Evaluation of Various Viticulture Practices on Soil Microbial Communities using Principal Component Analysis

Gabrielle Ilenstine

Statistics Department

California Polytechnic State University, San Luis Obispo

June 2018

Table of Contents

Introduction

In vineyard cropping, a nutrient-rich soil composition is desirable to achieve high production and quality of the fruit. There are various factors that contribute to the health of vineyard soil, but this study focuses primarily on the presence of microbial communities in the soil. The overall objective is to enrich the abundance of microbes in vineyard soil. This study evaluates the effect of eight different viticulture treatments on microbial abundance with the treatments split up into three experimental groups: herbicide, fertilizer, and cover crop. Microbial abundance is assessed at four seasonal dates during the year with two response variables: amount of DNA in the soil and biomass of each soil sample. We will be using principal component analysis on the collected data to explore two general research questions:

1. Which treatments have the most influence on abundance of microbes in the soil?
2. Which treatments are recommended for increasing microbial abundance, and thus improving soil health?

Experiment Design and Variables

The data for this experiment was collected before I was introduced to the study. In order to analyze the effects of herbicide use, fertilizer use, and cover crop cultivation, a 3.5-acre vineyard plot of land was divided into three separate experiments—one for each of the three different viticulture treatment categories. Table 1 shows the treatments within each of the experimental blocks.

Herbicide	Fertilizer	Cover Crop
• Herbicide • No Herbicide	• Organic Fertilizer • Synthetic Fertilizer • No Fertilizer	• High Water • Low Water • No Cover Crop

Table 1. Treatments categorized by experiment group

Each experiment contained three sample rows and six sampling blocks for each treatment. Treatments were assigned randomly to two sampling blocks in each row; the herbicide experiment contained four sample blocks per row while the fertilizer and cover crop experiments had six sample blocks per row. Data was collected seasonally at four significant viticulture dates during the year: bud break (mid-March), bloom (mid-May), veraison (mid-July), and harvest (late September). At each collection date, three soil core samples were obtained from each treatment block and then combined to form six total soil composites for each treatment to represent each sample block of land. In order to account for spatial differences between the treatment plots, these six composites were randomly combined to form a final three soil replicates for each of the eight treatments.

The main objective of this study is to identify which viticulture treatments contribute to an abundance of microbes in the soil. To measure this abundance of microbes, we analyzed two response variables. The first variable that was measured on each soil sample was the amount of DNA in the soil. Our data contains the amount of DNA that was recorded for each order (taxonomic rank) found in the soil; this allows us to see which orders are most prevalent in the microbial communities and how they vary across treatments. The amount of DNA in each order is a sum across the three soil samples. The second variable we analyzed was the biomass of each sample. Biomass is the total weight of double-stranded DNA per gram of dry soil, and is a representation of the overall abundance of microbes in each sample.

Principal Component Analysis

Principal component analysis is a statistical method used for dimensionality reduction in a data set when a large number of explanatory variables are measured. The procedure allows one to identify which variables are the most "important", or which variables contribute most to the total variability of the data (Rencher). In the context of our microbe data, through principal component analysis we can see which viticulture practices have the most influence on the variability of the amount of DNA and biomass in the soil. Principal components are linear combinations of explanatory variables that maximize the variance of the data, so we will focus on those principal components that explain most of the total variability.

Mathematical Procedure

For the data with **N** observations and **P** explanatory variables, in order to carry out the principal component analysis we start by organizing the data into an **N x P** matrix **X**.

$$X = \begin{bmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{np} \end{bmatrix}_{N\ x\ P}$$

Then, we calculate the variance-covariance matrix **C** of matrix **X**.

$$C = \frac{1}{N-1}(X - \bar{X}')'(X - \bar{X}')$$

This results in matrix **C** having dimensions of **P x P**. Notice how the dimensions of this variance-covariance matrix **C** are only dependent on the number of explanatory variables **P** and not the number of observations **N** from the original data set **X** (Quekovich).

$$C = \begin{bmatrix} c_{11} & \cdots & c_{1p} \\ \vdots & \ddots & \vdots \\ c_{p1} & \cdots & c_{pp} \end{bmatrix}_{P\ x\ P}$$

With this variance-covariance matrix **C,** we can calculate the eigenvalues and their corresponding eigenvectors that are necessary for our principal component analysis.

$$\text{Eigenvalues:} \quad \lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p$$

$$\text{Eigenvectors:} \quad \boldsymbol{e}_1, \boldsymbol{e}_2, \dots, \boldsymbol{e}_p$$

