site and Tibshirani 1991), and loess models (Cleveland 1979; Cleveland et al. 1993) , among others. In the case of wafers, formal spatial analysis techniques can be used to estimate the extent of spatial clustering and its relation to the covariates (for instance, Taam and Hamada, 1992), but the above graphical displays are more visually effective.

In both examples the models used did not include terms for longitudinal and spatial effects because it was strongly felt that there was no prior knowledge of the interaction between the longitudinal/spatial structure and the design factors. By allowing the effects to freely vary over the fiber and over the wafer surface we guarded against most types of misspecification; yet the graphics we used to display the effects and coefficients effectively reveal their longitudinal and spatial dependence.

If the data are not collected through designed experiments, techniques such as principal components or hierarchical clustering may be appropriate. These are some of the multivariate techniques whose graphical displays can easily be augmented with similar symbols or glyphs, e.g., imagine a cluster dendrogram with wafers at the leaves.

4. Acknowledgments

The fiber study was joint work with Daryl Pregibon, who introduced me to the idea of "pasting" ANOVA models; the wafer case study was joint work with Daryl and Mark H. Hansen – I want to thank both very much.

5. References

Becker, R. A., Clark, L. A., and Lambert, D. (1994). Cave plots: A graphical technique for comparing time series. *Journal of Computational and Graphical Statistics*, 3(3):277–284.

Becker, R. A., Eick, S. G., and Wilks, A. R. (1991). Basics of network visualization. *IEEE Computer Graphics and Applications*, 11(3):12–14.

Cleveland, W. S. (1979). Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association*, 74:829–836.

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

Cleveland, W. S., Mallows, C. L., and McRae, J. E. (1993). ATS methods: Nonparametric regression for non-gaussian data. *Journal of the American Statistical Association*, 88(423):821–835.

Cliff, A. D. and Ord, J. K. (1981). *Spatial Processes: Models and Applications*. Pion Limited, London, UK.

Goldman, A. I. (1992). Eventcharts: Visualizing survival and other timed-events data. *The American Statistician*, 46(1):13–18.

Hastie, T. and Tibshirani, R. (1991). Varying-coefficient models. Technical memorandum, AT&T Bell Laboratories, Murray Hill, NJ.

Hastie, T. J. and Tibshirani, R. J. (1990). *Generalized Additive Models*. Chapman and Hall, London.

Taam, W. and Hamada, M. (1992). Detecting spatial effects from factorial experiments: An application from integrated-circuit manufacturing. *Technometrics*, 35:149–160.

Tufte, E. R. (1983). *The Visual Display of Quantitative Information*. Graphics Press, Cheshire, Connecticut.

Tufte, E. R. (1990). *Envisioning Information*. Graphics Press, Cheshire, Connecticut.

David A. James
Bell Laboratories
dj@bell-labs.com

TOPICS IN INFORMATION VISUALIZATION

Simplifying Visual Appearance by Sorting: An Example using 159 AVHRR Classes

By Daniel B. Carr and Anthony R. Olsen

1. The Visual Intimidation Factor

A presumed goal of tables and plots is to communicate to a target audience. Unfortunately, many tables and some plots appear visually intimidating, so fail as a communication device. In the spirit of Tufte (1983), who introduced concepts such as the lie factor and the data ink to total ink ratio, we define a concept called the visual intimidation factor (VIF). The VIF (rhymes with whiff) is the reciprocal of the time (measured in seconds) it takes to decide that the study of a table (plot) is not worth the effort. If the reader studies the table and derives useful information, the time is infinite and the VIF=0. One can't decide faster than a preattentive vision sweep of the table (about 50 milliseconds) so a theoretical upper bound is 1/.050=20. More realistically it may take a whole second to make a decision, so a VIF of 1 is representative of a bad table. Driving down the

