three ways: by suppressing known inclusions smaller than the head of pencil on the finished map, by observing an SCS guideline that "limiting" and "non-limiting" inclusions ought not exceed 15 and 25 percent, respectively, of the area of a mapped soils unit, and by listing in the detailed description of each soil category the estimated overall percentage of inclusions and the names of included soils recognized in the field. More precise mapping is possible, to be sure, but the map sheets would be enormous, the survey could consume several lifetimes, and the field work might require more digging than landowners would tolerate.

### Margin of Error

Throughout the hearing the Town and PURE brought up, again and again, the phrase "margin of error." Each intervenor presented a retired SCS soil scientist as an expert witness, and both experts testified that the maps were not sufficiently accurate to calculate the area of prime soils to within 7 percent. One testified that the map's margin of error was "15 to 25 percent," whereas the other considered the margin narrower, "in the neighborhood of 10 percent." But neither witness addressed error as a task-specific concept.

Another term treated with equal obscurity was "accuracy"—the intervenors, and at times the ALJ, seemed unwilling to acknowledge that "accuracy" is meaningful only in reference to a specific measurement or problem. It was frustrating to hear adults fail to differentiate the internal homogeneity of soil mapping units from the positional accuracy of the delineated boundaries, or to distinguish the accuracy of a calculated percentage from the accuracy of determining preponderance. Judicial proceedings don't allow someone not testifying to inject a comment or raise a hand. So you roll your eyeballs, wring your hands, and think of nasty things to say about lawyers in a GIS column.

### The Outcome

OCRRA is confident the state ultimately will award the landfill permit, and that the courts will reject any subsequent legal challenge. After all, numerous complementary measurements indicate without exception that group 2 soils are not even close to predominant. And no evidence suggests that the Onondaga County soil survey is significantly less accurate than similar maps for other New York counties. Moreover, had the soil survey been demonstrably flawed or some plausibly appropriate measurements fallen on the other side of the 50-percent threshold, OCRRA's case would still be solid. The agricultural provision is based directly on the delineated land classification—making the existing soils map an "official map," which need not be a scientifically accurate map, and the standard of proof in civil proceedings is preponderance of evidence, not reasonable doubt. Even so, the Town and PURE were able to thwart the timely issuance of a permit, and with persistent legal maneuvering the intervenors might delay the process further, until OCRRA's options-to-purchase expire. Should the matter reach state Supreme Court and even the Court of Appeals, a legal precedent useful in similar adjudications might emerge. If soil scientists, statisticians, and GIS experts fail to develop well-defined, widely accepted procedures for addressing simple questions involving uncertain data, lawyers, judges, and jurors will develop the standards for them.

### References

Huff, D. 1954. *How to Lie with Statistics.* W.W. Norton. New York.

Monmonier, M. 1991. *How to Lie with Maps.* University of Chicago Press. Chicago.

Mark Monmonier
Syracuse University
mon2ier@syr.edu

## TOPICS IN SCIENTIFIC VISUALIZATION

# Converting Tables To Plots: A Challenge From Iowa State

by Daniel B. Carr and Sarah M. Nusser

### 1. Introduction

In October 1995 I gave a seminar for the Iowa State Statistics Department. My topic was converting tables into plots. Sarah Nusser, noting that Iowa is awfully close to Missouri (the show me state), suggested that I try my hand at converting one of her summary tables for water and wind erosion. I agreed, provided that she would help me write this article. This article shows my first attempt. Most of the water erosion information appears in two plots. Figure 1 shows the water erosion land classes and erosion rates while Figure 2 shows the corresponding acreage. For brevity, this article omits the similarly designed wind erosion plots. The plots are not trivial but with modest explanation many will be able to see the patterns. I conjecture that (1) more

readers will be willing to study the plots than the corresponding tables and that (2) most readers will be more confident about not missing basic patterns if they use the plots rather than the tables.

