Boxplot Variations in a Spatial Context: An Omernik Ecoregion and Weather Example

By Daniel B. Carr, Anthony R. Olsen,
Suzanne M. Pierson, Jean-Yves P. Courbois

1. Introduction

This article presents three variations on boxplots for use in a spatial context. The three variations are linked micromap boxplots, linked micromap bivariate boxplots, and angular boxplot glyphs. The specific examples show Omernik Level II Ecoregions. The variables chosen for illustration purposes are growing degree days and precipitation.

This article is closely related to papers by Carr et al (1998a, 1998b). The first paper provides a general description of linked micromap (LM) plots such as that in Figure 1b. The second paper puts LM plots to work in describing Omernik Level II ecoregions. It also promotes LM plots as useful methodology in the KDD pattern discover process and as overviews that provide an orienting context for drilling down to finer detail in extensive hierarchically organized summaries. This relatively brief article adapts graphics and text from the second paper while focusing attention on boxplots.

In this paper Section 2 provides context for the graphics with a description of Omernik ecoregions and the data sets. Section 3 provides background on boxplots and then presents both univariate and bivariate LM boxplots. Section 4 motivates two angular boxplot glyphs and presents the one that is more striking, the Portuguese man of war. Section 5 closes with connections to other work and challenges for the future.

2. Ecoregions and Datasets

While this paper focuses on graphics design, a scientific context exists behind the examples: the presentation of summary statistics for ecoregions and the attempt to learn about the intellectual structure that underlies ecoregion definition. In this section we provide background on ecoregions and associated datasets.

2.1 Ecoregions

Ecoregions are a way of codifying the recognizable regions within which we observe particular patterns or mosaics in ecosystems. The general consensus is that such ecological regions, or ecoregions exist. However, disagreement in how the regions should be constructed continues to be a scientific issue (Omernik, 1995). In describing ecoregionalization in Canada, Wiken (1986) stated: "Ecological land classification is a process of delineating and classifying ecologically distinctive areas of the earth's surface. Each area can be viewed as a discrete system which has resulted from the mesh and interplay of the geologic, landform, soil, vegetative, climatic, wildlife, water and human factors which may be present. The dominance of any one or a number of these factors varies with the given ecological land unit. This holistic approach to land classification can be applied incrementally on a scale-related basis from very site-specific ecosystems to very broad ecosystems."

Within the United States two alternative approaches in the construction of ecoregions are those developed by Omernik (1987, 1995) and Bailey (1995a, 1995b, 1998). Each constructs a hierarchy of ecoregions that corresponds to viewing the United States at different scales. Omernik's approach is conceptually similar to that described by Wiken, where the ecological regions gain their identity through spatial differences in the combinations of the defining factors and which factors are important vary from one place to another and at all scales. Bailey (1998) develops ecological regions hierarchically. First, he identifies ecological regions of continental scale based on macroclimate, where macroclimates influence soil formation, help shape surface topography, and affect the suitability of human habitation. The continent is subdivided with three levels of detail into domains, within domain divisions, and within division provinces. Domains and divisions are based largely on broad ecological climatic zones while provinces further sub-divide the divisions on the basis of macro features of the vegetation. Hence Bailey uses macroclimate as the controlling factor in the formation of ecoregions while Omernik uses all available factors where the importance of the factors varies among ecoregions. Some scientists question whether enough is known to delineate ecoregions. While our knowledge is limited, others proceed on the basis that our approximations of the "true" ecoregions will continue to improve as more information is gathered.

Our interest is not to define ecoregions or even to validate them. We simply believe that it is important to describe quantitatively some of the key characteristics associated with ecoregions to gain a better understanding of how ecoregions have partitioned these characteristics. In view of the newsletter page size, we show Omernik's level II ecoregions (Figure 1a) for the