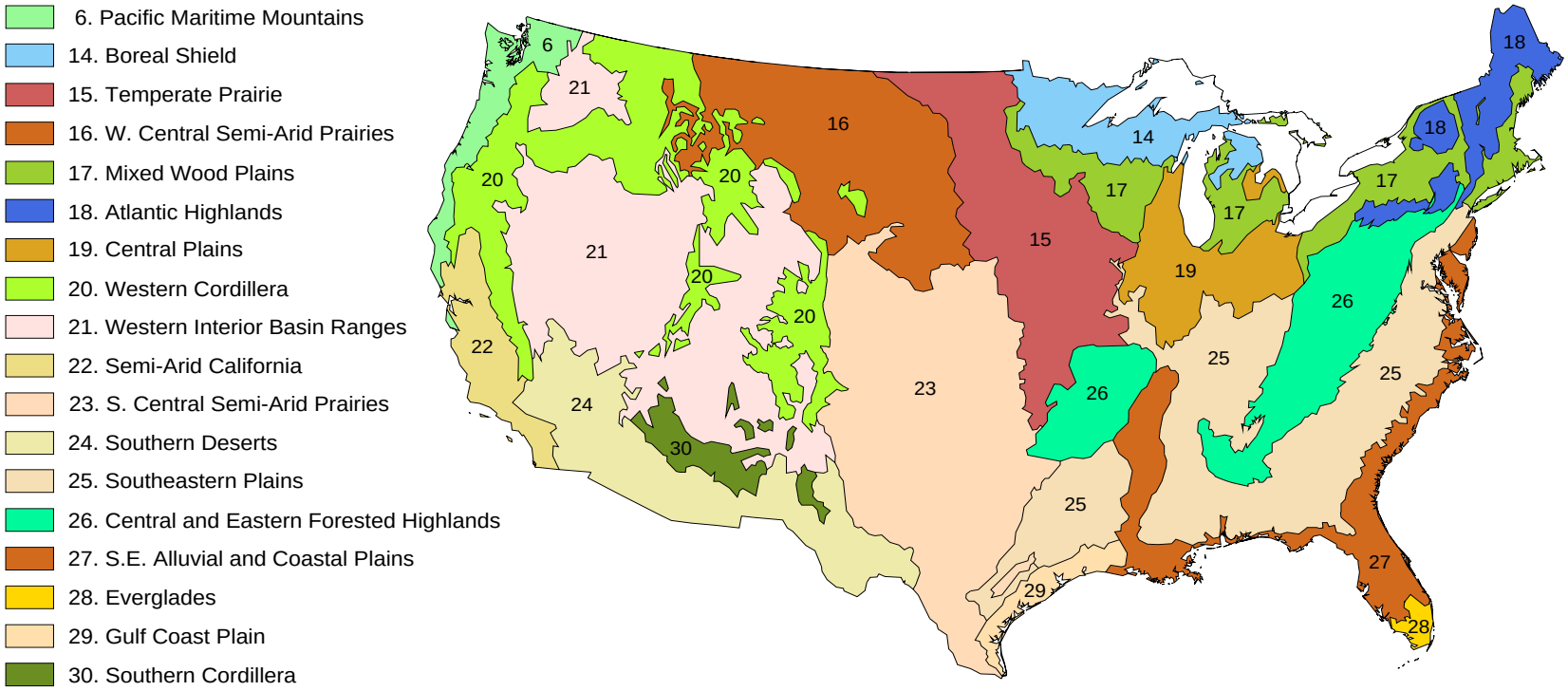

Figure 1. Level II Ecoregions of the Conterminous United States. (Omernik 1995).

VIF for complex tables and plots can be a challenge. Assuming an interested audience, Kosslyn's (1994) adage, "the spirit is willing but the mind is weak," is appropriate. This article focuses on a powerful tool for reducing the VIF, multivariate sorting.

2. Multivariate Sorting

Many statistical graphics writers are proponents of sorting. Cleveland (1985) describes research demonstrating that comparison accuracy increases with the nearness of the comparison items. Sorting brings similar items close together and they become easier to compare. Cleveland (1985, 1993) makes extensive use of sorting to bring out patterns in dot plots. Becker and Cleveland (1994) illustrate the advantage of sorting box plots by medians and Wainer (1993) discusses the advantage of sorting in tables.

