STAT 503X Case Study 3: Baker Field Data

1 Description

Since the time that the Global Positioning System (GPS) became available for public use nearly ten years ago, there has been an increased interest in precision farming. The Global Positioning System allows the agronomist or soil scientist to ascertain her location within a field to a high level of accuracy. This precise measure of location in a field, combined with soil characteristics, yield measurements and other data associated with the location can then be used to extract valuable information about the complex process of plant growth. This information collected at multiple locations within a field can be used to formulate yield-maximization strategies such as variable-rate fertilizer, herbicide or pesticide application.

The data to be analyzed in this study was drawn from part of a privately owned farm in southeastern Boone County, Iowa Colvin, Jaynes, Karlen, Laird & Ambuel (1997). A total of 224 sites within an approximately 350 m × 350 m portion of the field were studied. The sites are located along 8 equally-spaced east-west transects. Along each transect there are 28 sites, spaced approximately 12.2 m apart. At harvest time, a combine is driven down each transect, stopping every 12.2 m to measure yield in bushels/acre (Colvin, et. al., 1997).

The field of interest has been on a corn-soybean rotation since 1957. The 1997 data set to be discussed contains:

- x: location in the field
- y: location in the field
- Corn97BY: corn yields in bushels
- B: Boron (parts per million)
- Ca: Calcium (parts per million)
- Cu: Copper (parts per million)
- Fe: Iron (parts per million)
- K: Potassium (parts per million)
- Mg: Magnesium (parts per million)
- Mn: Manganese (parts per million)
- Na: Sodium (parts per million)
- P: Phosphorus (parts per million)
- Zn: Zinc (parts per million)

The corn was harvested on October 6, 1997 and the soil samples were taken on May 22-23, 1997. The soil samples consisted of 6 cores, 1 inch in diameter and collected to a depth of 8 inches. Samples were air dried, ground, extracted, and analyzed for the various nutrients using inductively-coupled plasma (ICP) techniques (Karlen and Colvin, 1998). Because soil nutrient data were not available at 9 of the sites, only 215 of the sites are included in the
analyses to follow.

The primary question is “What is the relationship between Corn Yield and the Soil Nutrients?”

2 Suggested Approaches

<table>
<thead>
<tr>
<th>Approach</th>
<th>Reason</th>
<th>Type of questions addressed</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data Restructuring</td>
<td>It may be necessary to use transformations because it would help to be able to model the data with a linear model.</td>
<td></td>
</tr>
<tr>
<td>Visual Inspection</td>
<td>Field Plot</td>
<td>“What is the layout of samples over the field?”</td>
</tr>
<tr>
<td></td>
<td>Plots of Yield vs Soil Nutrients</td>
<td>“Which variables are likely to best predict yield?”</td>
</tr>
<tr>
<td></td>
<td>Pairwise Plots of Soil Nutrients</td>
<td>“How do the presence of soil nutrients relate to each other?”</td>
</tr>
<tr>
<td></td>
<td>Tour Plots of Soil Nutrients</td>
<td>“How do the presence of soil nutrients relate to each other?”</td>
</tr>
<tr>
<td></td>
<td>Link Field Plot to Other Plots</td>
<td>“Is there a spatial pattern to the yield or soil nutrient concentrations?”</td>
</tr>
<tr>
<td></td>
<td></td>
<td>“How closely related are neighboring points in value?”</td>
</tr>
<tr>
<td></td>
<td></td>
<td>“Are there any points which have substantially different values to their neighbors?”</td>
</tr>
<tr>
<td>Numerical Analysis</td>
<td>Linear Regression Analysis</td>
<td>“Which variables best predict yield?”</td>
</tr>
<tr>
<td></td>
<td>Principal Component Analysis</td>
<td>“Are there dependencies amongst the soil nutrient concentrations?”</td>
</tr>
</tbody>
</table>
3 Actual Approaches

3.1 Data Structuring

Data Setup: Several files were constructed to be passed into XGobi:

- `baker.dat` the data matrix, each row contains measurements on one sample
- `baker.col` the variable labels, one label per row, so there are as many rows as there are variables

3.2 Visual Inspection

![Map of the yield/soil characteristic measurement locations in the field.](image)

Figure 1: Map of the yield/soil characteristic measurement locations in the field.

The field map (Figure 1) shows that the measurements were extremely regular, and there were a few missing observations.

The first step in our analysis was to look at the relationship or correlation between each of the soil nutrients and corn yield. This was done using XGobi software for interactive high-dimensional data analysis (see Swayne, Cook & Buja (1998)).

XGobi allows the user to identify points in a plot by selecting the “Identify” option and then clicking on the points of interest. This is a valuable tool for identifying outliers. Point
number 200 was identified as an outlier for this data set because it lies far away from the bulk of the points in several of the scatterplots that were examined.