Figure 1. a=Omernik Ecoregions b=Linked Micromap Boxplots

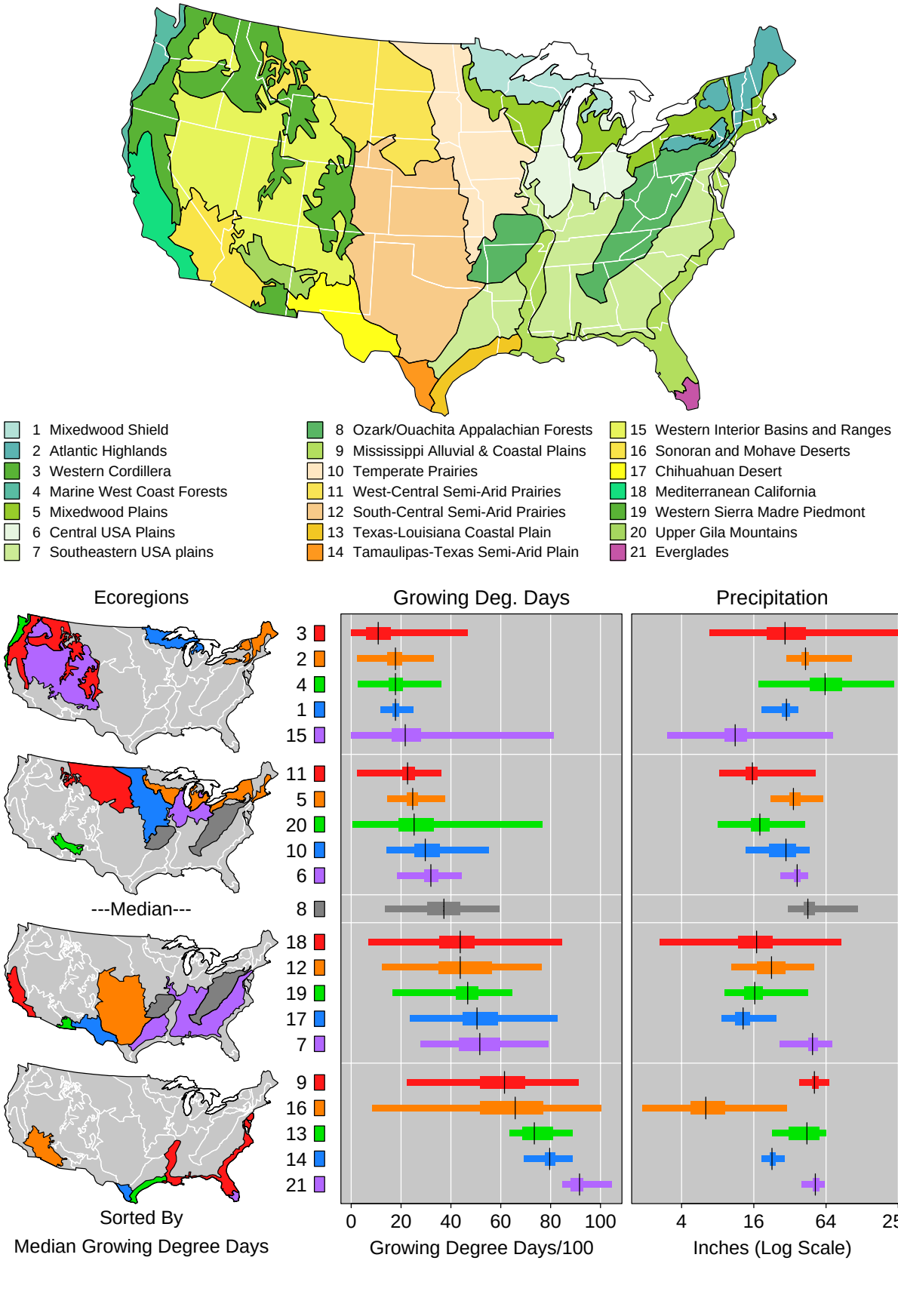
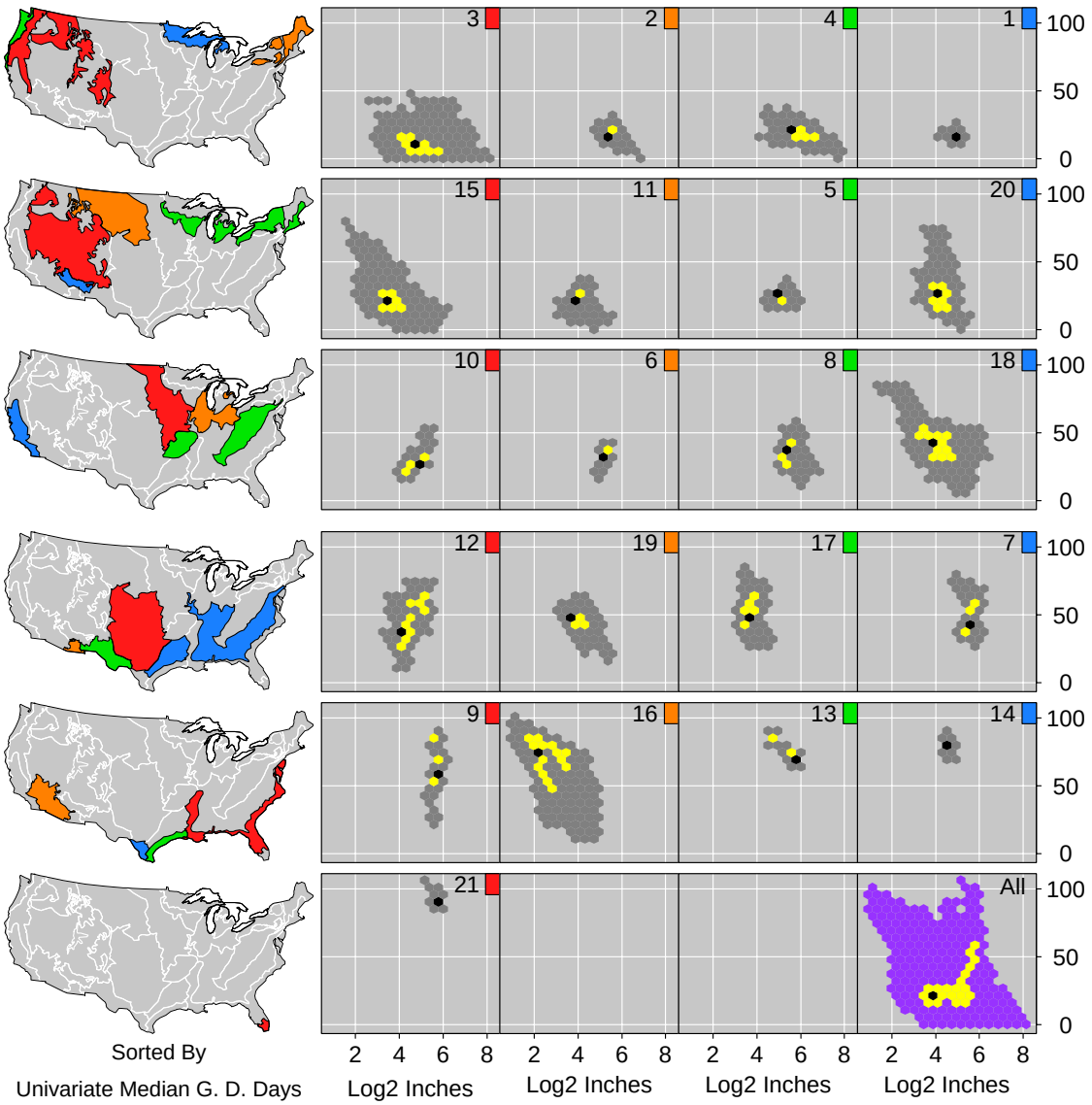