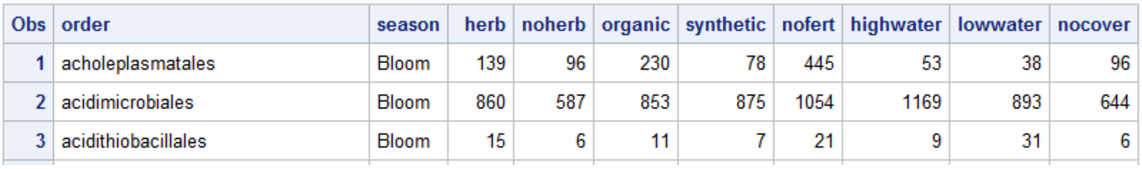
The total variance of the sample data equals the sum of the eigenvalues. This means that the largest eigenvalue will contribute the most to the total variability of the data, so we order the eigenvalues from largest to smallest to identify which eigenvectors to use to create our principal components.

$$\mathbf{var}(Y_i) = \mathrm{var}(e_{i1}X_1 + e_{i2}X_2 + \ldots e_{ip}X_p) = \lambda_i$$

This equation is the basis for our principal components. Each eigenvalue **λ_i** represents the variance of principal component **Y_i**. Eigenvector **i** will serve as the coefficients for the linear combination of the **P** x-variables. Because this entire process is very computational-heavy, it's standard to let software handle the calculations.

Using SAS

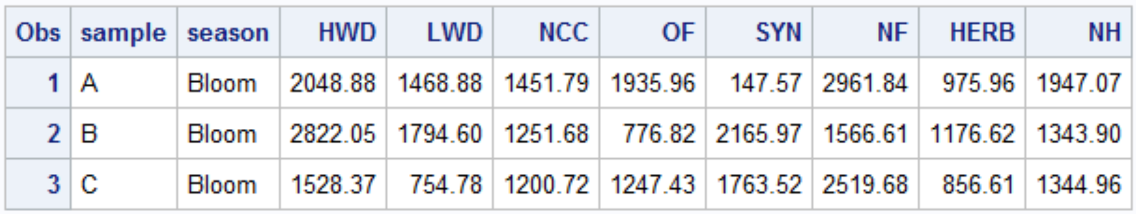
Table 2A displays the first three observations for the DNA data from the bloom season in a SAS table. Aside from the first two columns specifying order and season, the data is organized into an **N** x **8** matrix with the orders for the **N** rows (**N** varied for each season, but it hovered around 112) and the treatments for the **8** columns.



Obs	order	season	herb	noherb	organic	synthetic	nofert	highwater	lowwater	nocover
1	acholeplasmatales	Bloom	139	96	230	78	445	53	38	96
2	acidimicrobiales	Bloom	860	587	853	875	1054	1169	893	644
3	acidithiobacillales	Bloom	15	6	11	7	21	9	31	6

Table 2A. First three observations of DNA data set

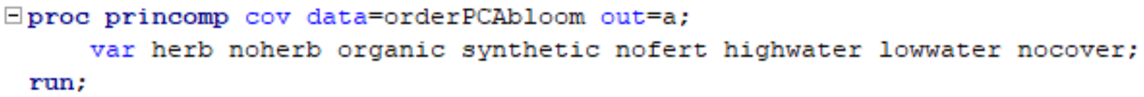
The biomass data was organized in the same format as the DNA data with only two differences: the biomass data has only three observations (one for each of the three soil samples) and the treatments in each column are named differently. The SAS output is presented in Table 2B.



Obs	sample	season	HWD	LWD	NCC	OF	SYN	NF	HERB	NH
1	A	Bloom	2048.88	1468.88	1451.79	1935.96	147.57	2961.84	975.96	1947.07
2	B	Bloom	2822.05	1794.60	1251.68	776.82	2165.97	1566.61	1176.62	1343.90
3	C	Bloom	1528.37	754.78	1200.72	1247.43	1763.52	2519.68	856.61	1344.96

Table 2B. Biomass data set

The principal component analysis can be done with the simple PROC PRINCOMP procedure in SAS. The code shown below will calculate the variance-covariance matrix and output eigenvalues, eigenvectors, variance-explained and Skree plot for the DNA data in the bloom season.

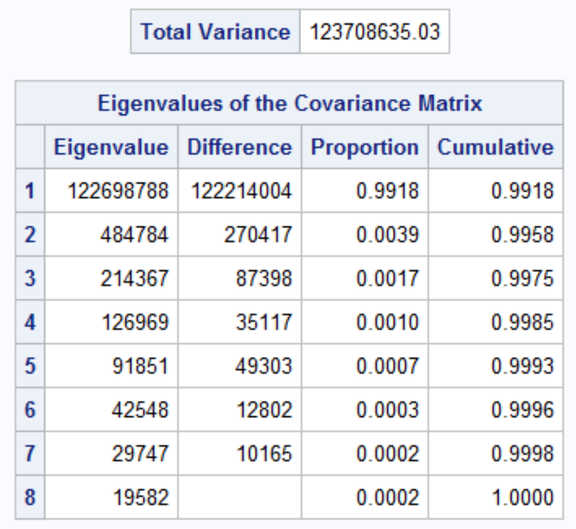


```
⊟proc princomp cov data=orderPCAbloom out=a;
     var herb noherb organic synthetic nofert highwater lowwater nocover;
 run;
```

Results for DNA Data

We begin by running a principal component analysis on the DNA data. In order to identify how many principal components to analyze, we start by seeing which eigenvalues contribute most of the total sample variability. We will also take a look at the correlation plots between principal components and make note of any extreme observations that are contributing to a lot of the variability.