Also, XGobi allows the user to make transformations on the fly. In the data we noticed many non-linear patterns, so we explored the use of transformations to linearize the variable relationships. This makes future analysis easier. All correlation coefficients discussed in this section are calculated in the absence of point number 200. When examining each soil nutrient, a transformation of the nutrient was often chosen in order to linearize its relationship with corn yield. A discussion of each nutrient’s relationship with yield follows.

**Boron.** The scatterplot of Boron versus corn yield is given in Figure 2. The correlation after removing the outlier (point number 200) is 0.137. There did not appear to be a transformation which could linearize this relationship, but it is interesting to note that higher amounts of Boron lead to consistently higher yields.

![Figure 2: Scatterplot of Boron versus corn yield. Levels of high, medium and low yield are brushed in green, navy and red, respectively.](image)

**Calcium.** Figure 3 shows scatterplots of Calcium versus corn yield and 1/Calcium versus corn yield. Note how the outliers (point numbers 81 and 82) become part of the main body of the data after applying the inverse transformation. Additionally, the relationship between 1/Calcium and yield is more linear than the relationship between Calcium and yield. This graphical observation is supported by the fact that the correlation improves from 0.329 to -0.457 when the Calcium values are inverted. (Without the points numbered 81 and 82, the correlation between the untransformed Calcium values and yield was 0.387.)
Figure 3: Scatterplots of Calcium versus corn yield and 1/Ca versus corn yield. Levels of high, medium and low yield are brushed in green, navy and red, respectively.

**Copper.** Figure 4 shows scatterplots of Copper versus corn yield and log(Copper) versus corn yield. Taking the log of Copper makes its relationship with corn yield more linear, as illustrated by the increase in correlation from 0.625 to 0.654.

**Iron.** The left plot in Figure 5 shows the highly nonlinear relationship between Iron and corn yield with a linear correlation coefficient of .134. This nonlinearity is accentuated some by taking the log of Iron, as shown in the second plot of Figure 5.

However, there exists an interesting relationship between Iron, Calcium and yield which can be visualized using a technique in XGobi called “brushing.” One can open two separate (but “linked”) XGobi windows and brush or highlight points in one window while viewing the associated points in the second window. Figure 6 shows a dotplot of Calcium on the left, with high values of Calcium (values in excess of 5200 ppm) brushed yellow-green and low values brushed pink. The plot on the right is the plot of log(Iron) versus yield which is linked to the dotplot to its left.

Hence, it appears that the “7”-shape of this plot is due to two separate phenomena. That is, the relationship of iron with yield is different for different values of Calcium. When Calcium is high (with values in excess of 5200 ppm), an increase in Iron results in little practical increase in yield. The correlation between log(Iron) and yield for the high-Calcium group is 0.309. However, when Calcium values are low (with values in less than 5200 ppm), an increase in Iron results in a substantial increase in yield. The correlation between log(Iron)
Figure 4: Scatterplots of Copper versus corn yield and log(Copper) versus corn yield. Levels of high, medium and low yield are brushed in green, navy and red, respectively.

Figure 5: Scatterplots of Iron versus corn yield and log(Iron) versus corn yield. Levels of high, medium and low yield are brushed in green, navy and red, respectively.
Figure 6: The plot on the left is a dotplot of Calcium with high values of Calcium (values in excess of 5200 ppm) brushed yellow-green and low values brushed pink. The plot on the right is the plot of log(Iron) versus yield which is linked to the dotplot to its left. The two high values of Calcium (point numbers 81 and 82) are labeled in both plots.

and yield for the low-Calcium group is 0.535.

**Potassium.** The plot on the left side of Figure 7 illustrates a nonlinear relationship between Potassium and yield. The plot on the right side of Figure 7 shows how inverting the Potassium values results in a fairly strong linear relationship between 1/Potassium and yield. The transformation improves the correlation coefficient from a value of 0.467 to -0.544.

**Magnesium.** The plot of Magnesium versus yield in Figure 8 is similar to the relationship between Potassium and yield in Figure 7. Applying the same transformation (1/x) yields a linear relationship between 1/Magnesium and yield. The transformation improves the correlation coefficient from a value of 0.491 to -0.528.

**Manganese.** Figure 9 illustrates that the relationship between the log of Manganese and yield is slightly more linear than the relationship between Manganese and yield. The transformation increases the correlation coefficient slightly, from a value of 0.456 to a value of 0.475.
Figure 7: Scatterplots of Potassium versus corn yield and 1/Potassium versus corn yield. Levels of high, medium and low yield are brushed in green, navy and red, respectively.

Figure 8: Scatterplots of Magnesium versus corn yield and 1/Magnesium versus corn yield. Levels of high, medium and low yield are brushed in green, navy and red, respectively.
Figure 9: Scatterplots of Manganese versus corn yield and log(Manganese) versus corn yield. Levels of high, medium and low yield are brushed in green, navy and red, respectively.