In what follows Sarah provides a description of the table and indicates some of the guidance she gave me. Then I describe my efforts to develop the plots. I generally follow the guiding principles of writers such as Cleveland (1985), Tufte (1990) and Kosslyn (1994). However the guidance does not capture all the interactions and special cases that arise in increasingly complex table conversions. (In fact Kosslyn suggests avoiding complex graphics.) While some design decisions come easily, others are struggles. Hopefully readers will pick up a useful idea or two while looking over my shoulder. Quite possibly ideas for better designs will emerge.

## 2. The National Resources Inventory Land Class and Erosion Table

The National Resources Inventory (NRI) is a longitudinal survey conducted by the USDA Natural Resources Conservation Service (NRCS) in cooperation with the Iowa State University Statistical Laboratory. The purpose of the survey is to assess the status and trends at 5-year intervals for natural resources on nonfederal lands of the United States. The agriculture variables of interest include land use patterns, soil types and properties, wind and water erosion, conservation practices, rangeland quality, and conversion of farmlands to nonfarm uses. Environmental concerns include wetlands, earth cover, and habitat measures.

The table in question displays changes over time in soil loss due to water and wind erosion as influenced in relation to land use dynamics. The data used to create the table contains a set of points with a time series of 1982, 1987 and 1992 data on land use, long-term average annual rates of soil erosion induced by water and wind, and NRCS erodibility indices. The six land use categories are cultivated, highly erodible cropland; cultivated, not highly erodible cropland; noncultivated, highly erodible cropland; noncultivated, not highly erodible cropland; Conservation Reserve Program (CRP) cropland; and other land use.

The typical change table considers two years at a time. The earlier year's land use categories define rows. The later year's categories define columns. The table has a two way layout with six entries per cell. The cell entries are the number of points belonging to the cell, the estimated number of acres that follow the pattern defined by the grid cell, and the estimated average water erosion rates and wind erosion rates for the two years. Considering t time points involves examining up to t choose 2 tables. It doesn't take long to become saturated even for those with a high table-reading IQ.

The challenge was to take the information contained in the 3 change tables (82-87, 87-92 and 82-92) and display it on one plot. Ideally it is desirable to look at the grid defined by initial condition (rows of land use categories for the earliest year) crossed with the most recent conditions (columns of land use categories for the most recent year) with some representation of the erosion rates for the different land use classes that were present during the interim years. The grid would be five rows by six columns because the Conservation Reserve Program was not in place in 1982. Since this was a pretty tall order, we decided that a good starting place would be to consider displaying information for only one type of erosion (wind or water) and the acreage associated with the particular land use pattern over time. The acreage is an important measure of the significance of the land use pattern in the dynamics of soil loss.

## 3. Design Thoughts for Figure 1

As Sarah posed the problem, it seemed natural to keep the 5 x 6 matrix for the beginning (1982) and ending (1992) classes. When focusing on water erosion, the challenge was to represent the land classification and erosion values for the time series within each cell. Since I had little space I immediately thought of using parallel coordinates plots to represent the erosion values.

Inselberg (1985) and Wegman (1990) introduced parallel coordinate plots. Parallel coordinate plots are natural to consider when plotting space is at a premium. Parallel coordinate plots are particularly useful for representing two variables (see Carr and Olsen 1995 for a map legend application ) and for representing a small number of time series. In comparison to scatterplot matrices that show all possible pairs of variables, parallel coordinate plots weakly express the relationship between non-adjacent variables. However, for time series the relationships between non-adjacent times are of lesser importance than those between adjacent times, so the compact representation will often suffice.

The next question was how to represent the land class information. My immediate ideas included use of different symbols, line textures and colors. (See Kosslyn 1994 for good symbol and line patterns.) Fortunately there are only six classes. This is crucial in the design considerations because most people can only remember seven plus or minus two chunks of information. With many more classes, creating subclasses and layering the information becomes important. To go for a simple ap-

pearance, I chose to use round dots and to avoid angles and end-points associated with other symbols and line textures. Thus puts all the burden on color to represent the six classes. As discussed later this choice returns to haunt me because the symbols are small and small areas make it difficult to see color differences.