Eigenvalues



Total Variance	123708635.03

Eigenvalues of the Covariance Matrix				
	Eigenvalue	Difference	Proportion	Cumulative
1	122698788	122214004	0.9918	0.9918
2	484784	270417	0.0039	0.9958
3	214367	87398	0.0017	0.9975
4	126969	35117	0.0010	0.9985
5	91851	49303	0.0007	0.9993
6	42548	12802	0.0003	0.9996
7	29747	10165	0.0002	0.9998
8	19582		0.0002	1.0000

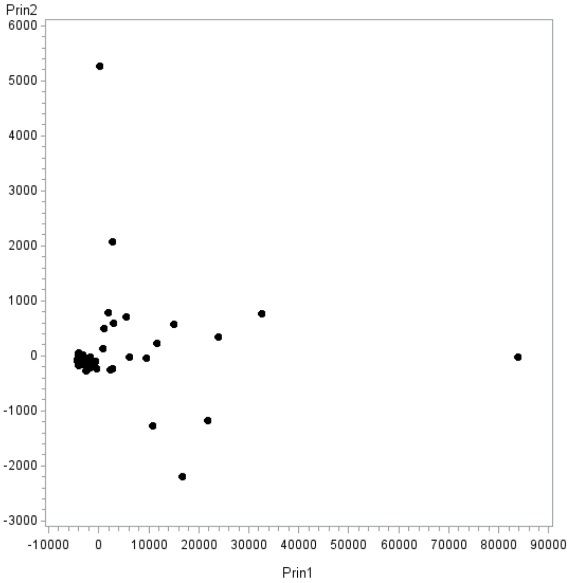
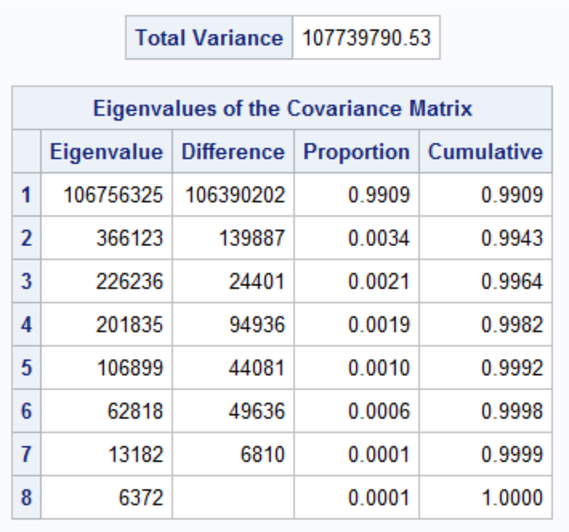


Table 3A. Eigenvalues (bud break)

Figure 1A. Correlation plot (bud break)



Total Variance	107739790.53

Eigenvalues of the Covariance Matrix				
	Eigenvalue	Difference	Proportion	Cumulative
1	106756325	106390202	0.9909	0.9909
2	366123	139887	0.0034	0.9943
3	226236	24401	0.0021	0.9964
4	201835	94936	0.0019	0.9982
5	106899	44081	0.0010	0.9992
6	62818	49636	0.0006	0.9998
7	13182	6810	0.0001	0.9999
8	6372		0.0001	1.0000

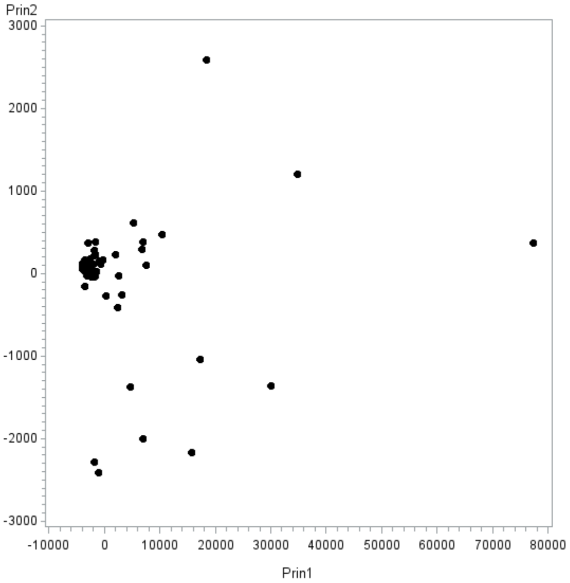
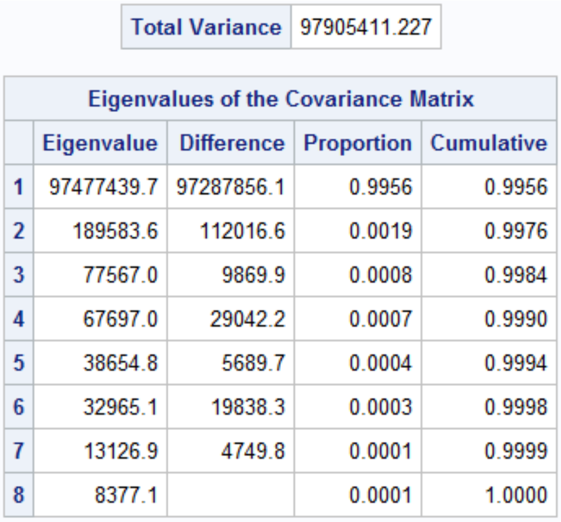