While increasing the perceptual accuracy of extraction may have provided Cleveland's explicit motivation to sort data, an amazing consequence is that plots look much simpler. Carr (1994) describes this in terms of shortening the eye traversal path and reducing the number of visual focal points. Sorting boxplots by the median reduces the eye traversal path in moving from median to median and increases the apparent simplicity of the plot. Creating localized blocks in two-way layouts reduces the number of visual focal points and increases the apparent simplicity. Since we process visual information simultaneously on different scales (Marr 1982), our eyes can be drawn to many different places in a plot. An amazing plot reprinted in Marr (1982, page 50) contains patterns at different scales that emerge and disappear as one gazes at the plot. We conjecture that sorting often reduces the number of comparison scales, that limiting the number of comparison scales helps us focus at the same places in repeated viewing, and that stability in repeated viewing is a key to apparent simplicity. Whether or not our conjecture is correct, sorting simplifies.

3. Examples Using Ecoregions and AVHRR Classes

We put sorting to work to simplify the appearance of a two way layout. The challenge comes from the USEPA Western Ecology Division. The levels of the first factor are ecoregions for the conterminous U.S. Omernik (1995) constructs maps that partition North America into ecoregions, on the premise that ecological regions can be identified through the analysis of the patterns and composition of biotic and abiotic phenomena. The partitions integrate extensive knowledge of geology, physiography, soils, vegetation, climate, land use, wildlife, and hydrology. Although ecoregions are available at several scales, our interest is in Level II, which has 18 ecoregions within the conterminous U.S. (Figure 1). We want to communicate the commonalities and differences in biotic and abiotic characteristics across ecoregions.

We focus on a land cover characterization to illustrate how multivariate sorting can reduce the VIF for a large two-way layout. Loveland et al (1991) derives a land cover classification relying mainly on satellite imagery. Using AVHRR imagery spectral intensities for 1 km square pixels and additional information, they assign approximately 8 million pixels into one of 159 land classes. This classification is pixel resolution dependent and does not necessarily reflect the diversity within a pixel. For example, few pixels are classified as water, since few bodies of water dominate a full pixel. At other resolutions, acreage associated with the land classes would differ. The levels of the second factor in the layout are these 159 AVHRR classes. The dependent variable is acreage.

Figure 2 is a line height (thin bar) plot showing the class acreage as a percent of each ecoregion total acreage. For compactness, the plot omits the labels for the 159 classes. To provide labels one could make a larger plot or in an interactive setting handle the labeling by brushing, progressive disclosure or selective magnification. We conclude that Figure 2 has a high VIF. No spatial pattern appears sufficiently interesting to induce further examination. The plot communicates a message of spikes located "randomly" throughout the plot. The order for the row factor levels reflects the general north to south ecoregion numbering pattern. The order for the column factor levels reflects a hierarchical land cover classification scheme that is partially described below.

Figure 3 is a first cut to simplify the appearance of the plot via bi-directional multivariate sorting. While we have not dealt with the column labels and interpretation, the patterns now seem simple enough that maybe we can understand some of the relationships without too much work. In other words there may be a few concepts that characterized the acreage for the ecoregions. With some luck the concepts will mesh well with the existing top levels of the existing hierarchical classification.

Figure 3 illustrates just one of several viable approaches to multivariate sorting are available. Consider sorting rows. One approach is to obtain the median across all the AVHRR classes for each ecoregion and then to sort rows using the median. Cleveland (1993a) has found several examples in which collapsing to one dimension is effective. In environmental applications, the approach often fails because the median is often zero.