Figure 2: LM Bivariate Boxplots

1961-1990 Precipitation (x) versus Growing Degree Days/100 (y)



continental U.S. examples here. This coarse level of aggregation bothers some people, for example when they see the Mississippi Valley in the same ecoregion with the Southeastern seacoast and Cape Cod. Level III and level IV ecoregions are progressively finer partitions of North America and are much more homogeneous in terms of familiar variables.

2.2 Data Sets

The data sets used in our graphics are readily available. Substantial thought and processing was involved in producing the data sets. Thus this paper builds upon the work of others.

Ecoregion delineation depends on climate as a major forcing function. We use nationally consistent climate data sets developed using PRISM (Parameter-elevation Regressions on Independent Slopes Model). PRISM, described by Daly et al (1994), is an analytical model that uses point data and a digital elevation model (DEM) to generate gridded estimates of monthly and annual climatic parameters. PRISM models data from individual climate stations to a regular grid through locally weighted precipitation/elevation regression functions. Orographic effects on precipitation and other climate parameters are well known and are used to define the local regions. PRISM has been shown to out perform other common spatial interpolation procedures such as

kriging. The regular grid used is 2.5 minute by 2.5 minute latitude/longitude grid, i.e., a nominal 5 km by 5 km grid. For our purposes having a regular grid of climate data ensures that the entire area of an ecoregion is represented in any statistical summary we construct. PRISM data we use are based on 30-year normal climate data for 1961-1990. Key annual parameters investigated are mean temperature, temperature range, growing degree days, and total precipitation, and monthly precipitation. Further information on PRISM and the data sets can be found at `www.ocs.orst.edu/prism/`.

Figures 1 and 2 focus attention on just two PRISM variables, growing degree days and precipitation. The delineation of ecoregions depends not only on the annual characteristics of climate but also on seasonal characteristics. Thus monthly precipitation is important to study. As indicated above many variables such as elevation, land cover, and soils also relate to ecoregions. A dated visual summary of level II ecoregion land cover classes is available in Carr and Olsen 1996. (A few boundaries have changed.)

Figure 3 (page 10) in this paper concerns precipitation trends. The figure makes direct use of annual weather station precipitation data for the period 1961-1996. The precipitation data (72 Mbytes) and weather station location descriptors (10.2Mbytes) are available from the National Climatic Data Center Web site `www.ncdc.noaa.gov/ol/climate/onlye/coop-precip.html#Files`.

Our data processing classified weather stations into ecoregions. We used each station's most recent latitude and longitude (some change over time) to locate it in one of 672 polygons that define the Level II ecoregions in the continental U.S. We omitted stations that had fewer than 24 annual values over the 36 years. Some 4075 weather stations satisfied the criteria of being in an ecoregion and having sufficient annual values. We then calculated trend using Sen's slope, which is the median of slopes computed using all pairs of available years for each station. This provides an estimate that is little influenced by an occasional extreme value. The angular boxplot glyphs in Figure 3 represent variation in 36-year trends at different weather stations with the 21 Level II ecoregions.

3. LM Boxplots

Figure 1b shows linked micromap univariate boxplots. We briefly discuss the basic LM plots elements: micromaps, linking labels, boxplots, and sorting. For more LM plot details see the previously mentioned papers.

The micromaps on the left are caricatures of level II ecoregions. The caricatures simplify the polygon boundaries and reduce the 672 polygons for the continental U.S. to 44 polygons. Arguments can be more for going even further and, for example, removing Cape Cod and the island in Lake Superior.

Omernik's Level II definition of ecoregions partitions the continental U.S. into 21 ecoregions. The long labels for the ecoregions appear in Figure 1a. Linking ecoregion name to location is difficult in Figure 1a. There are many ecoregions and some involve disjoint areas. The colors may be pleasant to view, but are not easy to discriminate. For most readers the colors do not have names, and this can make it much harder to remember the exact color that is to be found on the map. Plotting the ecoregion numbers on the map typically solves the memory problem, but serial search is still involved. With disjoint areas for the same ecoregion, finding one number does not mean the search task has been completed. The symbol congestion and search problem gets worse with Level III ecoregions.

Micromaps with color links provides a solution to the name and region linking problem. To find the polygons for Ecoregion 3 one simply looks for the red regions in the top micromap. The only difficulty concerns micromap caricatures with polygons so small that they do not reveal their color. The state map in Carr and Pierson 1996 enlarge small states to make sure their color is visible. In the current map caricature, Cape Cod (Ecoregion 9) is too small to reveal its color. The development of micromap caricatures involves compromise.