With a design in mind, the next step was to try it out. In the S-Plus context I chose my own row-labeled plot functions (Carr 1994a) to handle the layout and labeling but Trellis (see Cleveland 1993 for examples) might have been easier to use. A little effort went into splitting the data set into cells, some effort went into making a rudimentary function to plot the data in a cell, and a great deal of effort went into attending to details.

Developing the plotting function was not too bad since I had previously developed a parallel coordinate function. Attending to details meant providing global scaling so all plots had the same vertical scale, using a gray background with white grid lines to increase the perceptual accuracy of extraction, scaling the data to keep points inside each cell, and plotting axes labels on the right. Noting that color perception is difficult for small areas, I chose thick lines to connect dots over time. The dot-connecting line color used the most recent class assignment. Thus a line from a 1982 value to a 1987 value would use the class assignment for 1982.

Consider the color-driven class interpretation for the parallel coordinate panels show in Figure 1. The color of the row and column labels provide the color legend. By construction, the color of the row label matches the color of the all 1982 dots in that row. The lines to the 1987 time period also have the same color. The color of the column label matches the color of all the 1992 dots in the column. For example the top row second column dots all start in red and the time series all end in green. The 1987 dot color and subsequent line gives the intermediate transition class.

The first experimental plot (that I care to mention) had overplotted dots. Within a matrix cell, overplotting of 1982 dots did not cause an interpretation problem. For hidden dots the class membership and water erosion values were the same as that of the overplotting dots. Similarly the 1992 dot overplotting did not cause an interpretation problem. However the 1987 overplotted dots were problematic. In some cases I could not tell the class membership. Further I could not pair a line entering an overplotted point with the corresponding line leaving the point so two or more time series became confused with each other. To address this my second try changed the line color half way between the time periods. This paired the corresponding lines entering and leaving a dot since they were all the same color. The lines were often sufficiently visible to establish the color of hidden points. Unfortunately the color change between axes made the lines look more complicated and the strategy did not resolve the problem of overplotted lines near zero.

For my third try, I simply resolved the overplotting of six or fewer dots on the 1987 time axis. Carr (1994a) describes a point nudging algorithm that move dots slightly to avoid overplotting. The algorithm computes a small positive increment for a point if its nearest neighbor on the left (or below) is too close. Similarly the algorithm computes a small negative increment for a point if its nearest neighbor on the right (or above) is too close. At each iteration each point changes by the sum of its two increments. The increments cancel when neighbors on both sides are too close. The nudging spreads out the points that are too close. The "too close" criterion can stop moving points when they just touch. This approach almost works.

### *A point nudging algorithm moves dots slightly to avoid overplotting.*

Nudging can change small positive values to negative values. Some people find the shift from positive to negative erosion rates unacceptable. To avoid this interpretation shift, a further step increments small values so that the smallest value becomes zero. This step does not alter larger values that are separated from the lowest values by a sufficiently large gap.

Figure 1 shows the result of applying the modified nudging algorithm to all water erosion rates. While adjusting the 1987 values is sufficient for interpretation purposes, plots often look simpler without overplotting and Figure 2 exploits the lack of overplotting for the 1982 values. For those concerned about the graphical license taken, the most extreme rate adjustment for 1987 (green-cyan-cyan) increased by 1.49 tons per acre per year. The adjustment for 1992 provided more distortion since more values were small. The worst case was the (gray-red-cyan) increase of 4.1 tons per acre per year. This is not acceptable and such discrepancies needed to be flagged or fixed even though the stacked dots suggest the existence of a problem. The obvious solution for this plot is to drop the nudging for 1992. In future years, rescaling, partial dot overplotting and better nudging can help the approximation. Better nudging moves points horizontally and vertically toward a staggered or hexagonal pattern centered on the axis with considerations for line overplotting.

