USING MULTIVARIATE CLUSTERING TO CHARACTERIZE ECOREGION BORDERS

The authors' clustering technique unambiguously locates, characterizes, and visualizes ecoregions and their borders. When coded with similarity colors, it can produce planar map views with sharpness contours that are visually rich in ecological information and represent integrated visualizations of complex and massive environmental data sets.

Ecologists have long used massive data sets as the basis for visualizing ecoregion maps.[1,2] Ecoregions are areas containing similar environmental conditions that are classified for particular purposes. For example, the US Department of Agriculture publishes a map of Plant Hardiness Zones, which divides the US into different ecoregions so that gardeners can select appropriate plants and shrubs for their particular area. The ecoregion is one of the most important concepts in managing and understanding landscape ecology.[3,4] Unfortunately, ecologists have long struggled with exactly how and where to locate the dividing lines between ecoregions.[5,6]

Historically, the process of regionalization—drawing ecoregion borders—has been subjective: experts have attempted to integrate and weigh all of the environmental characteristics and draw the borders accordingly, often without being able to elucidate the logic behind their lines. This subjectivity leads to frequent revisions[5–8] and disagreements over particular locations,[9] and hampers widespread acceptance and use of such maps. Some experts, in fact, are unable to correctly identify actual ecoregion maps from synthetic maps simulated using a fractal technique.[10]

Part of the problem is the variable nature of the borders between ecoregions. Some borders are very sharp and distinct, and you can literally stand with your feet in two clearly different regions. Ecologists call such unequivocal and easy-to-locate borders *ecotones*, because they represent sharp cuts. However, most borders are more like the M.C. Escher woodcut *Sky and Water I* (*www.cs.rochester.edu/u/si/images/escher/birds_fish.gif*), in which black birds slowly transform into white fish. Although the picture clearly contains two distinct creatures, it's difficult to locate a line of demarcation between them. We define a new term, *ecopause*, to indicate the indistinct nature of such borders. But a border may actually change character along its length. For example, a border can begin in one geographic location as an ecotone and transform slowly along its length into an ecopause. Unfortunately, ecologists have had only simple lines with which to visualize these many types of borders.

Locating ecoregion borders is a multivariate decision process that must consider a large geo-

1521-9615/99/$10.00 © 1999 IEEE

WILLIAM W. HARGROVE
University of Tennessee
FORREST M. HOFFMAN
Oak Ridge National Laboratory

graphic data set for each of multiple environmental conditions. We have developed an objective technique called Multivariate Geographic Clustering, which objectively computes border placement between ecoregions, given maps of all environmental conditions under consideration. Our technique lets us locate and visualize ecoregion borders, and portray the instantaneous sharpness of those borders at every point along the line. Here, we present our technique and offer sample visualizations.

Multivariate Geographic Clustering

Rather than relying on expertise, Multivariate Geographic Clustering uses standardized values for each selected environmental condition in a map's individual raster cells as coordinates that specify the cell's position in environmental data space. The number of dimensions in data space equals the number of environmental characteristics. Two raster cells with similar environmental characteristics from anywhere in the map will appear near each other in data space; their closeness and relative position quantitatively reflects environmental similarities.

Our algorithm disassembles the map cells from geographic space and uses the standardized value of each of the environmental characteristics as coordinates to replot the cells in environmental data space. Because the density of cells in data space is variable, we use an iterative classification procedure to group nearby cells into clusters based on similar environmental conditions.

The process

To begin the process, the user specifies the desired number of clusters. The initial part of the algorithm then examines observations sequentially to find the most widely separated set of cells that will constitute the initial cluster "seeds." Each map cell is then compared against all cluster seeds and assigned membership in the cluster closest to it in terms of Euclidean distance. After all map cells are assigned, new cluster centroids are calculated as the mean of each coordinate in the cluster. At this point, the iterative assignment procedure repeats. Cells do not move in environmental data space; rather, the cluster centroids slowly migrate until they achieve equilibrium. When fewer than a specified number of map cells change cluster assignments in a particular iteration, the process halts.