The use of integers as ecoregion labels is another compromise. Descriptive labels always have merit. However, in this and other examples involving more statistical summary columns, we choose to reserve space for the statistical graphics panels. The integer labels take up little space and become familiar with repeated usage.

The color rectangle as a part of the label can be dropped in Figure 1b because the boxplots provide the color. For those experienced with LM plots, dropping the line of colored rectangles simplifies appearance and complicates color linking only slightly. We retain the rectangles in the current context.

The boxplot or box and whiskers plot is a well-known distributional caricature that has now appeared in some grade school curriculum. However, the choice of summary statistics and the graphical representation are not universal (see McGill, Tukey and Larsen 1978; Frigge, Hoaglin, and Iglewicz 1989; Tukey 1993; and Carr 1994). Figure 1b shows a five number summary: the extrema, the 1st and 3rd quartiles and the median. An-

other common choice represents adjacent values and outliers rather than just extrema. We note that the outlier portion of this choice does not scale well for large samples from thick tailed distributions. For example if estimates followed a t-distribution with 3 degrees of freedom roughly 2.75 percent of the estimates would be flagged as outliers on each end of the distribution. Outlier overplotting was problematic for precipitation in an alternative version of Figure 1b. When outlier overplotting is problematic a compromise caricature that we do not show uses adjacent values and adds an outlier dot for a maximum or a minimum that is not an adjacent value. The compromise uses a filled dot to indicate multiple outliers thus hiding the details of outlier multiplicity and location.

Our graphical representation follows the design of Carr (1994). That is, the thick rectangle extends from the 1st to the 3rd quartile, the thin rectangle extends to the extrema, and the vertical median line extends outside the thick rectangle. Comparison of adjacent medians can typically be based on the judgement of horizontal distance between line endpoints. Using dots for the median (Becker and Cleveland 1993) is a viable choice since the area symbol is easy to spot even though enclosed in a rectangle. However, the comparison of neighboring medians may not be as accurate since the direct distance between points is not the correct distance to judge.

The boxplot is suitable for describing a batch of data. The batch of data here reflects variation due to changing spatial location. However, one would like the items in a batch to be as comparable as possible. This is not the case here because the cells are defined using equal angle grids and not equal area grids. Carr et al (1997) question the use of equal angle grids in environmental sciences and in NASA's level 3 satellite products. They promote the use of discrete global equal area grids. While transformations can be made so the North Pole is not as wide as the equator, changing to equal area grids for reasons of comparability or polar modeling adds complexity and uncertainty. We choose to avoid the complexity of regridding the data for this article on boxplots. A consequence is that in large north to south ecoregions, such as Ecoregion 9, the northern portion is over represented in the boxplot due to smaller area grid cells. If the North is cooler than the South the growing degree day percentiles will be shifted a bit toward cooler values.

The notion of growing degree days many not be familiar, but the calculation is straight forward. If the average daily temperature is over 50 degrees Fahrenheit, the day counts. The degrees accumulated for each day that counts is the average daily temperature minus 50 degrees. As a calibration case suppose a grid cell in Florida had a daily average value of 75 degrees, each day for the whole year. The growing degrees days would then be 365 * (75-50) = 9125. To provide easier to remember two digit numbers, the scale for Figure 1b shows growing degree days divided by 100.

The precipitation data for Figure 1b has a thick right tail. A logarithmic transformation helps to pull in the tail, and provides better resolution for small precipitation values. The choice here was to use log base 2, since powers of 2 are familiar to many scientists (see Cleveland 1985).

Sorting helps to bring out patterns in LM plots. Carr and Olsen (1996) discuss sorting methods and show ecoregion examples. In Figure 1b sorting by increasing growing degree days arranges the ecoregions in the micromap panel sequence so that northern ecoregions tend to appear in the top panels and southern ecoregions appear in the bottom panels. This matches common expectation. Some of the north to south anomalies, such as Ecoregion 20, are easier to understand when elevation boxplots are included. While pattern details may be well-known to experts, those new to the field may delight in seeing patterns that confirm their limited knowledge and in finding anomalies. As an interpretation reminder, the patterns in growing degree days refers to the thirty year average and not to a particular day or year.

Sorting Figure 1b by increasing precipitation reveals a strong West to East pattern. With the exception of Ecoregion 4 that contains the Pacific Northwest rain forest, the high precipitation is in the East. A plot is available on our Web site.

The juxtaposed plots make it tempting to look at the growing degree day and precipitation medians to see if there is a relationship. The first impression is that if there is a relationship, it is weak. Of course juxtaposed univariate plots, such as bar plots, dot plots or boxplots, provide a poor way to look for a functional relationship even if one variable is sorted. While perhaps unknown to popular press, the scatterplot provides the standard for assessing a functional relation between two variables. We omit this scatterplot and proceed to represent the bivariate data for all grid cells and not just the univariate medians.

The current application with 481,475 grid cells warrants the use of density estimation as part of the graphic representation process. Scott (1992) provides methods for both density estimation and graphical representions that emerged from years of research with binning and other density estimation methods. Proponents of binning for dealing with large data sets also include Carr et

al (1987). Their space-time example included point in polygon selection on a map to defined two subsets, representation of density differences in a scatterplot matrix, and a representation of a temporal mismatch. Bivariate binning methods scale well in the number of cases. Carr (1998) shows a scatterplot matrix representing over billion point pairs. Such density estimates can vary over many orders of magnitude. Appreciating the density surface details is not necessarily a trivial task. For an overview we seek something far simpler and turn to a bivariate boxplot.