**Sodium.** Note from Figure 10 that the discrete Sodium measurements did not have a strong relationship with yield. Neither the plot nor the low correlation coefficient of .183 could be improved by a transformation, but it is interesting to note that higher amounts of Sodium lead to consistently higher yields.

**Phosphorus.** The nonlinear shape of the plot of Phosphorus versus yield in Figure 11 was changed to a more linear shape by taking the log of the Phosphorus values. This transform increases the correlation coefficient from 0.490 to 0.537.

**Zinc.** Because the shape of the plot of Zinc versus yield in Figure 12 is similar to the shape of the plots of Copper, Manganese and Phosphorus, a natural tendency would be to use the same (log) transform. But since some of the Zinc values are equal to zero, the transform log(Zinc + 1) was used. This increases the value of the correlation coefficient from 0.615 to 0.645.

**Relating Soil Variables to Each Other**

A convenient way to view the pairwise relationships between variables is using a matrix of scatterplots of all pairs of variables. This is shown in Figures 13 and 14. The pairwise plots are laid out in “matrix” format. For example, all the plots across the top row have Boron plotted vertically and all other variables are sequentially plotted horizontally. Down the first
Figure 10: Scatterplot of Sodium versus corn yield. Levels of high, medium and low yield are brushed in green, navy and red, respectively.

Figure 11: Scatterplots of Phosphorus versus corn yield and log(Phosphorus) versus corn yield. Levels of high, medium and low yield are brushed in green, navy and red, respectively.
Figure 12: Scatterplots of Zinc versus corn yield and log(Zinc+1) versus corn yield. Levels of high, medium and low yield are brushed in green, navy and red, respectively.

column, Boron is plotted horizontally, with the other variables plotted vertically. The second row shows Calcium plotted vertically and the other variables plotted horizontally.

In the plots of the raw variables it can be seen that almost all of the pairs of variables have a linear positive relationship. That is, the more of one nutrient the more of the other nutrient. There are a few noticeable exceptions: Calcium and Iron, and Copper and Iron. The plot of Copper vs Iron indicates a non-linear relationship, and the plot of Calcium vs Iron indicates a negative relationship.

The plots of the transformed variables are much clearer to read, but intuitively harder to interpret. For these plots it is clear that some of the variables are very strongly related to each other: Calcium, Copper, Magnesium, Manganese and Zinc, and, Potassium and Phosphorus, and Boron and Sodium. Note also, that the non-linear relationship between Copper and Iron is still visible, and the relationship between Calcium and Copper now looks somewhat non-linear. There are also a few noticeably unusual points in several of the plots.

A Grand Tour of the Soil Variables

The Grand Tour is a data analysis tool that can be helpful for identifying outliers, clusters and general structure in a data set. The XGobi software package describes the Grand Tour as a graphical method which “successively samples planes in p-space, where p is the number of variables presently selected, and connects the planes by moving along a geodesic interpolation path between them...By allowing the grand tour to run uninterrupted the viewer can get a global view of the linear combinations among the variables” (XGobi online help, see also
Figure 13: Matrix of pairwise scatterplots of all soil variables. Most variables have a positive relationship, with the exception of Calcium and Iron, and Copper and Iron.
Figure 14: Matrix of pairwise scatterplots of all transformed soil variables. Now it is clear that Calcium, like Copper, has some non-linear relationship with Iron. Also note that Calcium, Copper, Magnesium, and Zinc are strongly related. Similarly Potassium and Phosphorus are strongly related, and Boron and Sodium are strongly related.
Swayne, et. al., 1998).

As the Grand Tour progresses dynamically, one can notice observations that do not “move” with the rest of the data through the projection space. These observations might be considered to be outliers. Figure 15 shows three different views of the Grand Tour of the untransformed soil variables. In viewing this Grand Tour, one can readily notice that the points labeled 81 and 82 should be considered outliers when considering all ten untransformed soil variables jointly. Points 81 and 82 have the two highest values for Calcium. There are also a few more less extreme outliers but these also look to be mostly in the measurement of Calcium. Doing analysis on the inverse of Calcium makes the outlying nature of these observations less pronounced.

![Figure 15: Three different views of the Grand Tour of the untransformed soil variables.](image)

A cluster in high-dimensional data is a group of points that has “similar” values for all of the variables. We generally think of clustered data as having two or more such clusters, and generally there is a distinct separation between clusters. For detecting clusters we would look for groups of points which “move” together, but differently from points in other clusters in the Grand Tour. With this soil data there are no obvious clusters. Rather the untransformed data takes an interesting non-linear form, for example, there is a distinct “C” shape visible in a view shown in the first plot in Figure 16. What this means is that there are strong non-linear dependencies amongst the nutrients in the soil, and there are regions where the presence of some nutrients will be accompanied by the absence of other nutrients. Now a careful inspection of the axes in the lower left of the first plot indicates that almost all of the variables contribute to this view: it is a mess of variable axes, meaning that B, K, Ca, Mg, Zn, Cu, Mn and Fe all are important contributors to this view.