Figure 1 shows a visualization of 3,000 clusters in a 3D data space representing the US. In this case, the three dimensions are the first three principal component scores resulting from nine environmental characteristics (we discuss this in more detail later). Because showing individual map cells would obscure the view entirely, we show clusters instead. Cluster icons are sized and colored based on the number of member cells. Clusters with the largest membership tend to be centrally located in data space; cluster sizes follow a negative exponential distribution. Because the procedure generates clusters with nearly uniform variance in a cluster, the actual radius of all clusters is nearly equal, regardless of membership.

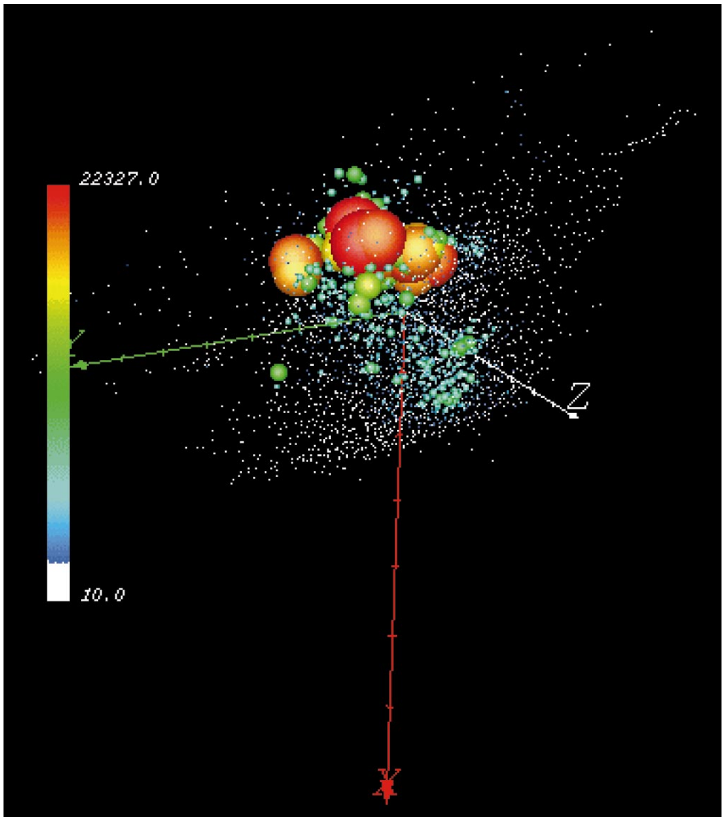

Figure 1. A visualization of 3,000 clusters in a 3D data space representing the United States. The number of member cells determines the cluster icon's size and color.

Map cells with their final cluster assignments are then reassembled into their proper geographic positions, and the resultant ecoregion map can be color-coded by cluster assignment. Because adjacent raster cells are likely to have similar environmental values, ecoregion clusters are often geographically contiguous. However, because the geographic location is not used for clustering, clusters can be spatially disjoint, and two map cells with similar environments could be classified in the same ecoregion even though

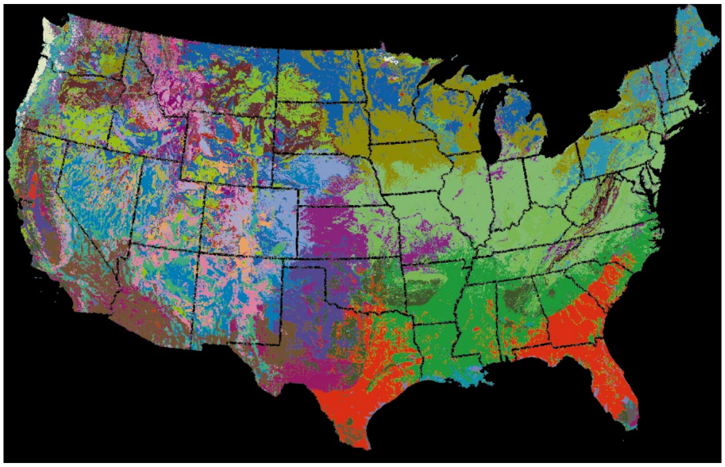

Figure 2. Using nine environmental conditions, the Multivariate Geographic Clustering technique segregated the continental US into 50 distinct ecoregions, represented by randomly assigned colors.

they are widely separated geographically. For example, two widely spaced but environmentally similar mountaintops could be classified in the same ecoregion cluster.