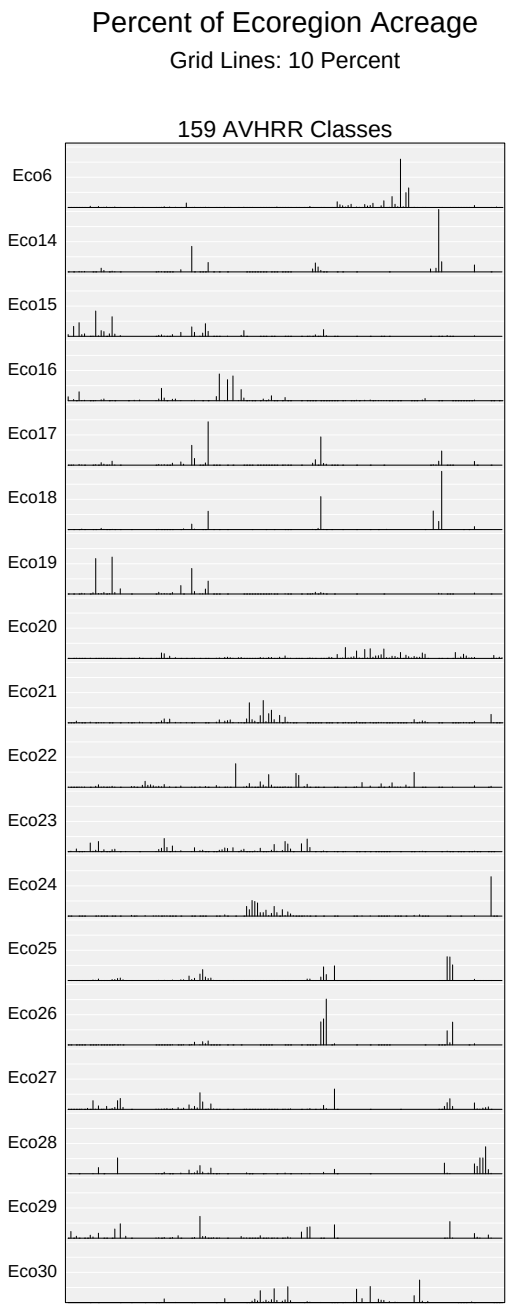

Figure 2:
Original Row and Column Order

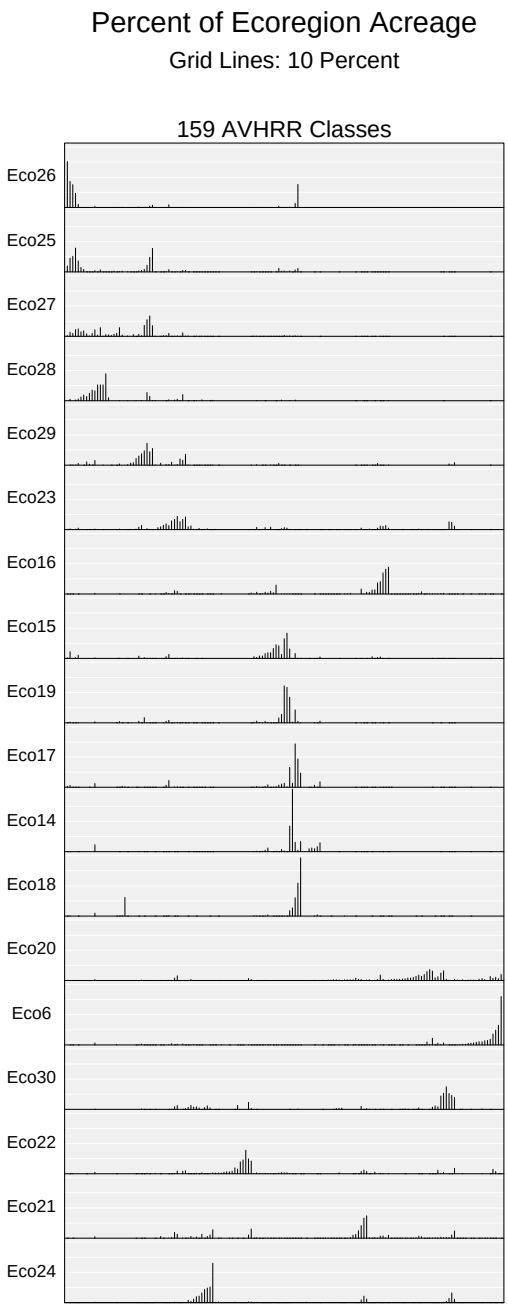

Figure 3:
Sorting of Rows and Columns