With this many variables it is usually not easy to make a simple interpretation of the non-linear shape, other than there is a strong non-linear dependency between the values of nutrients in the soil. But, there is some similarity to the relationship we saw in the last section between Copper and Iron, and these two variables have large contribution to this
Figure 16: Grand tour views illustrating shape of the point cloud. (Left) Initial view seen in grand tour, (middle) view refined by eliminating all variables except Copper, Calcium and Iron, keeps the structure intact, (right) the low yield values (red) are concentrated at one end of the “C” shape.

view. So it is probably safe to assume that the non-linear relationship that we see here is mostly due to the relationship between Copper and Iron, and also to some extent Calcium with Iron. If we remove all the variables from the view except for these three we can see that the non-linear structure remains.

Now an interesting aside to this structure is that it relates to the yield in the field. At the top end of the “C” the yield is almost exclusively low (red points), and as we move around the “C” to the bottom the yield gets progressively better (green). What does this mean? Clearly, the yield is different according to the combination of nutrients available. To understand this we would need to better characterize the constituents of the soil at both ends. A closer study of this phenomenon in the numerical studies helps to unravel the mystery a little.

Linking Yield to Spatial Location

Using linked brushing we can display the locations of the high, medium and low yields in the field, as shown in Figure 17. There appears to be very strong spatial contiguity between sites with the different yield classifications. Note that these sites also have fairly well-defined differences in the nutrient make-up: the high yield sites had more Calcium, Copper, Potassium, Magnesium, Manganese, Phosphorus and Zinc on average.
3.3 Numerical Analysis Methods for Multivariate Data (Ignoring Spatial Location)

This section will discuss several approaches for analyzing multivariate data. Throughout this section, we continue to ignore the location information associated with each observation. In the next section, we discuss an approach for incorporating spatial information in a statistical model.

3.3.1 Multiple Regression Analysis

Multiple regression entails the measurement of \( k \) explanatory variables and a response variable on each of \( n \) observations. For the purpose of discussion, we proceed here under the admittedly faulty assumption that the soil and yield data are independent observations.

Detecting and Handling Multicollinearity

It is often the case in a multiple regression setting that the explanatory variables exhibit high multicollinearity (that is, high correlations exist among the explanatory variables). When multicollinearity exists, it becomes difficult to accurately assess the importance of each of the explanatory variables due to the fact that standard errors for regression coefficients are inflated.

The use of Variance Inflation Factors (VIFs) or condition numbers are approaches to
identifying multicollinearity in a regression setting. The VIF for the $i$th regression coefficient can be calculated as $1/(1-R_i^2)$, where $R_i^2$ is the multiple correlation coefficient obtained by regressing the $i$th explanatory variable on all of the other explanatory variables. A simple rule of thumb is that any VIF which exceeds 10 should be cause for some concern (Myers, 1990, p. 369). Regressing the 10 transformed soil variables (with transformations described in Section 2.1) on corn yield gives VIFs in the range of 1.65 to 5.93, indicating that multicollinearity is not a serious problem here.

The largest eigenvalue of the explanatory variables’ correlation matrix divided by the smallest eigenvalue has been defined to be the condition number of the correlation matrix. When the condition number exceeds 100, one should be concerned about the effect of multicollinearity. For the transformed soil variables, the condition number is 44.43, indicating no serious multicollinearity problems.

When there is cause for concern about multicollinearity, one might consider using ridge regression (a biased estimation technique which overcomes the estimation problems introduced by the multicollinearity in the explanatory variables). Another estimation tool would be Principal Component Regression, a regression using a subset of the principal components. Principal components are artificial variables corresponding to orthogonal dimensions of the covariance matrix of the explanatory variables and will be discussed in greater detail in Section 3.2.

Regression of Corn Yield on the Transformed Soil Variables

Because we have no cause for alarm about multicollinearity among the 10 transformed soil variables, we proceed with a standard multiple regression. The results follow below.