Implementation

We have implemented Multivariate Geographic Clustering in a parallel algorithm coded in C using the Message Passing Interface. Our code is dynamically load-balancing and fault-tolerant, and it performs both initial seed-finding and iterative cluster assignment in parallel. The clustering algorithm is inherently parallelizable, because individual nodes can independently classify a portion of all cells, and then combine results at the end of the iteration.

We developed the Multivariate Geographic Clustering parallel algorithm and code on a highly heterogeneous Beowulf-class parallel machine constructed from surplus 486- and Pentium-based PCs. This 128-node "Stone SouperComputer" is described elsewhere[11] and online at *www.esd.ornl.gov/facilties/beowulf*.

We performed many empirical regionalizations for the conterminous US at one-square-kilometer resolution for up to nine environmental characteristics[12,13] and have divided the US into as many as 7,000 distinct ecoregions.[14] At this resolution, each of the nine US environmental-condition maps comprises more than 7.8 million cells. This map, data, and ecoregion resolution surpasses what ecoregion experts usually accomplish.

The example US ecoregions we show here are from a Multivariate Geographic Clustering at a resolution of four square kilometers on nine particular environmental characteristics important to plant growth. The environmental characteristics we considered included elevation, slope, soil bulk density, mineral soil depth, bedrock depth, mean annual temperature, mean annual precipitation, soil water-holding capacity, and mean annual solar insolation, including cloud interception.

Our first *principal-component analysis* grouped soil density, soil depth, and bedrock depth into a principal component encompassing soil factors. The second PCA grouped temperature and precipitation, and inverse elevation and slope. The third PCA grouped solar insolation and inverse soil water-holding capacity. We used the three principal components as the axes for the environmental data space and the basis for this ecoregionalization.

Figure 2 shows the resulting map of the conterminous US, which is divided into 50 distinct ecoregions based on the nine environmental conditions and identified by randomly assigned colors. Although this map contains about half a million cells, each with nine characteristics, parallel Multivariate Geographic Clustering can efficiently handle much larger problems.

Color-coding similarities

Visualizing ecoregions with random colors emphasizes the location of the borders. However, ecologists might also want to see the relative mix of conditions in bordering ecoregions. Because the cluster centroid's final location is, by definition, central, its coordinates describe the average ecological conditions in the cluster ecoregion. Comparing centroid coordinates from two ecoregions quantifies the differences between the average environments in each.

If, through PCA, we condense numerous "raw" environmental variables into three orthogonal principal-component axes in the environmental data space, we can perform a one-to-one scalar mapping of the first, second, and third principal component scores to a red-green-blue (RGB) color triplet. In this way, we can combine the three coordinates for each cluster centroid to specify a unique color for that ecoregion. Under this similarity-colors scheme, each ecoregion's color indicates the relative mix of each environmental factor. Comparing adjacent ecoregions is thus simple: ecoregions of similar colors have similar environments.

Figure 3 shows Figure 2's ecoregions using the similarity-colors scheme. With Figure 2's random-colors scheme, all intervening borders are easily seen. However, with the new RGB similarity colors, the borders between some adjacent and similar ecoregions nearly disappear. The map thus becomes a gradient of slowly changing colors that quantitatively reflect the mix of environmental conditions found at each point. In Figure 3, green, blue, and red represent the three factors: soil properties, temperature and precipitation, and solar and water-holding capacity, respectively. Black results from small but balanced values of all factors, and white from large but balanced values of all factors. Thus, white areas in Florida, Texas, and California's Central Valley reflect high solar insolation, low water-holding capacity, high bulk density, deep soils and bedrock, high temperature and precipitation, low elevation, and gentle slopes.

RGB-encoded similarity-colors maps converge rapidly to show the same large regional trends in ecological relationships. For example, if two ecoregionalizations are produced from the same environmental conditions, but one is divided finely into many ecoregions while the other is divided coarsely into relatively few, the similarity-colors versions of each very different map will be indistinguishable from each other. This convergence occurs despite the fact that the polygons underlying each map are completely different—only the RGB coding technique is the same. Thus, beyond some minimum number of ecoregions, the same regional ecological patterns are revealed regardless of the number of ecoregion divisions.