Table 3B. Eigenvalues (bloom)

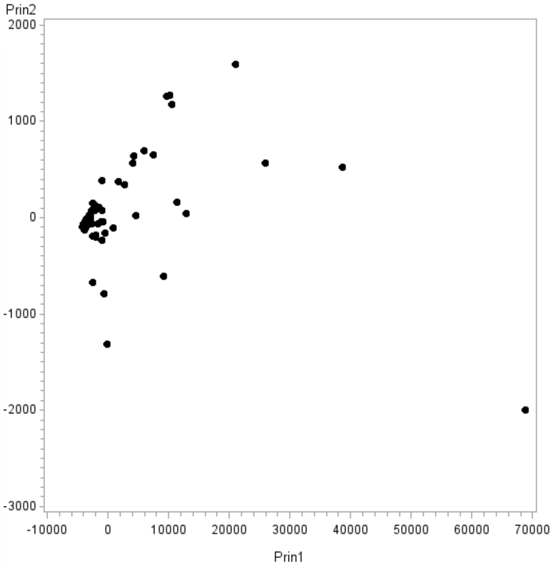
Figure 1B. Correlation plot (bloom)



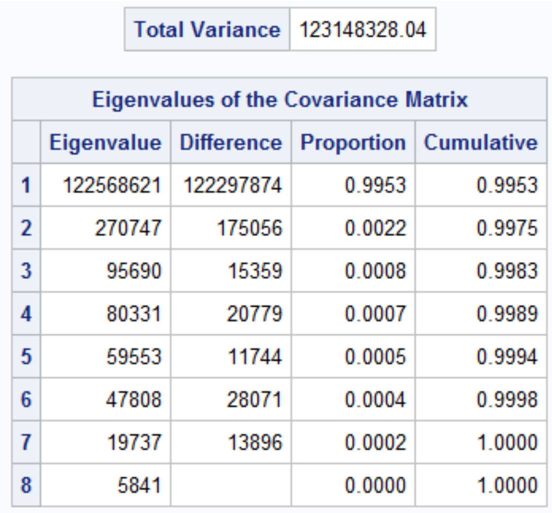
Total Variance	97905411.227

Eigenvalues of the Covariance Matrix				
	Eigenvalue	Difference	Proportion	Cumulative
1	97477439.7	97287856.1	0.9956	0.9956
2	189583.6	112016.6	0.0019	0.9976
3	77567.0	9869.9	0.0008	0.9984
4	67697.0	29042.2	0.0007	0.9990
5	38654.8	5689.7	0.0004	0.9994
6	32965.1	19838.3	0.0003	0.9998
7	13126.9	4749.8	0.0001	0.9999
8	8377.1		0.0001	1.0000

Table 3C. Eigenvalues (veraison)



1C. Correlation plot (veraison)



Total Variance	123148328.04

Eigenvalues of the Covariance Matrix				
	Eigenvalue	Difference	Proportion	Cumulative
1	122568621	122297874	0.9953	0.9953
2	270747	175056	0.0022	0.9975
3	95690	15359	0.0008	0.9983
4	80331	20779	0.0007	0.9989
5	59553	11744	0.0005	0.9994
6	47808	28071	0.0004	0.9998
7	19737	13896	0.0002	1.0000
8	5841		0.0000	1.0000

Table 3D. Eigenvalues (harvest)

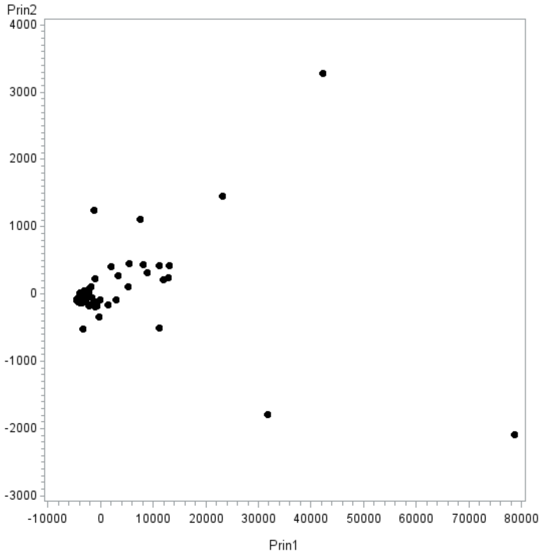


Figure 1D. Correlation plot (harvest)

For every season, the first eigenvalue (displayed in Tables 3A-3D) represents over .99 of the total variability of the data. This means that the variance of the first principal component is extremely close to the actual sample variance. In this case, we don't need to consider the second principal component because the first will provide us with enough information for analysis. The correlation plots shown in Figures 1A-1D graph the first and second principal components of every data point. We can see that every season shows a certain extreme observation out to the very far right of the graph, at the end of the x-axis. These particular points represent the actinomycetales order. For a biochemist, it may be useful for them to know that this is a very abundant microbe in the soil. We will now delve further into the analysis by taking a look at the first principal component for each season.