Residual Standard Error = 13.6748,  Multiple R-Square = 0.5581

$N = 214$,  $F$-statistic = 25.6406 on 10 and 203 df,  $p$-value = 0

<table>
<thead>
<tr>
<th>Variable</th>
<th>coef</th>
<th>std.err</th>
<th>t.stat</th>
<th>p.value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>77.0352</td>
<td>28.6212</td>
<td>2.6915</td>
<td>0.0077</td>
</tr>
<tr>
<td>B</td>
<td>0.1831</td>
<td>0.7662</td>
<td>0.2390</td>
<td>0.8114</td>
</tr>
<tr>
<td>1/Ca</td>
<td>28668.3481</td>
<td>11654.0203</td>
<td>2.4600</td>
<td>0.0147</td>
</tr>
<tr>
<td>log(Cu)</td>
<td>14.6165</td>
<td>3.7833</td>
<td>3.8634</td>
<td>0.0002</td>
</tr>
<tr>
<td>log(Fe)</td>
<td>-0.0121</td>
<td>0.0201</td>
<td>-0.6014</td>
<td>0.5483</td>
</tr>
<tr>
<td>1/K</td>
<td>6.7580</td>
<td>4.5422</td>
<td>1.4878</td>
<td>0.1383</td>
</tr>
<tr>
<td>1/Mg</td>
<td>-4755.3761</td>
<td>1869.0567</td>
<td>-2.5443</td>
<td>0.0117</td>
</tr>
<tr>
<td>log(Mn)</td>
<td>3.2404</td>
<td>4.6383</td>
<td>0.6986</td>
<td>0.4856</td>
</tr>
<tr>
<td>Na</td>
<td>0.0498</td>
<td>0.5456</td>
<td>0.0913</td>
<td>0.9274</td>
</tr>
<tr>
<td>log(P)</td>
<td>5.7629</td>
<td>3.5636</td>
<td>1.6171</td>
<td>0.1074</td>
</tr>
<tr>
<td>log(Zn+1)</td>
<td>7.8570</td>
<td>5.1387</td>
<td>1.5290</td>
<td>0.1278</td>
</tr>
</tbody>
</table>
We choose a subset of four explanatory variables based on the fact that \( R^2 \) increases substantially as the subset size increases from one to five and then increases only slightly as more variables are added. The subset of five explanatory variables that yields the largest multiple correlation coefficient is the set containing \( \text{1/Ca, log(Cu), 1/Mg, log(P), and log(Zn+1)} \). The results from regressing these five variables on corn yield follow below.

\[
\text{Residual Standard Error} = 13.5997, \quad \text{Multiple R-Square} = 0.5522
\]
\[N = 214, \quad \text{F-statistic} = 51.2982 \text{ on } 5 \text{ and } 208 \text{ df, p-value} = 0
\]

<table>
<thead>
<tr>
<th>coef</th>
<th>std.err</th>
<th>t.stat</th>
<th>p.value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>111.4641</td>
<td>6.9274</td>
<td>16.0904</td>
</tr>
<tr>
<td>1/Ca</td>
<td>25102.5392</td>
<td>9740.9125</td>
<td>2.5770</td>
</tr>
<tr>
<td>log(Cu)</td>
<td>15.1719</td>
<td>3.3668</td>
<td>4.5063</td>
</tr>
<tr>
<td>1/Mg</td>
<td>-4821.0464</td>
<td>1657.3948</td>
<td>-2.9088</td>
</tr>
<tr>
<td>log(P)</td>
<td>9.2383</td>
<td>2.2675</td>
<td>4.0743</td>
</tr>
<tr>
<td>log(Zn+1)</td>
<td>8.4403</td>
<td>4.7217</td>
<td>1.7876</td>
</tr>
</tbody>
</table>

Note that very little is lost in terms of \( R^2 \) by omitting the other five soil variables. The \( p \)-value for log(Zn+1) indicates that it could justifiably be dropped from the equation. The subset containing \( \text{1/Ca, log(Cu), 1/Mg, and log(P)} \) is the best subset of four soil variables in terms of \( R^2 \).

However, there is something counter-intuitive about this result: the coefficient for \( \text{1/Ca} \) is positive, which says that the more Calcium in the soil the lower the yield. This doesn’t match the original plots, where it is clear that Calcium has a positive relationship with yield. Moreover when we do a regression with only \( \text{1/Ca} \) as the independent variable the coefficient is correctly negative. So the sign switches when more variables are included. This can be explained very simply: Copper is highly correlated with Calcium, and hence multicollinearity does exist and is a serious problem even though the numerical procedures don’t declare it. Hence it is important here to use the best four variables, rather than the best five to get a sensible result. The best four variables give as good a model as the best five variables.

\[
\text{Residual Standard Error} = 13.5984, \quad \text{Multiple R-Square} = 0.5501
\]
\[N = 214, \quad \text{F-statistic} = 63.8952 \text{ on } 4 \text{ and } 209 \text{ df, p-value} = 0
\]

<table>
<thead>
<tr>
<th>coef</th>
<th>std.err</th>
<th>t.stat</th>
<th>p.value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>141.4148</td>
<td>11.4003</td>
<td>12.4045</td>
</tr>
<tr>
<td>log(Cu)</td>
<td>13.1430</td>
<td>2.6807</td>
<td>4.9028</td>
</tr>
<tr>
<td>1/K</td>
<td>-986.9279</td>
<td>358.4932</td>
<td>-2.7530</td>
</tr>
<tr>
<td>1/Mg</td>
<td>-2617.0983</td>
<td>1488.1863</td>
<td>-1.7586</td>
</tr>
<tr>
<td>log(P)</td>
<td>7.0551</td>
<td>2.7388</td>
<td>2.5759</td>
</tr>
</tbody>
</table>
The regression equation for the best subset of five soil variables tells us that when all the soil variables are at their lowest - Copper is 0.240ppm, Potassium is 33ppm, Magnesium is 178ppm and Phosphorus is 5ppm - the estimated Yield will be \(141.4148+13.1430\log(0.240) - 986.9279/33 - 2617.0983/178 + 7.0551\log(5) = 89.4\) bushels/acre. Whereas when the soil variables are at their highest - Copper is 4.67ppm, Potassium is 363ppm, Magnesium is 875ppm and Phosphorus is 114ppm - the estimated Yield will be \(141.4148 + 13.1430\log(4.67) - 986.9279/363 - 2617.0983/875 + 7.0551\log(114) = 189.4\). Their is a substantial gain in Yield when these nutrients are at a high level.