Chances are that many people have proposed bivariate boxplots. So far bivariate boxplots have not seemed to catch on, likely due to lack of promotion. Possibly disagreement on details is behind this lack of promotion. Certainly there are issues concerning what generalization is most appropriate for the 1st, 2nd, and 3rd quartiles. (See Small [1989] concerning bivariate median options that were available years ago.) Perhaps it is better to show density modes and two alpha contours. Our attitude is to get on with the graphics. We illustrate the approach of Carr (1991) but are quite willing to substitute other methods.

Carr (1991) binned the data using hexagon grids to speed the processing. He directly used the high density cells containing 50 percent of the data in place of the interquartile range. To obtain a median he smoothed the high density cells (an optional step), and then eroded the cells using gray level erosion. The gray level erosion process removes counts from the cells proportional to exposed surface area. The last cell eroded provides location of the median in the binned density representation. (S-PLUS™ now provides the needed tools for hexagon grids.) This simple procedure generalizes to other regular grids in two and higher dimensions. With translucence or the see through contour representations in Scott (1992), showing 3-D boxplots is not a problem.

If one were to use the 50% high density region in univariate boxplots rather than the interquartile range, comparison becomes harder because the high density region can be disjoint. Carr (1991) addresses the increased difficulty of comparison in the bivariate context by developing a difference plot. Cells in two gray levels distinguish the two mismatched portions of the high density region. An arrow shows the change in the median. The example interleaves difference plots between bivariate boxplots in an age and sex conditioned two-way layout of bivariate boxplots. With more space we could interleave the bivariate difference plots in the linear sequence shown in Figure 2.

Figure 2 shows a bivariate boxplot of the combined data in the lower right corner. The purple shows the support of the bivariate pairs, (30-year average precipitation on a log scale, 30-year average growing degree days). The yellow cells are the 50% high density cells and the black cell is the median cell. The positive trend of the high density cells contrasts strongly to the negative slope suggested by the region of support. The latter is all one would see in an overplotted view. How useful it is to study this aggregate plot can be argued. Scientists tend to be interested in detail. When scientists see a plot splitting into non-homogeneous pieces as in Figure 2 they naturally want to study the pieces. Nonetheless overviews are often valuable for providing a frame of reference as one approaches the pieces.

We mention several patterns in Figure 2, but the existence of Level IV ecoregions suggests that scientists would prefer to focus attention on more highly stratified ecoregion views. Already suggested was the fact the variation of individual ecoregions is generally much smaller than that of the composite. The small amount of yellow in the Ecoregion plots indicates a high concentration of bivariate values within tight domains. Several boxplots suggest that they are mixtures of tight patterns. Few bear much resemblance to a bivariate normal distribution. A moderately elliptical exception, Ecoregion 10, shows no yellow so the cell with yellow overplotting black median dot contains over 50 percent of the observations. Panels 15, 18, and 16 (top to bottom order) catch the eye with relatively big bivariate support regions (gray) that have negative slopes. The high density yellow cells reflect the negative slopes as well. The high density cells in Region 16 show a bifurcation that motivates further investigation and suggests possible subdivision of the ecoregion. (Carr et al 1998b note that Region 16 is homogeneous in terms of land cover). Regions 10 and 9 show high positive slopes. This suggests that most of the variation is in terms of growing degree days and that growing degree days is not major factor in their definition. Note that both regions cover a large range of latitude. In general the growing degree day variation in the bivariate boxplot appears associated with latitude variation in the micromaps. This motivates putting growing degree days on the y axis.