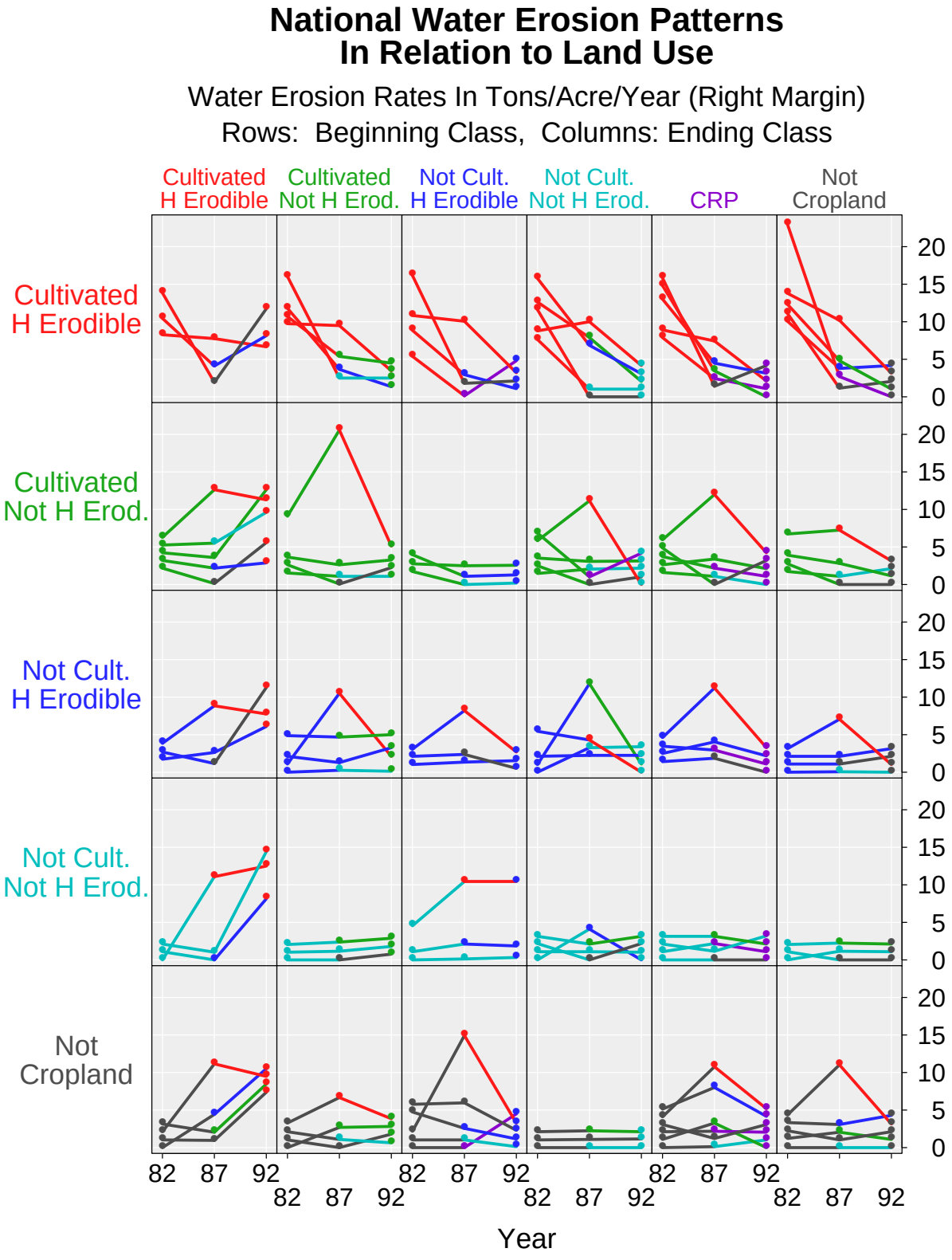


**Figure 1: Land Class By Color, Erosion Rate by Parallel Coordinates**

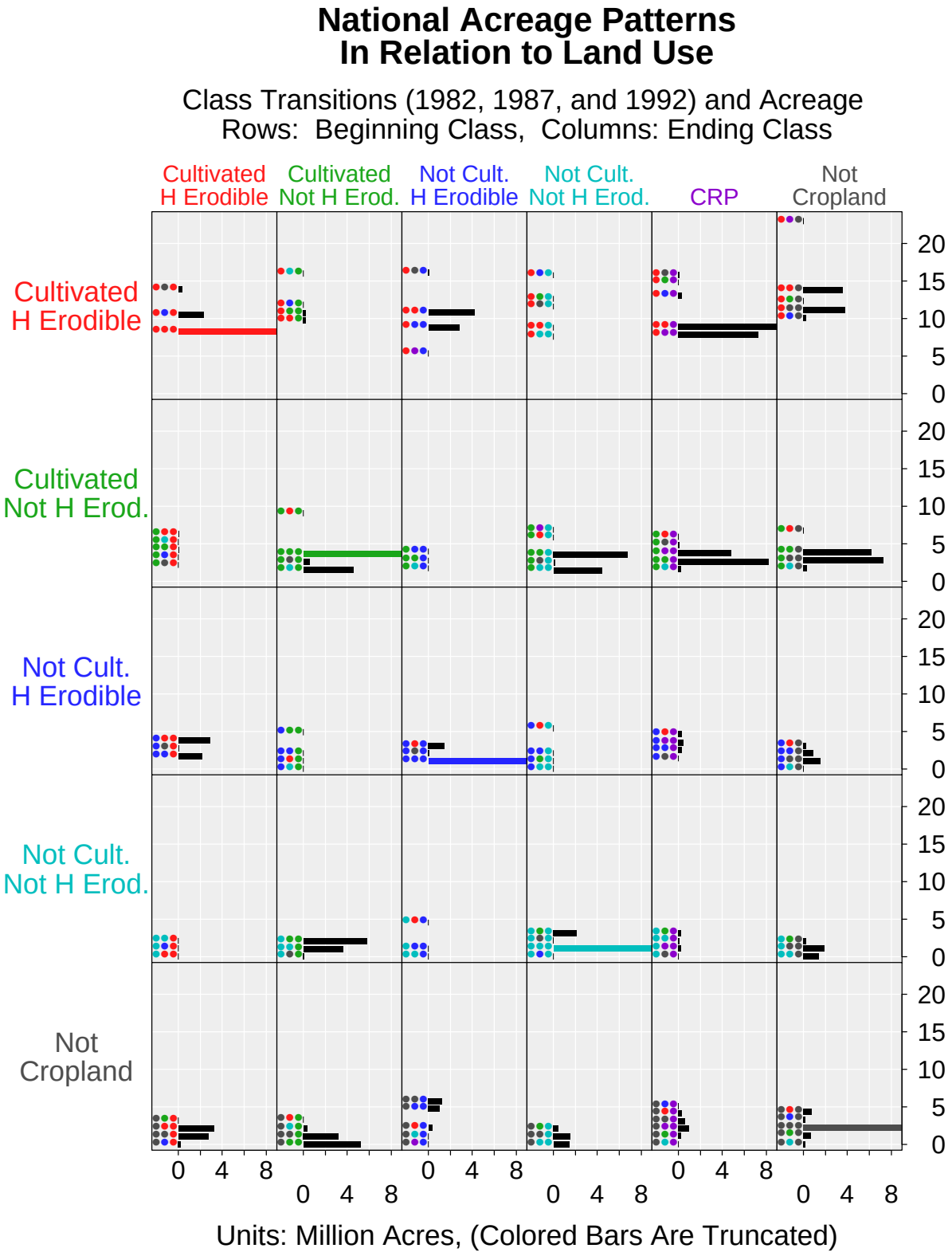


**Figure 2: Linked Plot Using 1982 Water Erosion Values**

Next steps included matrix labeling, plot titling, sizing of the plot to maximize use of the available space, and staggering the time axis labels to avoid overplotting. As indicated above, Figure 1 avoids the need for a separate classification legend by using color for the row and column labels. This keeps the labels very close to the data, minimizes memory burdens and reminds the reader about the matrix construction. However, plotting labels in color raises the issue of unequal contrast with the background. Outlined fonts with background-contrasting outlines increase the colors that can be used in this way but such fonts were not available. (I tried overplotting a font on top of a font outline but the outline wasn't quite big enough. Multiple drawing of the background text in a tight circular pattern and then overplotting will work.) The available software handled titling, color, resizing and staggering of axis tic labels but had to be modified to support multiple line row labels.