Now although the other variables were not included in the regression model, they are implicitly represented, and remain important to obtaining high yields. For example, Calcium is highly correlated with Copper and Magnesium. We could estimate the amount of Calcium needed to get a high yield by regressing Calcium on Copper and Magnesium. Similarly, we could obtain recommended amounts for Manganese by examining Copper and Phosphorus, and amounts for Zinc by examining Copper, Potassium, Magnesium and Phosphorus. On the other hand it is not possible to get estimates for the amount of Boron and Sodium needed from this model from the graphical analysis it is clear that the upper limits of these nutrients consistently produced high yield, so we could roughly recommend limits for these nutrients, although we couldn’t estimate the increase in yield. Iron is only important to consider if the Calcium or Copper content is low, so we could consider it a substitute for these nutrients.

### 3.3.2 Principal Component Analysis

**Introduction to Principal Component Analysis**

In principal component analysis, we seek to maximize the variance of a linear combination of the variables. This linear combination, called the first principal component, can then be thought of as the “most important” dimension of the data and can be used as a one-dimensional simplification of the entire data set. Similarly, the second principal component might be thought of as the second most important dimension of the data. Principal component analysis (PCA) is often used in situations where dimension-reduction would facilitate the understanding of a high-dimensional process.

In terms of our data, we wish to simplify the relationship between soil characteristics and yield. Though we will not use Principal Component Analysis to relate soil to yield directly, it is our belief that understanding the important relationships among the soil variables will eventually lead to a better understanding of the relationship between soil and yield.

The first principal component is simply the linear combination of the original variables defined by the first eigenvector of the covariance matrix for the data. That is, if \(a_1\) is the first eigenvector of the covariance/correlation matrix and \(x\) is the \(p\)-dimensional observation vector, then the first principal component is equal to \(a_1'x\). Similarly, the 2nd through \(p\)th components are defined by the 2nd through \(p\)th eigenvectors.
The degree to which the first $k$ components adequately represent the variability in the full data set is assessed by either graphically or numerically examining the relative magnitude of the variances of the components. Since the variability of the $i$th principal component is equal to the $i$th largest eigenvalue ($\lambda_i$) of the covariance matrix, the somewhat ad hoc notion of “proportion of variance explained” by the first $k$ components is often quantified with

$$\frac{\sum_{i=1}^{k} \lambda_i}{\sum_{i=1}^{p} \lambda_i}$$

Principal component analysis was conducted using both the standardized untransformed soil data set and the standardized transformed soil data set. For succinctness, we discuss only the analysis of the standardized transformed soil data set. The transformation for each variable in the transformed data set was chosen via the graphical analysis described in Section 2.1. Note that although these analyses will not involve yield in any way, it will be shown that the first principal component may be of use in understanding the relationship between soil quality and yield.

**PCA of the Standardized Transformed Data**

A PCA was conducted using the standardized transformed data. Recall that the transformation for each variable in the transformed data set was chosen via the graphical analysis described in Section 2.1. The outlier (point number 200) identified in the plots of Section 2.1 was omitted for this analysis.

The eigenvalues of the covariance matrix were: 4.288, 1.906, 1.593, 0.768, 0.624, 0.259, 0.222, 0.138, 0.114, and 0.089. Thus, 43% of the total variance is explained by the first principal component, 62% of the total variance is explained by the first two principal components, and 86% of the total variance is explained by the first four principal components. Figure 18 also shows a scree plot of the eigenvalues. Where the plot makes an elbow indicates that this is a reasonable number of principal components to use; in this case we should consider using 2, 4 or 6 principal components. The first four eigenvectors of the covariance matrix are given in Table 1.

The coefficients associated with the first principal component in this analysis indicate that all of the variables contribute. The order of importance from greatest to least is Copper, Zinc, Calcium, Magnesium, Manganese, Phosphorus, Potassium, Sodium, Boron and Iron. Perhaps the order is not as important as the fact that there is a positive loading for each variable on the first principal component. (Note that the negative coefficients match with the inverse transformations so that the interpretation is that this is a positive coefficient on the original data.)