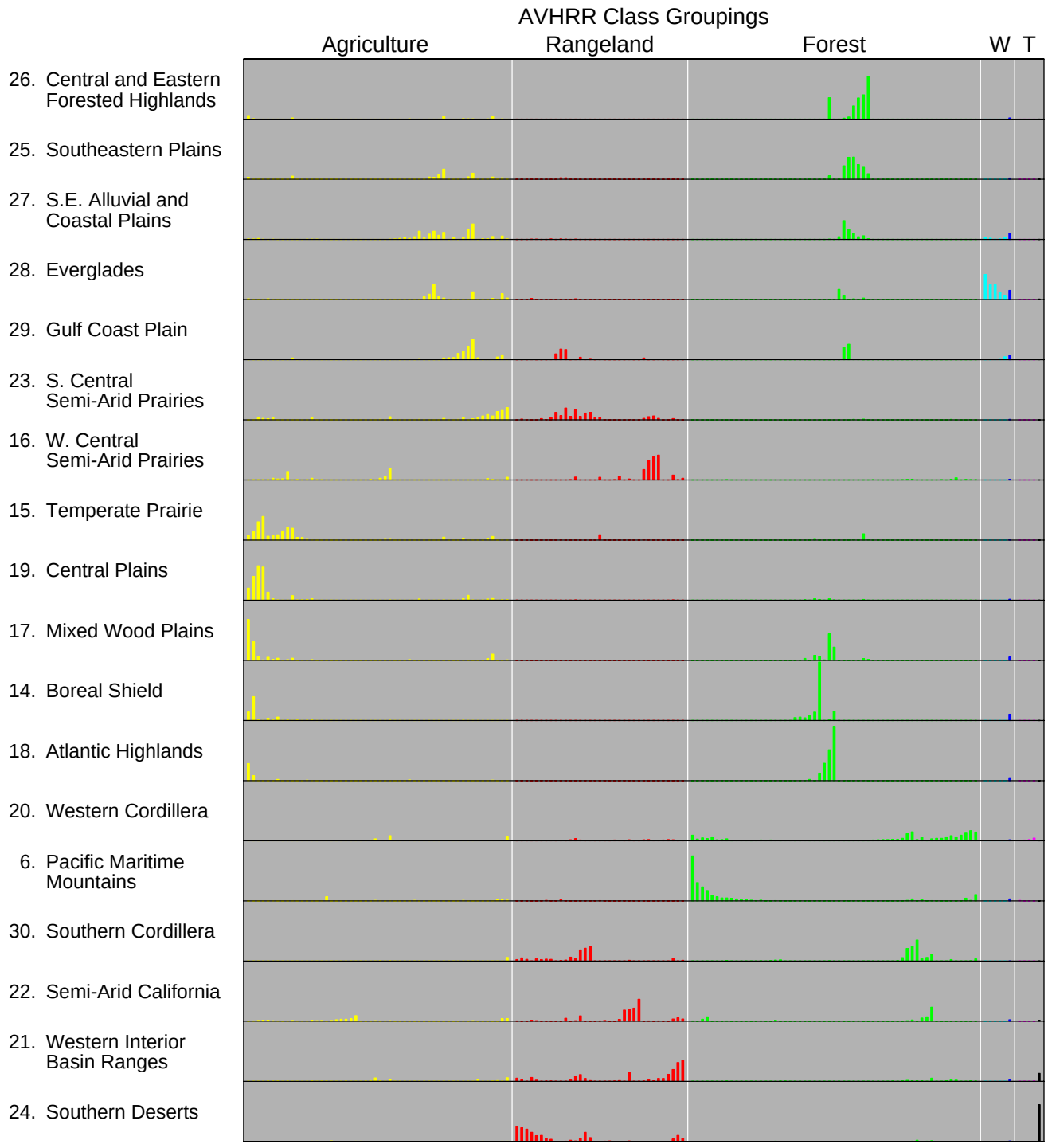

Figure 4: Sorting Classes Within Groupings
W: Wetlands and Water (Cyan and Blue)
T: Tundra and Barren (Magenta and Black)