Detailed logical labeling conflicts with space constraints. Selecting a compromise involved some iteration. Putting the label for the water erosion axis in the title is not entirely satisfactory. (Later I may try a vertical label that uses horizontal letters.)

Having the option of color is wonderful but attempting to use color well can be frustrating. Every device I used produced different colors from the same postscript file. Hopefully, the newsletter version will be close to my preferred version. Since the task is to represent different classes, hue differences are appropriate. Some have suggested choosing color from among cyan, magenta, yellow, black, red, green and blue to avoid half-toning on many color printers. I hate to admit it but my hue choices ended up coming from this set. I lightened and darkened some colors in the attempt to make them more distinguishable. The color assignment uses red to call attention to the high risk cultivated highly erodible class. Blue, the darkest color on the light background, calls attention to the other highly erodible class. Gray represents the class of least interest, the not cropland class.

In looking for color guidance. Tufte (1990) says to use natural colors and minimize on highly saturated colors. Kosslyn (1994) warns about the different focal lengths for blue and red, and the fact that 8% of men aren't going to see the red-green distinction. He recommends selecting colors that follow the conventions for the specific problem and audience. Brewer (1994) provides guidance for several one variable and two variable color schemes. Her 2 x 3 layout with two shades of green, blue and magenta is one of several elegant examples. This problem is a 2 x 2 layout plus 2 classes. I would like to combine all of the above advice. Perhaps two shades of brown could convey cultivated land and two shades of green could convey uncultivated land. Guidance is easier to find than to apply.

The current plot problem is tougher than common mapping problems. Brushing aside issues associated with two different backgrounds (white and light gray) there are two basic differences. First, the dots and lines have small area. Color perception is dependent on having adequate area. In this problem subtle color distinctions are not going to work. Second, the plot has crossing lines while map regions do not typically overplot. Mapping techniques, such as putting black lines around regions, can mitigate surround induced variation in color perception. If one uses black outlines to make, for example, yellow lines visible, crossing lines generate new, unwanted patterns. The monochrome Figure 3 illustrates the problem. The line crossings in the corresponding (green row, cyan column) cell in Figure 1 draw much less attention. Since the objective is to follow the lines in a flat plot and not through space, outlining is undesirable.

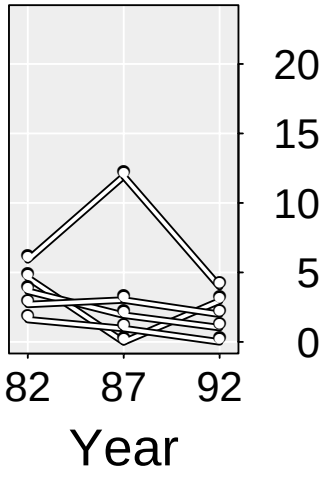


**Figure 3. Outlines**

Graphical design is often open ended. The thick lines in Figure 1 allow me to readily distinguish among all six colors in my prints. In Figure 2 there are just dots. I can still make red/magenta and green/cyan distinctions but have to pay attention. It may be safer to add small white (or black) dots to the magenta and cyan dots or to use larger areas and different symbol shapes to designate class membership. The design for Figure 2 may not have stopped at the best point and that has implications for Figure 1 due to the linkage.

## *4. Interpretation and Acreage Plots*

Patterns in Figure 1 are easy to spot. High erosion rates typically correspond to the cultivated highly erodible class. High positive slope lines typically end in a red dot. That is, a marked increase in erosion generally goes with a transition to the cultivated highly erodible status. Most of the high erosion rates (above 15 tons/acre/per year) appear in 1982. Tracking the (red, red, red) and

(green, green, green) dots over time suggests a slight reduction in erosion rates for land staying in the same class. Trends are flat for the other non-transition lands. Figure 2 aids the interpretation by providing the acreage associated with the land classes.