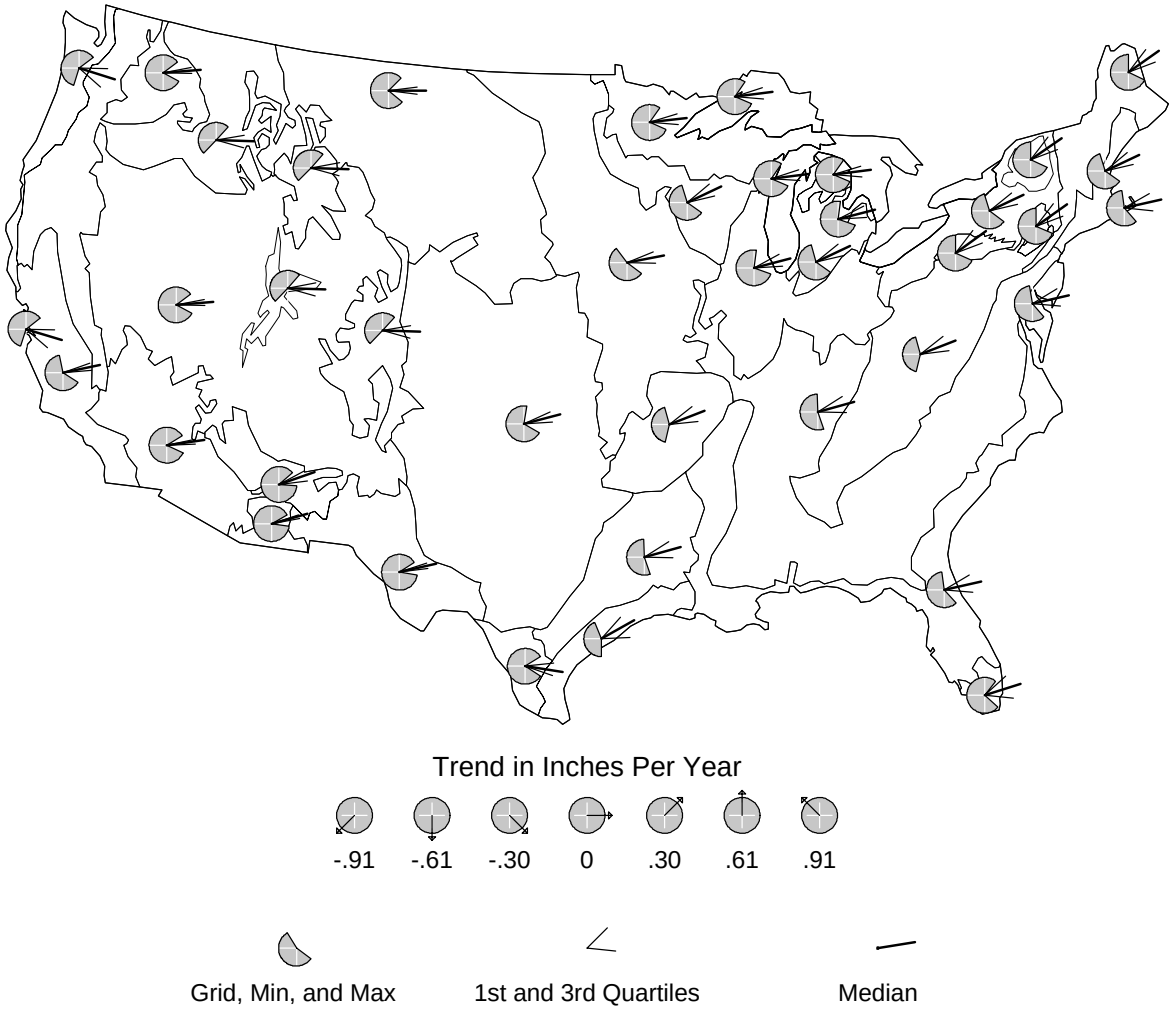
Figure 2 uses the univariate growing degree days as the basis for sorting, but the micromap layout is different than in Figure 1. Due to the size of bivariate boxplots, Figure 2 shows four or fewer ecoregions per micromap. The vertical list of names and color links is absent since they are mostly redundant with the names and color tags in each bivariate boxplot panel. The list would clarify the bivariate boxplot order that is left to right, top to bottom. Careful comparison of bivariate median values

for growing degree day values against the reference grid indicates the bivariate and univariate ranking are in reasonable agreement. Other sorting can bring out other patterns.

4. Angular Glyph Boxplots and a Precipitation Trend Example

Carr and Pierson (1996) favored LM plots over choropleth maps in terms of representing statistical summaries more faithfully. Their reasons include better perceptual accuracy of extraction and easy visual representation of uncertainty. However, classed choropleth maps are not the only way to represent spatially indexed statistical summaries. Glyph plots provide a viable alternative for representing summaries so they deserve consideration. Carr et al (1998b) propose two angular glyph boxplots but illustrate only one. Figure 3 shows the more controversial glyph that we call the Portuguese man of war. The glyph extends the trend glyph design of Carr, Olsen, and White (1992). In their trend glyph design, ray angle represents the sulfate deposition trend and the straight edges of a circle sector represent 90 percent confidence bounds. A local scale, a circle with lines at regular angles, provides the basis for judging ray angles. We continually attempt to put Cleveland and McGills (1984) guidance concerning perceptual accuracy of extraction into practice. They promote representations that use position along a scale as being the best. Our encoding uses position along a scale, albeit an angular scale.

Figure 3: 1961-1996 Precipitation Trends
Angular Boxplots For Stations Within Ecoregions



Then boxplot glyph shown in Carr et al (1998b) uses the circle sector to represent the interval from first to third quartile. Two shorter length rays (but longer than the reference circle) encode the extrema. The Portuguese man of war glyph shortens the extrema rays to the length of the reference circle and cuts away the circle sector between extrema rays. As can be seen in Figure 3, the glyph also removes the arc between lines showing the first and third quartiles. The gestalt is striking. Some symbols are aligned and some twist. The departures from symmetry about the horizontal zero trend line and about the median line are clear. The gestalt is so striking that it may impede people looking at the detail. Also cutting away part of the reference circle makes it a bit harder to judge angles. Thus this Portuguese man of war glyph may lose in careful testing for perceptual accuracy of extraction.

The angular limits for the boxplot glyph in Figure 3 are from 135 degrees to 135 degrees. There is some advantage in restricting the limits to 90 degrees so, for example, the largest local y coordinate indicates the largest increase. The extension here is to increase angular resolution. The legend scale has limits of -.91 and .91 inches per year. This implies a change of roughly 32 inches over the 36 years. The scale is defined symmetrically so both extremes are not necessarily achieved. However, a study of the glyphs indicate that this is nearly so. Time series plots for individual stations confirm such substantial change.

The existence of multiple polygons for the same ecoregion complicates examination of Figure 3. The same glyph appears for each polygon that comprises a given ecoregion. This gives more visual weight to ecoregions with multiple polygons. If unequal visual weight is given, a more reasonable choice is to favor large area ecoregions. The glyph design itself could be modified to indicate the number of weather stations with each ecoregion.