First Principal Component

Now we are able to analyze the eight treatments by breaking down the first principal components for each season. The first principal component is the linear combination of x-values that has the maximum variance (Lesson 11); we use our eigenvectors as the coefficients (in the following equations, we are only going to be looking at the first principal component $\mathbf{Y}_i$).

$$\begin{aligned}
\hat{Y}_1 &= \hat{e}_{11}X_1 + \hat{e}_{12}X_2 + \cdots + \hat{e}_{1p}X_p \\
\hat{Y}_2 &= \hat{e}_{21}X_1 + \hat{e}_{22}X_2 + \cdots + \hat{e}_{2p}X_p \\
&\vdots \\
\hat{Y}_p &= \hat{e}_{p1}X_1 + \hat{e}_{p2}X_2 + \cdots + \hat{e}_{pp}X_p
\end{aligned}$$

	Bud Break	**Bloom**	**Veraison**	**Harvest**
Herbicide	0.3633	0.3194	**0.3923**	0.3110
No Herbicide	0.3658	0.2761	0.3586	**0.4289**
Organic Fertilizer	**0.4173**	0.3693	0.2972	0.3098
Synthetic Fertilizer	0.3393	0.3338	0.3608	0.3365
No Fertilizer	0.3138	**0.3912**	0.3509	**0.4259**
High Water	0.3330	**0.4351**	0.3759	0.3211
Low Water	**0.3917**	0.3386	0.3383	0.3408
No Cover Crop	0.2866	0.3419	0.3465	0.3310

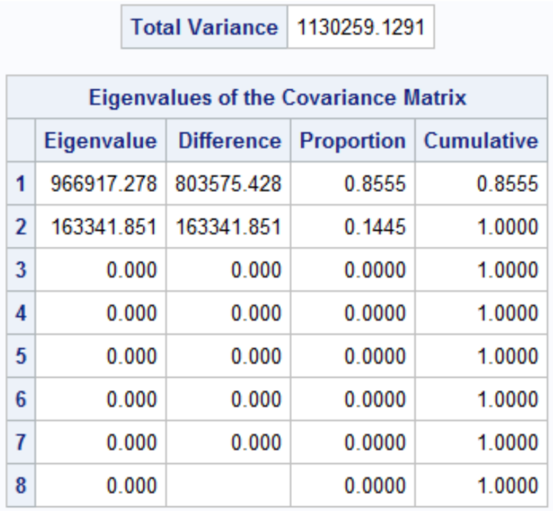
Table 4. DNA data first principal component coefficients for each season

The greater in magnitude the coefficient, the more that variable contributes to the total variance. Since one of our objectives here is to identify which treatments have the most influence on abundance of microbes in the soil, we are looking for the treatments with large corresponding coefficients. The treatments that contribute the most to the total sample variability for each season are bolded in Table 4. These are the recommended treatments for each season (based off of DNA data). It's interesting to note that the treatments that contribute to the largest microbial abundance are not always the same for each season. This makes sense if one thinks about it—different treatments are more appropriate, and therefore more effective, depending on the season.

Results for Biomass Data

The biomass data shows some differences from the DNA data. We are still going to identify which eigenvalues contribute to most of the total variability, but we won't be looking at the correlation plots for this data. The sample size is only three here (one observation for each of the three samples) as opposed to the nearly 120 various different orders in the other data set, so a correlation plot with only three points would be unable to display any extreme observations. However, the Skree and Variance Explained plots are included here as graphical representations of the eigenvalues to demonstrate more of what SAS outputs.