Figure 2 is basically a horizontal bar plot showing acres. The design is similar to that in Carr (1994b). New variations include the addition of dots to indicate the class transitions, the placement of the bars to match the position of the 1982 values in Figure 1, and the truncation of large acreages for non-transition land. Another choice would be to link erosion rates at the last time period. In either case the dot nudging in Figure 1 provides the space for the horizontal bars in Figure 2.

Scanning Figure 2 for the black transition acreage bars reveals eight substantial bars at the top right. These are transitions to CRP and not cropland classes from the cultivated classes. Somewhat smaller not highly erodible acreages make the transitions between being cultivated and not cultivated. Knowing the acreage helps in assessing the importance of the transitions.

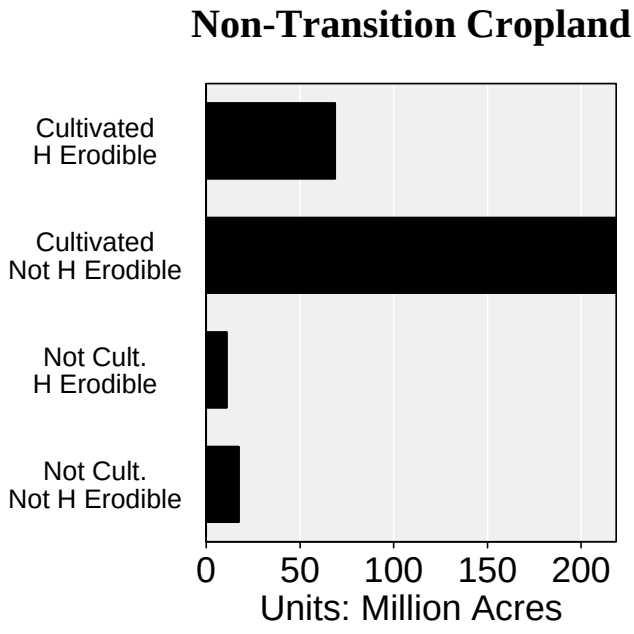


**Figure 4. Bar Plot**

Figure 4 shows the non-transition cropland acreage truncated in Figure 2. To provide resolution this figure omits the not cropland class with its 1.49 billion acres. The plot interpretation is straight forward.

## 5. Closing Remarks

Soil erosion affects long-term agricultural production and historically has concerned farm policy makers. More recently, effects of the degradation of soil resources is being understood in the broader context of environmental concerns for water quality. Degradation of soil function in natural and agroecosystems is an emerging policy issue. Improved abilities to understand soil erosion dynamics are an important part of assessing soil function in ecosystems..

Hopefully the design comments above suggest some new possibilities for displaying time series in relation to a matrix of transition states. We can represent rather complicated relationships graphically and have them understood by scientific audiences. As the relationships get more complicated and detailed, graphical design becomes more crucial and the need for graphical approximation increases. Cartographers and geographers have long recognized that communication requires hiding and simplifying information as a function of map scale. They are taking up the challenge of representing data quality (see Van Der Well, Hootsman and Ormeling 1994, MacEachren and Pickle 1995, and Howard and MacEachren 1995). Similarly the statistical graphics community could address graphical approximations as a function of information complexity and work further in such areas such as representing estimate variability and sampling plan adequacy. Figure 1 is not the full story because nothing has been said about estimate quality.

Figure 1 will generalize to at least one more time period. Adding new time axes is straight forward. A potential problem is that new transition combinations could generate many new lines. With two intermediate time periods there are 36 possible classes per cell, Fortunately future NRI data are likely to add only a few new combinations. I have a few more years before having to face the too many combinations challenge.

This article continues the theme of converting tables into plots. Further challenges are welcome, as are gentle constructive suggestions. Those wanting to obtain the data or adapt the software can do so by anonymous `ftp` to `galaxy.gmu.edu`. The directory is `/pub/submissions/erosion`.

**Acknowledgments**

Graphics research related to this article was supported by EPA under cooperative agreement No. CR8280820-01-0. The article has not been subject to the review of the EPA and thus does not necessarily reflect the view of the agency and no official endorsement should be inferred.