Gauging representativeness

To characterize the sharpness of the borders, we must be able to quantify *representativeness*: how representative a particular location is of the parent ecoregion. As we described earlier, each cluster's final centroid is the best single way to represent that cluster ecoregion because it represents the arithmetical average of all member cells. Map cells that are in the cluster's interior, close to the mean centroid, are highly representative of this ecoregion; map cells in the cluster's outer "shell" are less representative. These outlying cells are the ones that might change cluster assignments in another iteration of the classifying algorithm.

Given this, we quantify representativeness by measuring the Euclidean distance from each cell to its assigned cluster's centroid. We can thus compute a representativeness value for all map locations. Also, because more ecoregions mean more (and closer) centroids, the metric takes the number of ecoregion divisions into account. Cells close to their centroids are always more representative of their cluster ecoregions than outlying cells.

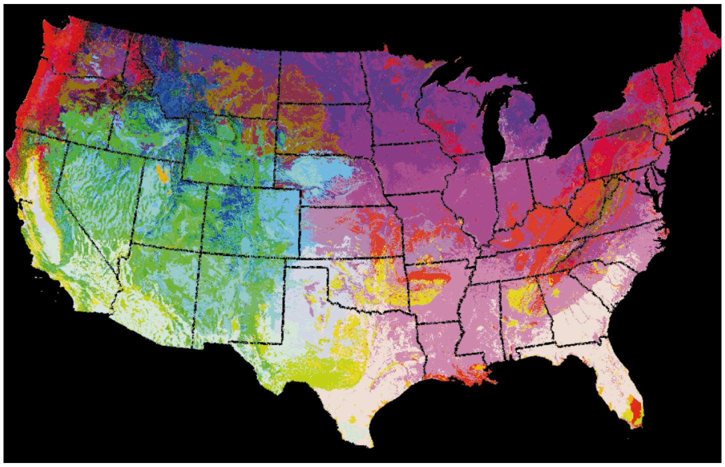

Figure 3. A similarity color scheme of the US ecoregions shown in Figure 2. Factor 1 (soil properties), is shown in green; factor 2 (temperature and precipitation) in blue; and factor 3 (solar and water holding) in red. Black represents regions where there are small, but balanced values of all factors; white represents areas in which there are large but balanced values of all factors.

Defining elevation

If we map the distance from each cell to its cluster centroid back into geographic space and depict these values as elevations, we can create a surface whose height inversely corresponds to the cell's representativeness at that geographic location. Because we can calculate such a value for all cells, this representativeness surface will be complete and continuous across the map. This theoretical elevation surface reflects representativeness, and represents something different from the locations' actual topographic elevations.

In such an elevation surface, idealized hypothetical cluster ecoregions would appear as a series of depressions or craters, with border regions tracing along the tops of the crater rims. The crater's deepest spots would correspond with cells at or near the cluster's centroid, representing the lowest geographic locations.

Edge characteristics

We use elevation profile cross-sections to characterize adjacent borders, which can be sharp, fuzzy, or a combination of the two. For example, a border might be steep-sided and "U"-shaped, with sharp borders characteristic of an ecotone, or it might descend more gradually, in a "V"

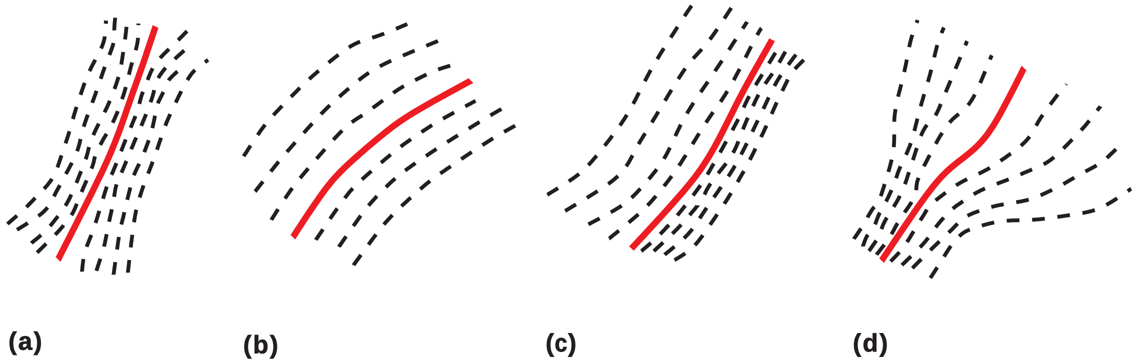

Figure 4. Using contour lines, we can visualize borders that are (a) sharp on both sides, (b) fuzzy on both sides, and (c) mixed. We can also represent (d) borders that change sharpness characteristics along their length.

