7, 335–336. doi:10.1038/nmeth.f.303 Fierer, Noah, and Robert B. Jackson. "The Diversity and Biogeography of Soil Bacterial Communities." Proceedings of the National Academy of Sciences of the United States of America 103, no. 3 (January 17, 2006): 626–31. doi:10.1073/pnas.0507535103. Fierer, Noah, Mya Breitbart, James Nulton, Peter Salamon, Catherine Lozupone, Ryan Jones, Michael Robeson, et al. "Metagenomic and Small-Subunit rRNA Analyses Reveal the Genetic Diversity of Bacteria, Archaea, Fungi, and Viruses in Soil." Applied and Environmental Microbiology 73, no. 21 (November 1, 2007): 7059–66. doi:10.1128/AEM.00358-07. Griffiths, Robert I., Bruce C. Thomson, PHillip James, Thomas Bell, Mark Bailey, and Andrew S. Whiteley. "The Bacterial Biogeography of British Soils." Environmental Microbiology 13, no. 6 (June 1, 2011): 1642–54. doi:10.1111/j.1462-2920.2011.02480.x. Hartmann, Martin, Beat Frey, Jochen Mayer, Paul Mäder, and Franco Widmer. "Distinct Soil Microbial Diversity under Long-Term Organic and Conventional Farming." The ISME Journal 9, no. 5 (May 2015): 1177–94. doi:10.1038/ismej.2014.210. Johnson, Richard Arnold., and Dean W. Wichern. Applied multivariate statistical analysis. 6th ed. Upper Saddle River: Prentice Hall, 2007. Print. Legendre, Pierre, and Louis Legendre. Numerical Ecology. Amsterdam: Elsevier, 2012. Print. Myrold, David D., Pamela A. Matson, and David L. Peterson. "Relationships between Soil Microbial Properties and Aboveground Stand Characteristics of Conifer Forests in Oregon." Biogeochemistry 8, no. 3 (1989): 265–81. O'Brien, Sarah L., Sean M. Gibbons, Sarah M. Owens, Jarrad Hampton-Marcell, Eric R. Johnston, Julie D. Jastrow, Jack A. Gilbert, Folker Meyer, and Dionysios A. Antonopoulos. “Spatial Scale Drives Patterns in Soil Bacterial Diversity.” Environmental Microbiology, March 1, 2016, n/a-n/a. doi:10.1111/1462-2920.13231. Rognes, T., Flouri, T., Nichols, B., Quince, C., Mahé, F., 2016. VSEARCH: a versatile open source tool for metagenomics. PeerJ 4, e2584. doi:10.7717/peerj.2584 Rousk, Johannes, Erland Bååth, Philip C. Brookes, Christian L. Lauber, Catherine Lozupone, J. Gregory Caporaso, Rob Knight, and Noah Fierer. “Soil Bacterial and Fungal Communities across a pH Gradient in an Arable Soil.” The ISME Journal 4, no. 10 (October 2010): 1340–51. doi:10.1038/ismej.2010.58. Running, S. W., D. L. Peterson, M. A. Spanner, and K. B. Teuber. “Remote Sensing of Coniferous Forest Leaf Area.” Ecology 67, no. 1 (1986): 273–76. doi:10.2307/1938532. Wang, Xiaobo, Joy D. Van Nostrand, Ye Deng, Xiaotao Lü, Chao Wang, Jizhong Zhou, and Xingguo Han. “Scale-Dependent Effects of Climate and Geographic Distance on Bacterial Diversity Patterns across Northern China’s Grasslands.” FEMS Microbiology Ecology 91, no. 12 (December 1, 2015): fiv133. doi:10.1093/femsec/fiv133.