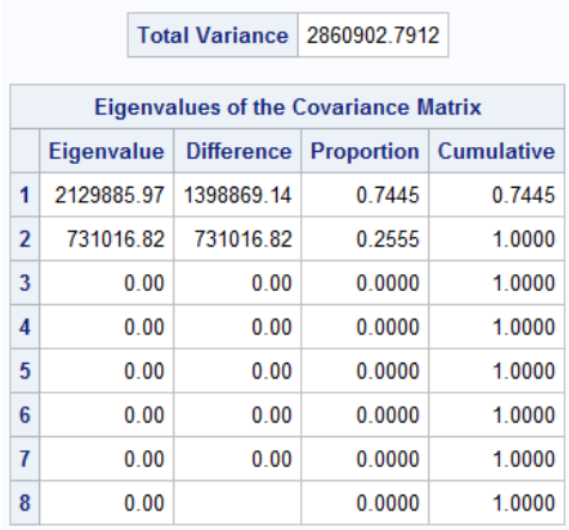
Eigenvalues



Total Variance	1130259.1291

Eigenvalues of the Covariance Matrix				
	Eigenvalue	Difference	Proportion	Cumulative
1	966917.278	803575.428	0.8555	0.8555
2	163341.851	163341.851	0.1445	1.0000
3	0.000	0.000	0.0000	1.0000
4	0.000	0.000	0.0000	1.0000
5	0.000	0.000	0.0000	1.0000
6	0.000	0.000	0.0000	1.0000
7	0.000	0.000	0.0000	1.0000
8	0.000		0.0000	1.0000

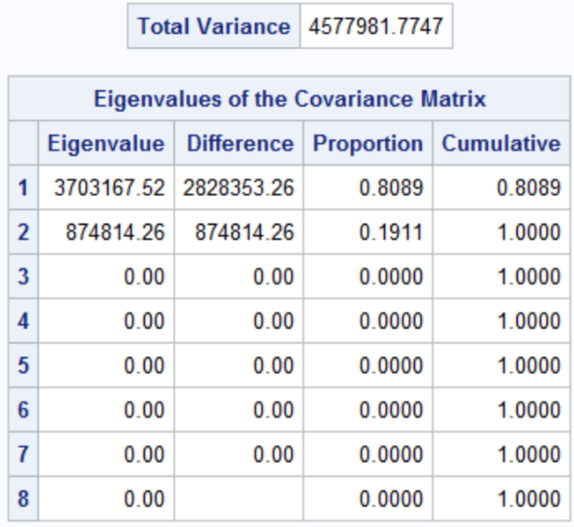
Table 5A. Eigenvalues (bud break)



Total Variance	2860902.7912

Eigenvalues of the Covariance Matrix				
	Eigenvalue	Difference	Proportion	Cumulative
1	2129885.97	1398869.14	0.7445	0.7445
2	731016.82	731016.82	0.2555	1.0000
3	0.00	0.00	0.0000	1.0000
4	0.00	0.00	0.0000	1.0000
5	0.00	0.00	0.0000	1.0000
6	0.00	0.00	0.0000	1.0000
7	0.00	0.00	0.0000	1.0000
8	0.00		0.0000	1.0000

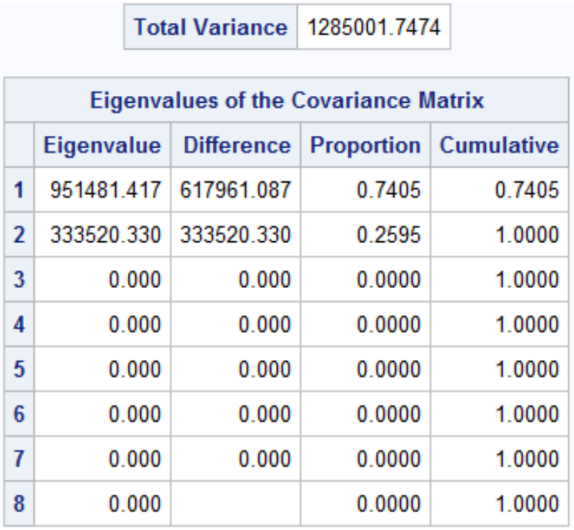
Table 5B. Eigenvalues (bloom)



Total Variance	4577981.7747

Eigenvalues of the Covariance Matrix				
	Eigenvalue	Difference	Proportion	Cumulative
1	3703167.52	2828353.26	0.8089	0.8089
2	874814.26	874814.26	0.1911	1.0000
3	0.00	0.00	0.0000	1.0000
4	0.00	0.00	0.0000	1.0000
5	0.00	0.00	0.0000	1.0000
6	0.00	0.00	0.0000	1.0000
7	0.00	0.00	0.0000	1.0000
8	0.00		0.0000	1.0000

Table 5C. Eigenvalues (veraison)



Total Variance	1285001.7474

Eigenvalues of the Covariance Matrix				
	Eigenvalue	Difference	Proportion	Cumulative
1	951481.417	617961.087	0.7405	0.7405
2	333520.330	333520.330	0.2595	1.0000
3	0.000	0.000	0.0000	1.0000
4	0.000	0.000	0.0000	1.0000
5	0.000	0.000	0.0000	1.0000
6	0.000	0.000	0.0000	1.0000
7	0.000	0.000	0.0000	1.0000
8	0.000		0.0000	1.0000

Table 5D. Eigenvalues (harvest)