Glyph plots have some problems. We have slightly revised the GIS-provided representative point for each polygon to eliminate glyph overplotting in Figure 3. Symbol congestion control becomes increasingly problematic as the number of polygons increase, for example with level III ecoregions. (The trend glyph plot in Carr, Olsen and White [1992] was restricted to a regular hexagon grid to control symbol congestion.) The glyphs provide a spatial overview, but in this gray level example do not connect strongly to ecoregions and their names. LM plots provide a strong summary to name connection and avoid multiple glyphs, placement and symbol congestion problems. Also the straight line scale boxplot is a bit easier to assess than an angular scale boxplot. Still, the gestalt spatial impression of glyph plots has value.

5. Connections and Future Challenges

The article builds upon previous work by many people. Carr et al (1998b) cite many that inspired our development of LM plots. Our shortened list here includes Cleveland and McGill (1984), Cleveland (1985), Monmonier (1988, 1993), Tufte (1990, 1993) and Kosslyn (1994). Our scholarship in regard to bivariate boxplots is at best dated. We welcome references to the work of others and constructive comments about better designs.

The Splus script files we used to produce the graphics are available by anonymous ftp to `galaxy.gmu.edu`. The subdirectory is `pub/dcarr/newsletter/boxplots`. The contexts include the statistical summaries used in the graphics but not the prior data. For more information on the data or on algorithms such as gray level erosion in 3-D and 4-D please contact us directly. There is also a Web site, `www.galaxy.gmu.edu/~dcarr/graphgall/ecoregions/ecoregions.html` that contains a growing number of Omernik level II and level III ecoregion examples.

We want to make a point relative to our first presentation (Olsen et al 1996) of Omernik ecoregions. The 4 x 8 foot poster example provided a beginning description of Omernik's Level III ecoregions. This poster included over one hundred micromaps, boxplots of ecoregion digital elevations and even Loveland's 8 million pixel map of 159 land cover classes for the continental U.S. (See Loveland et al 1995). While large high-resolution graphics that integrate different views would seem to provide a powerful tool in the attempt to understanding large data sets, such graphics fall outside the mainstream. Today's emphasis is on Web graphics. Tufte (`www.clbooks.com/nbb/tufte.html`) comments that, "the problem with the Web is that it is low resolution in both space and low in time." As one way to appreciate this statement, note that a typical workstation monitor shows only 1024 * 1280 pixels (roughly 1.3 million pixels). Thus without reducing resolution, Loveland's map requires over six complete screens. Something is lost in comparison to high quality printed maps. There is no denying the power of human computer interface methodology such as logical pan and zoom, but the challenge of large data sets and complex conceptual structures motivates use of all powerful tools including good old human pan and zoom of large high quality graphics.

The task of developing and presenting statistical overviews for massive data sets is a big challenge. The environmental science community has worked on the task longer than many disciplines. How does one represent environmental facts and knowledge? Two recent book length overviews, Stone et al (1997) and Wahlstrom, Hllanaro, and Maninen (1996), involved the work of many people, including in one case layout specialists and graphics designers. The integration of photographs, text, and statistical graphics serves as an inspiration. Our challenge is to learn when others have blazed the trail, to make the improvements when we see the opportunity and to adapt the methods to other applications. Perhaps an even harder challenge is to see the patterns in important data not collected and take action.

6. Acknowledgements

S-PLUS is a registered trademark.

EPA funded the majority of the work behind this paper under cooperative agreements No. CR8280820-01-0 and No. CR825564-01-0. Additional federal agencies, BLS and NCHS, supported some facets of this work. The article has not been subject to review by BLS or NCHS so does not necessarily reflect the view of the agencies, and no official endorsement should be inferred. It has been subjected to EPA's peer and administrative review and approved for publication. The conclusions and opinions are solely those of the authors and are not necessarily the views of the Agencies.

7. References

Bailey, R.G. (1995a) "Description of the ecoregions of the United States," Misc. Publ. No. 1391 (rev.), USDA Forest Service.

Bailey, R.G. (1995b) *Ecosystem Geography*, Springer-Verlag, New York.

Bailey, R.G. (1998) "Ecoregions map of North America: Explanatory Note," Misc. Publ. 1548, USDA Forest Service.

Becker , R.A. and Cleveland, W.S. (1993) "Discussion of Grapical Comparison of Several Linked Aspects: Alternative and Suggested Principles," JCGS, Vol. 2, No. 1, 41-48.

Carr, D.B. (1991) "Looking at Large Data Sets Using Binned Data Plots," in *Computing and Graphics in Statistics*, eds. Buja, A. and Tukey, P., 7-39.

Carr, D.B. (1994) "A Colorful Variation on Boxplots," *SC& G Newsletter*, Vol. 5, No. 3, 19-23.

Carr, D.B. (1998) "Multivariate Graphics," *Encyclo. of Biostat.*, eds. Armitage, P. and Colton, T., Vol. 4, 2864-2886.

Carr, D.B. and Olsen, A.R. (1996) "Simplifying Visual Appearance By Sorting: An Example Using 159 AVHRR Classes," *SC& G Newsletter*, Vol. 7, No. 1, 10-16.

Carr, D.B., Olsen, A.R. and White, D. (1992) "Hexagon Mosaic Maps for Display of Univariate and Bivariate Geographical Data," Cartography and Geographic Information Systems, Vol. 19, No. 4, 228-236,271.