shape, characteristic of an ecopause. Also, because edge properties are dependent on each adjacent cluster, each side has distinct (and possibly different) properties. Although initially counterintuitive, this "sidedness" property is logical, given that we are characterizing the transition from the border to the centroid independently on each side. Thus, for example, a border might be sharp on one side and fuzzy on the other.

To visualize border sharpness, we use contour lines. As Figure 4 shows, closely spaced contours reflect steep sides and therefore a sharp ecotone; widely spaced contours indicate gradually sloping crater walls and a fuzzy, gradual ecopause. We can also represent borders that change from fuzzy to sharp or vice versa, as the figure shows.

Visualization examples

We have used our clustering technique to produce ecoregion maps for many areas. Here we examine in detail ecoregions and borders from the southeastern US and southern and central California. Full-size high-resolution versions of these visualizations, along with other examples, are available online at *www.esd.ornl.gov/~hnw/borders*.

Figure 5 shows a 3D visualization of the representativeness surface for Alabama, southwest Georgia, and northern Florida. Each square in the mesh represents a single four-square-kilometer raster cell; we find the cell's elevation by measuring the Euclidean distance between it and the centroid of its cluster. Cluster membership is shown in Figure 5 as the (random) color of

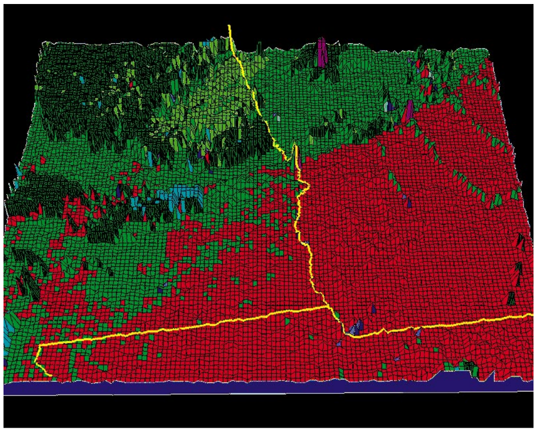

Figure 5. Representativeness topography for Alabama (upper left), southwest Georgia (upper right), and northern Florida. Ecoregions are shown as random colors.

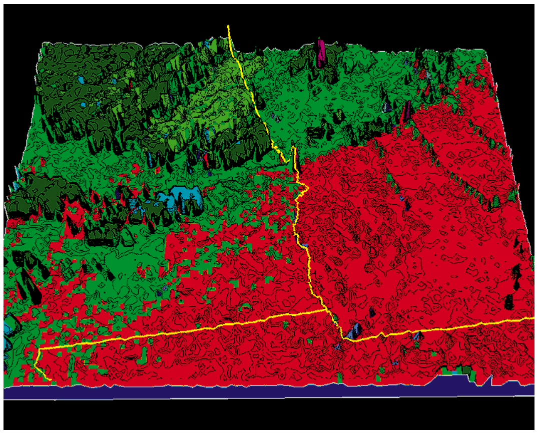

Figure 6. Southeastern ecoregions represented with equal-elevation contours draped onto the representativeness surface to visualize the sharpness of the ecoregion borders.

each cell. The representativeness topography is continuous and interpretable at this resolution.

The region's major cities (Atlanta, Macon, and Columbus) are shown as a discontinuous purple urban cluster. From west to east, four rivers (the Flint, Ocmulgee, Oconee, and Ogeechee) are seen as linear extensions of central Georgia's kelly-green piedmont ecoregion, flowing into the state's red coastal plain. In southern Alabama, the red coastal plain and kelly-green piedmont ecoregion colors interdigitate, showing single cells of red within green and vice versa. The light-green southwestern Appalachians pass through northeast Georgia into eastern Alabama. In northwestern Alabama, the olive-drab ridge-and-valley ecoregion forms a higher-elevation representativeness plateau.