The goal is to create a smaller number of perceptual groups or focal points. Figure 3 illustrates our use of Friedman and Rafsky's (1979) minimal-spanning-tree breadth-traversal algorithm to sort rows and columns. To sort rows, the algorithm starts by building a minimal spanning tree in 159 space. Then the algorithm establishes two nodes on the tree that have the largest traversal path in going from node to node. The breadth traversal algorithm starts at one of these two nodes and visits nearest subtrees until it eventually arrives at other node. The visiting of subtrees tends to provide effective visual groupings.

Two other approaches for separate row and column sorting are worth mentioning. One simple procedure is to sort rows (columns) by the first principal component scores. Another is to invoke a clustering algorithm such as a single-link algorithm, (see Banfield and Raferty 1992 for some modern clustering options), and to borrow the ordering from the ensuing dendrogram.

Separate row and column sorting is applicable to many crossed two-way layouts. Logical constraints may prohibit sorting both rows and columns, but any time a color matrix appears (from genetic algorithm population descriptions to protein descriptions) one should think about sorting. We don't know of research establishing a perceptually best sorting method. With today's computational power one can optimize over all permutations of rows and over all permutation of columns and iterate if necessary. It is easy to propose various clustering indices for optimization. Anything that puts low values together and high values together will likely help.

Another facet of making row (and column) labeled plots look simpler is to break the labels into groups. Kosslyn (1994) suggests that groups of size four or fewer are best for making within group comparisons. A list of length 12 is more visually intimidating than three groups of size four. In addition, creating smaller groups provides edges. The edges draw visual attention and when readers happen to notice a label of interest at an edge, they begin to get involved. This audience is probably not attuned to "home" ecoregions as it would be to a home state or county. Thus, the comment is not so important for this particular example. We do note in passing that more can be done with the rows in Figure 3. If such were available, a classification based on the labels provides one way of clustering rows. Another approach is to use a clustering algorithm as suggested above. For example, one can make clusters by cutting at the long links in the spanning tree traversal.

Several options provide a graphical representation of the clusters. The natural choice is to add space between clusters. When space is at a premium, we might try indenting every other cluster or alternating two background colors behind the labels. In an interactive setting, a mousing operation might reveal the whole dendrogram. Explicitly showing clusters is not always advantageous. A one-dimensional layout is not conducive to preserving among cluster distances, and the clusters themselves can be somewhat arbitrary. When the explicit clusters are not well supported by the rest of the graphic, the VIF may increase.

In this example, the column labels come with a hierarchical classification. There are seven classes at the lowest resolution, twenty-five at the second resolution, and 159 classes at the highest resolution. The seven classes are agriculture, rangeland, forest, water, wetland, barren and tundra. Figure 4 shows five groups, lumping water with wetland and barren with tundra. Colors indicate all seven groups, using yellow, red, green, cyan, blue, magenta and black, respectively. With grouping by position, the use of different hues in not required for this plot. (Gray and black bars can distinguish the combined groups.) However, the different hues help to strengthen the perceptual grouping and tend to add visual appeal, thus reducing the VIF. Representing as many as seven classes with different hues presses the limits for reducing the VIF. In this case the redundant encoding by position and the almost too-short-to-see magenta lines lessen the interpretation demands. Note that we have sorted the columns within each of the classes. One could also sort the placement of the five classes by the class totals.

Some patterns emerge in Figure 4. The Everglades have wetlands. The Pacific Maritime Mountains have a distinctive pattern of forest types. The Western Cordillera has great diversity. The Southern Deserts are substantially barren. This is partly consistent with what the reader already knows, but perhaps adds some new information. That's a good starting point. The omitted class labeling offers to provide additional information.