Carr, D.B., Olsen, A.R., Courbois, J.P., Pierson, S.M., and Carr, D.A. (1998a) "Linked Micromap Plots: Named and Described," *SC& G Newsletter*, Vol. 9, No. 1, 24-32.

Carr, D. B., A. R. Olsen, S. M. Pierson, and J. P. Courbois (1998b) "Using Linked Micromap Plots To Characterize Omernik Ecoregions," *Data Mining and Knowledge Discovery*, submitted.

Carr, D.B., Littlefield, R.J., Nicholson, W.L., and Littlefield, J.S. (1987) "Scatterplot Matrix Techniques For Large N," *JASA*, Vol. 82, 424-436.

Carr, D.B., Kahn, R., Sahr, K., and Olsen, A.R. (1998a) "ISEA Discrete Global Grids," *SC& G Newsletter*, Vol. 8, No. 2/3, 31-39.

Carr, D.B. and Pierson, S. (1996) "Emphasizing Statistical Summaries and Showing Spatial Context with Micromaps," *SC& G Newsletter*, Vol. 7, No. 3, 16- 23.

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

Cleveland, W.S. and McGill, R. (1984) "Graphical Perception: Theory, Experimentation, and Application to the Development of Graphical Methods," *JASA*, Vol. 79, 531-554.

Daly, C., Neilson, R.P., and Phillips, D.L. (1994) "A statistical-topographic model for mapping climatological precipitation over mountainous terrain," *Jour of Appl Meteorology*, 33, 140- 158.

Frigge, M., Hoaglin, D.C. and Iglewicz, B. (1989) "Some Implementations of the Box Plot," *The American Statistician*, 43,50-54.

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 of the Assoc of American Geographers*, Vol. 85, 339-355.

McGill, R., Tukey, J.W. and Larsen, W.A. (1978) "Variation of Boxplots," *The American Statistician*, Vol. 32, 12-16.

Monmonier, M. (1988) "Geographical Representations in Statistical Graphics: A Conceptual Framework" in *1998 ASA Proc of the Section on Stat Graphics*, 1-10.

Monmonier, M. (1993) *Mapping It Out*, University of Chicago Press, Chicago, IL.

Olsen, A. R., Carr, D.B., Courbois, J.P, and Pierson, S.M. (1996) "Presentation of Data in Linked Attribute and Geographic Space," Poster presentation, ASA Annual Meeting.

Omernik, J.M. (1987) "Ecoregions of the conterminous United States," *Ann Assoc Am Geographers*, Vol.77, 118-25.

Omernik, J.M. (1995) "Ecoregions: a spatial framework for environmental management," in *Biological Assessment and Criteria: Tools for Water Resource Planning and Decision Making*, eds. Davis, W.S. and Simon, T.P. Lewis Publishers, Boca Raton, FL, 49-62.

Omernik, J.M. and Bailey, R.G. (1997) "Distinguishing between watersheds and ecoregions," *J of the Am Water Resources Assoc*, Vol. 33, 935-49.

Scott, D.W. (1992) *Multivariate Density Estimation, Theory, Practice and Visualization*, John Wiley and Sons, New York, NY.

Small , C.G. (1989) "A Survey of Multidimensional Medians," Tech Report Stat-89-08, Dept of Statistics and Actuarial Science, University of Waterloo.

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

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

Tukey, J.W. (1993) "Graphic Comparisons of Several Linked Aspects: Alternatives and Suggested Principles," *JCGS*, Vol. 2. No. 1, 1- 33.

Dan Carr
Institute for Computational Statistics and Informatics
George Mason University
dcarr@galaxy.gmu.edu

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

Suzanne M. Pierson
OAO Corporation
spierson@mail.cor.epa.gov

Pip Courbois
Oregon State University
courbois@stat.orst.edu

Exploring Time Series Graphically

By Antony Unwin and Graham Wills

Introduction

Interactive graphical tools are available in a number of packages (Data Desk, JMP, SAS Insight and to a lesser extent LispStat, S+ and others), but none provide interactive tools for working with time series. This is surprising, as graphical displays have always been considered important for time series analysis. The natural time ordering of the data makes graphs very informative and the lack of independence of the data makes simple summary statistics less useful than in other contexts.

Diamond Fast was designed to provide some basic tools for exploring multivariate, irregular, short time series where analytic methods could not be applied, but it proves to be valuable for working with univariate, regular, long series too [Unwin and Wills]. The main idea is to display series in a common window, to query them, to move them around, and to compare them. Its a bit like working with icons in a window. As the package has not been widely publicized but is still in use, we thought it would be helpful to summarize its main features here.

Scaling

Most books on statistical graphics explain at length how scales should be chosen and drawn. There are some specific recommendations for time series in Cleveland [1993], where he points out that different aspect ratios may be necessary for studying local and global structure. This is not so relevant for interactive graphics where a series can be interactively rescaled both vertically and in time and where not only the scales, but also individual points can be directly queried. Figure 1 shows weekly values of the Dow Jones Industrial average between 1929 and 1932. The series has been queried at its maximum before the crash. The same mouse click can call up information on a point, on an axis or on the graph itself, depending on where you click.

The main method of changing the vertical scaling in Diamond Fast is to grab the series graph and pull it upwards to magnify or downwards to shrink. This is very effective as you can see how the series changes and judge the effect you want to obtain. Technically this is achieved by distorting the current screen image and when the mouse button is released an accurate picture of the graph is drawn at the new resolution.