The top seven contributions come from variables flagged to be important in the graphical analysis; the ones that were most highly correlated with yield. The plot of the first principal component versus yield (given in Figure 19) shows a strong positive linear relationship with

20
Figure 18: A scree plot of the eigenvalues for the soil nutrients indicates we should consider using 2, 4 or 6 principal components.

Table 1: The first two eigenvectors from the covariance of the standardized transformed data. Relatively large coefficients are in bold.

<table>
<thead>
<tr>
<th>Variable</th>
<th>$a_1$</th>
<th>$a_2$</th>
<th>$a_3$</th>
<th>$a_4$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Boron</td>
<td>0.103</td>
<td>0.254</td>
<td><strong>0.659</strong></td>
<td>0.087</td>
</tr>
<tr>
<td>1/Calcium</td>
<td>-0.399</td>
<td><strong>0.350</strong></td>
<td>-0.044</td>
<td>0.018</td>
</tr>
<tr>
<td>log(Copper)</td>
<td><strong>0.448</strong></td>
<td>-0.125</td>
<td>-0.009</td>
<td>0.044</td>
</tr>
<tr>
<td>log(Iron)</td>
<td>0.026</td>
<td><strong>0.471</strong></td>
<td>-0.161</td>
<td><strong>0.822</strong></td>
</tr>
<tr>
<td>1/Potassium</td>
<td>-0.256</td>
<td>-0.477</td>
<td>0.227</td>
<td>0.296</td>
</tr>
<tr>
<td>1/Magnesium</td>
<td>-0.394</td>
<td>0.212</td>
<td>0.038</td>
<td>-0.232</td>
</tr>
<tr>
<td>log(Manganese)</td>
<td><strong>0.363</strong></td>
<td>-0.191</td>
<td>-0.078</td>
<td>0.227</td>
</tr>
<tr>
<td>Sodium</td>
<td>0.161</td>
<td>0.142</td>
<td><strong>0.666</strong></td>
<td>-0.043</td>
</tr>
<tr>
<td>log(Phosphorus)</td>
<td><strong>0.262</strong></td>
<td><strong>0.489</strong></td>
<td>-0.161</td>
<td>-0.328</td>
</tr>
<tr>
<td>log(Zinc+1)</td>
<td><strong>0.425</strong></td>
<td>0.087</td>
<td>-0.096</td>
<td>-0.114</td>
</tr>
</tbody>
</table>
yield. Using the transformed data, the correlation between the first principal component and yield is 0.713.

Recall from Section 2.1 that Iron, Copper and Calcium have an interesting interaction with yield. This nonlinear relationship seems to be picked up by the first vs the fourth principal component: the first has strong components of Copper and Calcium, and the fourth is mostly composed of Iron. Notice also, that the third principal component is composed mostly of Boron and Sodium. These two nutrients affect yield in a rather different manner to the other nutrients. Low values don’t automatically mean lower yield, but it does mean considerably more variability in yield. High amounts of these nutrients assures consistently higher yields. It is interesting that principal component analysis detects this unique relationship between the two.

**Principal Component Regression Analysis**

In an earlier section, we concluded that Principal Component Regression Analysis is unnecessary for the purposes of overcoming multicollinearity problems in the data. However, it is instructive to determine which principal components are the best predictors of yield. The output from the regression of corn yield on the first four principal components follows below. Given that we have considered four principal components sufficient to describe the variation in this data we use just the first four for the regression analysis.

Residual Standard Error = 13.9506,  Multiple R-Square = 0.5265
N = 214,  F-statistic = 58.1053 on 4 and 209 df, p-value = 0

<table>
<thead>
<tr>
<th></th>
<th>coef std.err</th>
<th>t.stat</th>
<th>p.value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>153.0874</td>
<td>0.9536</td>
<td>160.5295</td>
</tr>
</tbody>
</table>
Interestingly, the first principal component is overwhelmingly important for modeling yield. Recall, that the first principal component is composed of a positive contribution of ALL soil nutrients, so this is essentially saying that all the nutrients are important for modeling yield.

The second principal component has the next most important contribution to yield. It is not significant at the 5% level but still the p-value is small enough to consider the contribution that this component makes to modeling yield. Now it is not so easy to understand the second principal component. It mostly has contributions from Calcium, Iron, Potassium, and Phosphorus, and notably Calcium and Iron contribute in opposite ways. This component may represent the way Calcium and Iron interact in affecting yield.

The third principal component is mostly composed of Boron and Sodium. It doesn’t contribute significantly to modeling yield, either, but the p-value is also small enough to take careful note of how the component affects yield. This is fairly clear: this component represents the contribution of both Boron and Sodium to yield.

3.4 Exploring Spatial Dependence

When measurements are collected in such a systematic spatial manner, the assumption of random sampling, and hence independent errors in a regression analysis is not valid. It is interesting to explore the nature of the similarity of variable values based on their spatial proximity. We study this here using a geographic information system (GIS, called ArcView) linked to XGobi (Symanzik, Cook, Lewin, Majure & Megretskaia 1999).