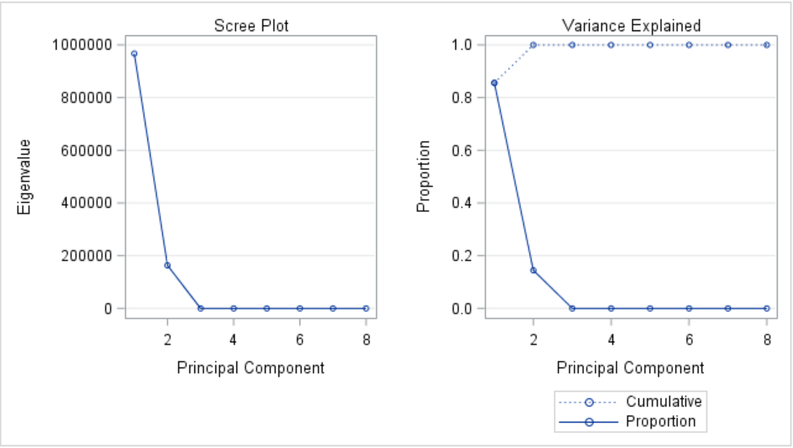


Figure 2A. Scree and Variance Explained plots (bud break)

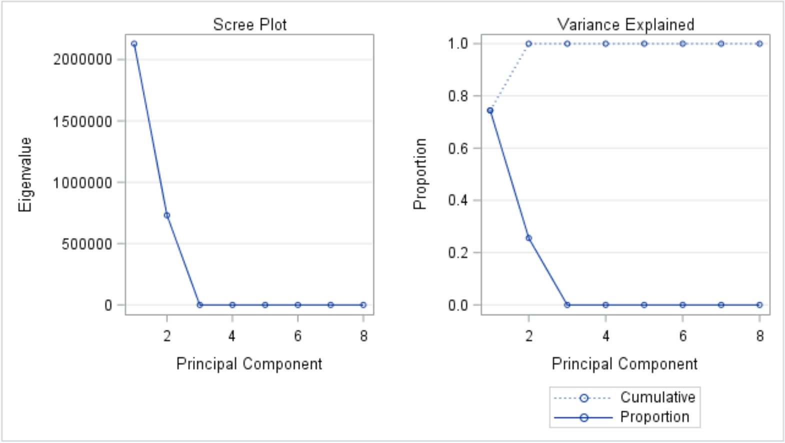


Figure 2B. Scree and Variance Explained plots (bloom)

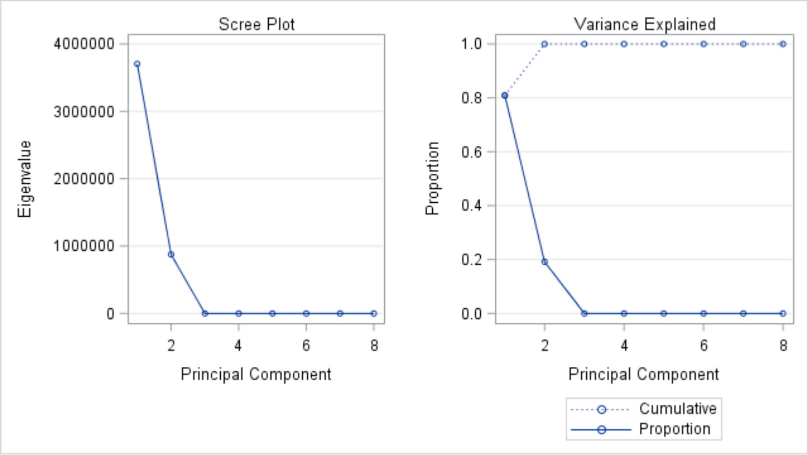


Figure 2C. Scree and Variance Explained plots (veraison)

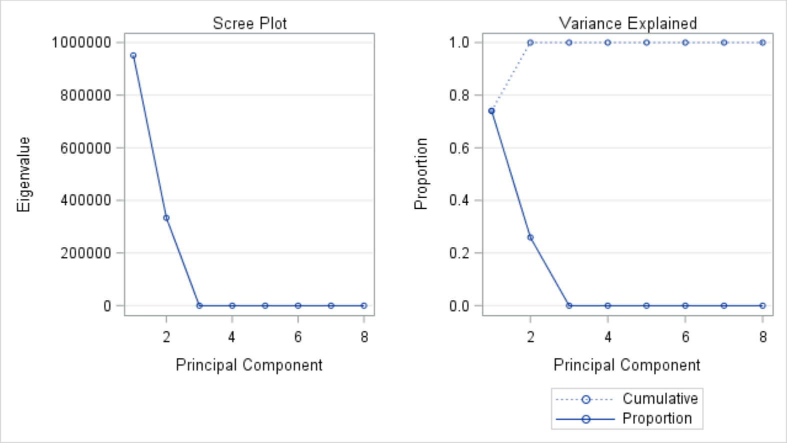


Figure 2D. Scree and Variance Explained plots (harvest)

What is important to notice here is that the first principal component for the biomass data only explains at most .85 of the total variability (Table 5A) and as little as .74 of the total variability (Table 5D). This means that while we will put the strongest emphasis on our first principal component analysis, we will also look to the second principal component to explain the remaining variability. The scree and variance plots in

Figures 2A-2D display the sharp drop in variance explained from the first to the second principal components.

