Chapter 4

Space and Time

4.1 Background

Space and time variables arise when measurements are taken over a geographical region, or over a number of time points. In each case the assumption of random sampling no longer applies. There are several additional steps to exploring data having a space or time component. First to explore the multivariate trends, treat space and time as separate and different variables, and link plots of these variables to the multivariate variable plots. Secondly, once the trend has been removed, explore spatial dependence by constructing new plots, such as lag plots, or variogram clouds. These are especially powerful if they are linked to the other plots as well.

This chapter will describe the approach as we use them to explore data from agriculture, and oceanographic records.

4.2 Case Study: Comprehensive Ocean-Atmosphere Data

4.2.1 Data Description

In situ weather observations taken by merchant marines over the past 150 years. We use January 1980-December 1991, 30°S to 30°N and 160°E to 75°W (Pacific Ocean). Various cleaning and processing was done to get regularly gridded data. Variables are sea surface temperature (SST), sea level pressure (SLP), wind speed (WndSpd), wind direction (WndDir). We use the long term means, 1800 data points here.

Source: http://ingrid.ldgo.columbia.edu/SOURCES/.COADS
4.2.2 Comparing Traditional with Interactive

In the traditional approach taken by atmospheric scientists with this data the winds are plotted as vectors on the map (Figure 4.1). The vector arrow points towards direction wind is blowing to, and length represents speed (longest vector represents $9 m/s$). The approach is good for perceiving the general trends across the globe, such as, above the equator, winds tend to come from the north and east, and below the equator the winds tend to come from the south and east. Is less successful for exploring small deviations and anomalies.

With some restructuring of the variables we can get the data in a format suitable for interactive querying. The variable wind direction is re-coded as cosine and sine of wind direction (Figure 4.2). This allows us to plot it in a modular format, recognizing the equality of 0 and 360. It also facilitates brushing around the circle, like brushing around the points of a compass. Jitter is added to alleviate the heavy overplotting of easterly winds.

Ideally the map view is displayed with a good map drawing program, so that geographic boundaries and landmarks are included in the view. There is a link between XGobi and ArcView\(^1\) which allows linked brushing between plots displayed in XGobi and map views in ArcView. Here we use the link (Symanzik, Cook, Lewin-Koh, Majure & Megretskaia 1999). Figure 4.3 displays a series of stills from a brushing session on this data. The brush moves from north-westerly winds clockwise in the XGobi plot of wind direction. In the absence of a good map drawing program XGobi can be used to display the latitude and longitude

\(^1\)ArcView is a trademark of Environmental Systems Research Institute, Inc.
Figure 4.2: Textured dot plot of wind directions (0° to 360°) ⇒ Sine versus cosine plot ⇒ “jitter” added.

of the measurements, as well as the plots of other variables.

4.2.3 Exploring Other Variables

The plot of sea level pressure and wind speed has some unusual features. These are explored in Figure 4.4. It is interesting to see that the Pacific near Australia, and near central America have very similar patterns of sea level pressure and wind speed (in the average, since these are long term means over 11 years). There is an unusual combination of pressure and wind speed in the south eastern Pacific, off the coast of south America.

4.2.4 Exercises

1. Examine the relationship between sea level pressure and sea surface temperature, and the relationship of patterns to spatial location.

4.3 Case Study: Baker Field Data

4.3.1 Data 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
Figure 4.3: Brush moves from north-westerly winds clockwise in XGobi, linked to map displayed in ArcView.
Figure 4.4: Exploring anomalies in sea surface pressure and wind speed.
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 x 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 bushells
- 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?"

### 4.3.2 Scatterplots

The field map (Figure 4.5) 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 4.6. 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.

**Calcium.** Figure 4.7 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 transforma-
Figure 4.6: Scatterplot of Boron versus Corn Yield. Levels of high, medium and low Yield are brushed in green, navy and red, respectively.

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 0.387m but the relationship is still non-linear.)

Figure 4.7: Scatterplots of Calcium versus Corn Yield and 1/Calcium versus Corn Yield. Levels of high, medium and low Yield are brushed in green, navy and red, respectively.

Copper. Figure 4.8 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.
Figure 4.8: 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.

4.3.3 Exercises

1. Explore the relationships between the remaining nutrient variables and Yield. Use transformations as appropriate. Which variables are the most important for obtaining high yield?

2. There is something very interesting about Iron in relation to Calcium and Yield. Can you find what it is?

3. Use the smoother tool to explore the relationships between variables and Yield.

4.3.4 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 4.9 and 4.10. 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 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.
Figure 4.9: 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 4.10: 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.
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.

4.3.5 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. In this data, from the grand tour, one notices several observations that do not “move” with the rest of the data through the projection space. These observations are outlying from the main concentration of points. Figure 4.11 shows three different views of the grand tour of the untransformed soil variables illustrating the outlying nature of points 81 and 82. These have the two highest values for Calcium. There is a simple explanation here. The axes suggest that these points are outlying in Calcium and to some extent Magnesium. In general, outliers can be more complex than this, in that they may not be outlying in any one (or two) variable(s) but may be extreme in the multivariate space, when three or more variables are used. The grand tour can help find and identify such points and the linear combination of variables where the extremeness occurs. In this data, taking the inverse of Calcium makes the outlying nature of observations 81 and 82 much less pronounced.

![Figure 4.11: Three different views of the Grand Tour of the untransformed soil variables, illustrating outliers, and general skewness in the data.](image)

It is typical to explore this type of data for cluster structure also. 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 4.12. 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. 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 view. So we explore these two variables and look one-by-one for other variables to find the important contributors to this structure (Figure 4.12). Note also that the shape is present whether the variables are transformed or not.

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? The variables that contribute to the structure are Fe plotted orthogonally to a contrast of $1/Ca$ and $ln(Cu)$. The high to low yield mostly falls in the direction of the contrast, which means that high yield is associated with high Calcium and high Copper. The curvature is primarily due to the relationship of Iron and these other two nutrients.
4.3.6 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 4.13. 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.

4.3.7 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.

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 ??, 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 4.14 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).

Figure 4.14: Sequence of rotations of the Yield against the spatial coordinates.

Another common type of plot to use is a variogram cloud plot, the two variables are computed as follows: 
\[ d_{ij} = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2}, \gamma_{ij} = |yields_i - yields_j|^{1/4}. \]
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 4.15 examines the variogram cloud plot for Yield, and 4.16 is the smoothed variogram cloud using the smoothing tool.

Variogram cloud plots can also be used to explore for spatial anomalies: points that have remarkably different values from their close neighbors, especially if they are not unusual in the overall scale of the variable values. Figure 4.17 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.

4.3.8 Exercises

1. Use the smoother on the variogram cloud plots of the nutrient variables to examine the spatial dependence.
Figure 4.15: Variogram cloud plots of Yield, with differing degrees of jittering.
Figure 4.16: Smoothened Variogram cloud plot of Yield, using a spline smoother.

Figure 4.17: Variogram cloud plots of Yield, brush is exploring the spatial anomalies.