Initially it is interesting to look at the yield values over the field, essentially something called a yield map. We did this roughly in Figure 1, by coloring the yield according to low, medium and high values, plotted in the spatial domain. Another approach would be to display the yield by space in a perspective plot, or contour plot, or a 3D rotating plot. Figure 20 displays a sequence of rotations of the 3 variables. There are not overt trends of yield over the field, although there is a difference in variance. One area of the field has very low values, as well as high values. Removing trends is important before exploring the data for spatial dependence (next paragraph).

Another common type of plot to use is a variogram cloud plot, the two variables are computed as follows: \( d_{ij} = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2} \), \( \gamma_{ij} = |\text{yield}_i - \text{yield}_j|^{1/4} \). This type of plot allows us to examine how the variable values change as points get spatially farther away. You would expect that the closer two points are in space, the closer the variable value will be. Figures 21 and 22 examines the variogram cloud plot for yield.
Figure 20: Sequence of rotations of the yield against the spatial coordinates.

Variogram cloud plots can also be used to explore for spatial anomalies: points that have remarkably different values from their close neighbors, but not unusual in the overall scale of the variable values. Figure 23 examines yield for spatial anomalies. We brush the larger values at each of the distances, and explore the corresponding pairs of locations of these points in the field (displayed by a line in the ArcView plot). Larger values means that the yield values are unusually large in respect to the spatial distance.

Figures 24 and 25 explores the spatial dependence in copper measurements across the field.
Figure 21: Variogram cloud plots of yield, with differing degrees of jittering, and the last one is smoothed using a spline smoother.
Figure 22: Smoothed Variogram cloud plot of yield, using a spline smoother.
Figure 24: Variogram cloud plots of copper, brush is exploring the spatial anomalies.

Figure 25: Smoothed Variogram cloud plot of copper, using a spline smoother.
4 Conclusions

Graphical Analysis

From the graphical analysis of the soil nutrient data, it appears that all the nutrients have a positive relationship with yield, with the exception of Iron. Seven of the ten (transformed) nutrients have a strong linear relationship with yield. Ordered from strongest to weakest, the seven transformed variables are log(Copper), log(Zinc+1), 1/Potassium, log(Phosphorus), 1/Magnesium, log(Manganese), and 1/Calcium. That is, higher yield was obtained with higher Copper, higher Zinc, higher Potassium, higher Phosphorus, higher Magnesium, higher Manganese, and higher Calcium.

With Boron and Sodium, the yield is uniformly higher when the content is higher; with lower content yield ranges across the full spectrum. There is one exception to this: one site had very low yield despite the relatively high quantities of Boron and Sodium (and also Phosphorus). There is clearly something unusual happening at this site.

Now there are some interesting relationships between the nutrients when Iron is involved. Yield is good regardless of Iron content if Calcium content is high; however, if Calcium is low Yield improves with higher Iron! Similarly this holds for Copper and Iron also.

Numerical Analyses (Ignoring Spatial Location)

The regression analysis indicated that four variables were important for modeling yield: Copper, Potassium, Magnesium, Phosphorus. But we know from the graphical analysis that Calcium, Manganese and Zinc are also very important. These variables implicitly factor into the regression model through their close correlation with the four variables in the model. The contribution of Boron, Sodium and Iron to yield were too difficult to describe by this model, but the principal component analysis explained well the importance of these variables.

All the soil nutrient variables had positive loadings on the first principal component. This principal component was found to be the most important predictor of yield, in the principal component regression. This says that all the soil nutrients are important in increasing yield. With the first three principal components in a regression model we can describe how each of the nutrients can be used to improve yield.

We conclude that based on the regression analysis, the variables which seem to be most important in prediction of corn yield are: Copper, Potassium, Magnesium, and Phosphorus. Further, we conclude that the first principal component is a reasonable one-dimensional measure of “soil quality.”

Based on the graphical and numerical analysis we may be inclined to explore if yield depends on a low-dimensional process of possibly one or two latent factors that would adequately represent the soil’s “quality” or “capacity for high corn yield.”

Graphical Exploration of Spatial Dependence
There is fairly strong spatial dependence in yield up to about 30m, and then the spatial
dependence tapers out. For copper, there is fairly strong spatial dependence up to about
40m. (Other variables need to be examined similarly.) It would also be interesting to examine
cross-variogram cloudsto examine spatial dependence between variables.

One area of the field appears to have spatial anomalies in the yield variable, yield values
that vary dramatically from close neighbors. This region of the field did have the lowest
values of yield. We might want to examine other variables such as elevation, soil moisture
in this part of the field to understand this variability.

References

variability within a central iowa field’, Transactions of the ASAE 40, 883–889.


3.0 and XGobi: Insight Behind the Front End’, Journal of Computational and Graphical
Statistics. Forthcoming.