## References

Brewer, C. A. 1994. "Color Use Guidelines for Mapping and Visualization," *Visualization in Modern Cartography*, Eds. MacEachren and Taylor, Pergamon/Elsevier Science Ltd., pp. 123-147.

Carr, D. B. and A. R. Olsen. 1995. "Parallel Coordinate Plots For Representing Distribution Summaries in Map Legends." Proceedings 1 of the 17th International Cartography Association Conference 10th General Assembly of the ICA. pp. 733-742.

Carr, D. B. 1994a. "Converting Plots to Tables," Technical Report No. 101, Center for Computational Statistics, George Mason University, Fairfax, VA. 22030.

Carr, D. B. 1994b. "Using Gray in Plots." *Statistical Computing & Graphics Newsletter*, Vol. 5, No. 2, pp. 11-14.

Cleveland, W. S. 1985. *The Elements of Graphing Data*, Monterey CA: Wadsworth Advanced Books and Software.

Cleveland, W. S. 1993. *Visualizing Data*, Summit NJ: Hobart Press.

Inselberg, A. 1985. The Plane With Parallel Coordinates, *The Visual Computer*, 1, pp. 69-91.

Howard, D and A. M. MacEachren. 1995. "Constructing and Evaluating an Interactive Interface For Visualizing Reliability," Proceedings 1 of the 17th International Cartography Association Conference 10th General Assembly of the ICA. pp. 320-329.

Kosslyn, S. M. 1994. *Elements of Graphic Design*, New York, NY: W. H. Freeman and Company.

MacEachren, A. M. and L. Pickle. 1995. "Mapping Health Statistics, Representing Data Quality," Proceedings 1 of the 17th International Cartography Association Conference 10th General Assembly of the ICA. pp. 311-319.

Van Der Well, F. J. M., R. M. Hootsman, and F. Ormeling, 1994. "Visualization of Data Quality," *Visualization in Modern Cartography*, Eds. MacEachren and Taylor, Pergamon/Elsevier Science Ltd., pp. 313-331.

Tufte, E. R. 1990. *Envisioning Information*, Cheshire, CT: Graphics Press.

Daniel B. Carr
George Mason University
dcarr@voxel.galaxy.gmu.edu



# UNIX Commands

by Phil Spector

One of the most attractive features of UNIX is its extensibility. Even a novice UNIX user can add commands to the operating system, and those commands will be immediately accessible, and can be accessed in the same way system commands are accessed, namely by typing the name of the command. This is because when the UNIX shell encounters a command name, it searches through a set of directories known as the search path until it encounters a file with the name of the command it encountered. If that file has execute permissions appropriately set, then it is executed. (Recall that to set the execute permission for a file, the appropriate UNIX command is `chmod +x filename`).

Two types of files can serve as commands on UNIX systems. First, a text file which has appropriate execute permissions is interpreted as a collection of commands. By default, these commands are executed by the Bourne shell, `/bin/sh`. A curious mechanism is used to inform the operating system of an alternative interpreter for commands within an executable file; if the first line of the file is of the form

```
#! program_name
```

then `program_name` will be used to interpret the commands instead of the Bourne shell. Since that line would have to be ignored under other circumstances, this technique will work for any program which uses the pound sign (#) as a comment delimiter. Such programs include all the UNIX shells (`/bin/sh`, `/bin/csh`, `/bin/ksh`, etc.) as well as the perl (usually `/usr/local/bin/perl`) text processing language.

The second type of file which is executable is a binary file created as the result of compiling a program in some low-level language like C or Fortran. The operating system recognizes such files by the so-called "magic number" which is a bit pattern inserted at the beginning of the file when it is first created, and which is different for each different type of file which is supported on the system. (See the man page for the `file` command or the `magic` file format for more information about magic numbers.) Since the compiler or loader deals with this detail, you will rarely have to worry about it. Executable text and binary files can be freely mixed in directories without causing any confusion.

If you write a command of your own, and you want