4. Variations and Extensions

We have looked at the next higher resolution classification involving 25 groups. One plot variation aggregated acreage into the 25 groups. With only 25 groups, labeling was easy. To save horizontal space and facilitate reading we started the labels over the columns and rotated them counterclockwise 40 degrees from horizontal. With the columns a bit wider than character height thickness, the thin bars became thick bars or else were widely separated. The visual effect was disappointing and the aggregation story was of questionable interest. While including column labels is an important key to

deeper interpretation, we have relegated the plot to electronic access.

Sorting, grouping and labeling are powerful tools. The primary drawback to sorting rows is that people often look up values by their labels (Cleveland 1993b). When labels are not in an alphabetic or another familiar order, then the look-up task becomes complicated. Linking items to a new ordering or back to a map helps people to put the information together. Different hues are only an effective link for a few items. For many items, other approaches are better. Beyond providing written grid coordinates, visual linking methods include marked microplots of 90 degree rotated row-labels, and marked postage stamp maps (see Eddy and Mockus 1995 for a discussion of stamp-sized images). However, linking is a topic for another paper.

5. Access and Comments

Data, Splus functions and script files producing the current examples are available via anonymous ftp to `galaxy.gmu.edu`. Change directory to `submissions/newsletter/sorting`. Examples from other newsletter articles are now stored under this newsletter directory. As always I (Dan) welcome constructive comments plus new graphics challenges.

Acknowledgments and Disclaimer:

This work is part of collaborative research with the USEPA Western Ecology Division and involves Sue Pierson and Pip Courbois. We thank them for their contributions of data and ideas. We invite the readers to see our collective results involving digital terrain data and other variables at the ASA annual meeting in Chicago.

Although the research described in this article has been funded in part by the U.S. Environmental Protection Agency through cooperative agreement CR820820 to George Mason University, it has not been subjected to Agency review. Therefore, it does not necessarily reflect the views of the Agency.

References

Banfield, R. D. and Raferty, A. E. (1992). Model-based Gaussian and non-Gaussian clustering. *Biometrics*, 49:803-822.

Becker, R. A. and Cleveland, W. S. (1993). Discussion of Graphical Comparisons of Several Linked Aspects by John W. Tukey. *Journal of Computational and Graphical Statistics*, 2(1):41-48.

Carr, D. B. (1994). Converting Tables to Plots. Technical Report 101, Center of Computational Statistics, George Mason University, Fairfax VA, 22030.

Cleveland, W. S. (1985). *The Elements of Graphing Data.* Wadsworth Advanced Books & Software, Monterey, CA.

Cleveland, W. S. (1993a). *Visualizing Data.* Hobart Press, Summit, NJ.

Cleveland, W. S. (1993b). A Model for Studying Display Methods of Statistical Graphics. *Journal of Computational and Graphical Statistics*, 2(4):323-343.

Eddy, W. F. and A. Mockus (1996). An Interactive Icon Index: Images of the Outer Planets. *Journal of Computational and Graphical Statistics*, 5(1):100-111.

Friedman, J. H. and L. C. Rafsky (1979). Multivariate Generalizations of the Wald-Wolfowitz and Smirnov Two-Sample Tests. *The Annual of Statistics*, 7(4):697-717.

Kosslyn, S. M. (1994). *Elements of Graph Design.* W. H. Freeman and Company, New York, NY.

Loveland, T. R., Merchant, J. W., Reed, B. C., Brown, J. F., Ohlen, D. O., Olson, P., and Hutchinson, J. (1995). Seasonal land cover regions of the United States. *Ann. Assoc. Amer. Geog.*, 85:339-355.

Marr, D. (1982). *Vision.* W. H. Freeman and Company, New York, NY.

Omernik, J. M. (1995). Ecoregions: A framework for managing ecosystems. *The George Wright Forum*, 12:35-51.

Tufte, E. (1983). *The Visual Display of Quantitative Information.* Graphics Press, Cheshire, CT.

Wainer, H. (1993). Tabular Presentation. *Chance*, 6:52-56.

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

Tony Olsen
US EPA
National Health and Environmental Effects Research Laboratory
`tolsen@mail.cor.epa.gov`