First Principal Component

	Bud Break	Bloom	Veraison	Harvest
Herbicide	**0.8806**	-0.0656	-0.2702	0.1012
No Herbicide	-0.3667	0.2108	-0.0094	-0.0495
Organic Fertilizer	0.1125	0.3984	0.0233	**-0.4493**
Synthetic Fertilizer	-0.0522	**-0.6997**	0.0747	0.0327
No Fertilizer	-0.0126	**0.4740**	0.3002	-0.0817
High Water	-0.2721	-0.2518	**0.9033**	-0.1837
Low Water	-0.0203	-0.0990	-0.0765	**0.7989**
No Cover Crop	0.0107	0.0707	-0.0939	0.3252

Table 6A. Biomass data first principal component coefficients for each season

Just like with the DNA data, the coefficients with the largest magnitude have the most influence of the variability of the data. Our results here are quite different because we have a few negative coefficients. This can be interpreted as having a negative relationship between the treatment and abundance of microbes. Thus, if we're looking to increase the abundance of microbes in the soil, we want to choose the treatments that have large *positive* coefficients and avoid the treatments that have large *negative* coefficients.

Second Principal Component

	Bud Break	**Bloom**	**Veraison**	**Harvest**
Herbicide	0.0078	0.1525	**0.5307**	-0.0945
No Herbicide	-0.1181	0.1901	**0.7458**	**0.6806**
Organic Fertilizer	**0.6248**	0.0490	0.1148	0.0732
Synthetic Fertilizer	**0.6713**	-0.3668	0.2953	-0.0593
No Fertilizer	-0.0910	-0.2019	0.1589	-0.4125
High Water	0.3102	**0.6285**	0.1043	0.0949
Low Water	0.1641	**0.5987**	0.0167	0.2845
No Cover Crop	0.1165	0.0977	0.1592	**-0.5088**

Table 6B. Biomass data second principal component coefficients for each season

By definition, the second principal component accounts for as much of the remaining variance as possible after the first principal component, with the constraint that the correlation between the first and second component is 0. In this case, we can still interpret the coefficients the same as with the first principal component, but we will not give them as much weight in our final conclusions of the best viticulture treatments. Table 6B displays the second principal component coefficients for each season, with the values greatest in magnitude bolded.

Extension

Principal component analysis is limiting in that the method does not account for time variables; since season is a time variable, we are only able to analyze the effects of differing vineyard practices within each season individually. In order to see how microbes in the soil change over time, we could use an alternative statistical method known as multiple factor analysis (MFA). MFA "is an extension of principal component analysis tailored to handle multiple data tables that measure sets of variables collected on the same observations, or, alternatively, multiple data tables where the same variables are measured on different sets of observations" (Abdi). The procedure can be thought of as a multi-table PCA, and thus would be appropriate for our consistently formatted, seasonal data.

The multiple factor analysis procedure can be loosely explained in a few steps. In the context of our microbe data, a principal component analysis is performed for each season individually, and then the resulting eigenvectors are divided by their respective eigenvalues. This "normalizes" each season and gives greater weight to the seasons that contribute more variability. These “normalized” data tables for each season are then combined and another PCA is run on the combined data set to analyze the original explanatory variables (herbicide, fertilizer, and cover crop treatments). Because multiple factor analysis accounts for discrepancies and commonalities between seasons, this procedure would be an appropriate extension to principal component analysis.

Conclusion

Principal component analysis provides sufficient analysis of which viticulture treatments contribute the most to a variety of microbial communities in vineyard soil. Table 7 provides a summary of the treatments in each season that were found to be the most influential on the abundance of microbes.

	Bud Break	**Bloom**	**Veraison**	**Harvest**
Amount of DNA	Organic Fertilizer	High Water	Herbicide	No Herbicide No Fertilizer
Biomass	Herbicide	No Fertilizer	High Water	Low Water

Table 7. Summary of recommended treatments

It’s interesting to note that the recommended treatments are not the same for the DNA data and for the biomass data. This could be because the biomass sample size was so small, or it could be because the DNA data was grouped by order. Either way, we should consider all recommended treatments. Principal component analysis serves as a great jumping-off point for further exploration, and the potential benefits for vineyard cropping are motivation to continue with this study.

Works Cited

Abdi, Herve, et al. *Multiple factor analysis: principal component Analysis for multitable and multiblock data sets.* Wiley Periodicals, Inc., 2013.

"Lesson 11: Principal Components Analysis." *STAT 505: Applied Multivariate Statistical Analysis,* Pennsylvania State University, 2018, onlinecourses.science.psu.edu/stat505/node/51/.

Quekovich. "PCA: example – Steps 1 & 2." *Youtube,* 5 Jan. 2017. youtube.com/watch?v=Ao_iYZ50RNY.

Rencher, Alvin C. *Methods of Multivariate Analysis.* J. Wiley & Sons, 2002.