Figure 6 shows equal-elevation representativeness contours visualizing the sharpness of these ecoregion borders. In southern Alabama, the contours' random orientation and meandering character near the red coastal plain and kelly-green piedmont ecoregions clearly indicate that this border is an ecopause. In contrast, northern Alabama's closely-spaced, parallel contour lines separating the kelly-green piedmont from the olive-drab ridge-and-valley represent this border as a sharp ecotone.

Figure 7 shows the same southeastern ecoregions using the RGB-encoding similarity scheme with border sharpness contours. Although the colors appear to simply reflect the elevations, they are actually derived from the centroid coordinates from each ecoregion. Abrupt color changes are accompanied by the numerous parallel contours of an ecotone, while subtle color changes are accompanied by the meandering contours of an ecopause. In a smaller region such as this, colors correlate with height; however, distant locations with equal representativeness elevations might have substantially different environment colors. Once we interpret the sharpness contour lines, we can create a simple plan-view map with random ecoregion colors (see Figure 8) that adequately captures both the location and the characteristics of ecoregion borders. Dense adjacent contours create thick black lines along sharp ecotone borders.

Figure 9 shows a hallucinogenic planar view of California from Los Angeles to San Francisco. The representation uses sharpness contours and randomly selected ecoregion colors. San Francisco Bay can be seen at the upper left; the San Joaquin Valley is represented as a purple ecoregion in the north and a gray ecoregion in the south. Figure 10 shows the jagged topography as a mesh, again with randomly selected ecoregion colors. The Coastal Range appears to the west and the Sierra Nevada mountains to the east, separated by the much flatter and more representative San Joaquin Valley. As the meandering sharpness contours in Figure 11 show, there is little representativeness difference between the northern purple and southern gray ecoregions of this valley. Figure 12 shows similarity colors based on the nine environmental characteristics; the flat yellow plateau at the upper right is Mono Lake. This body of water is relatively homogeneous, and substantially different from the surrounding land ecoregions.

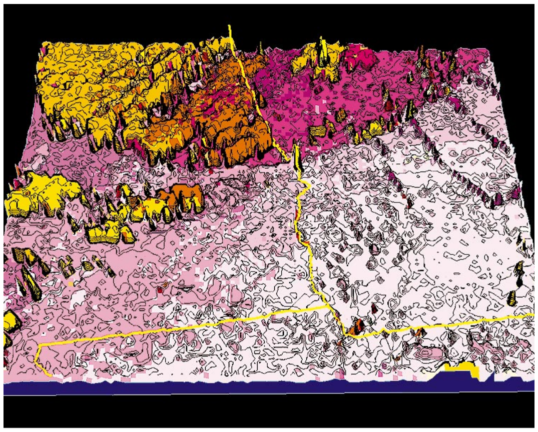

Figure 7. Southeastern ecoregions colored by similarity, with border sharpness contours. Parallel contour lines indicate the sharp ecotone border between Northern Alabama's piedmont and ridge-and-valley regions.

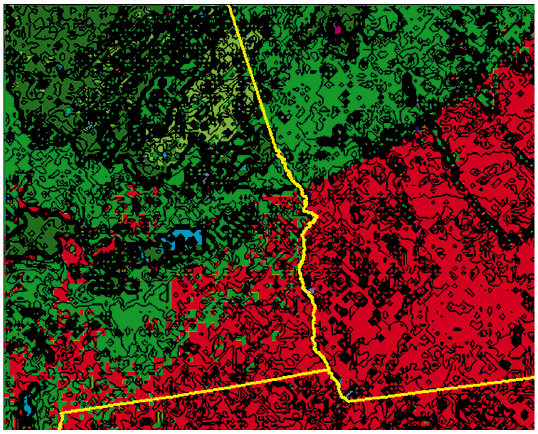

Figure 8. Plan view of southeastern ecoregions with random colors and representativeness contours. Dense adjacent contours create thick black lines along sharp ecotone borders.

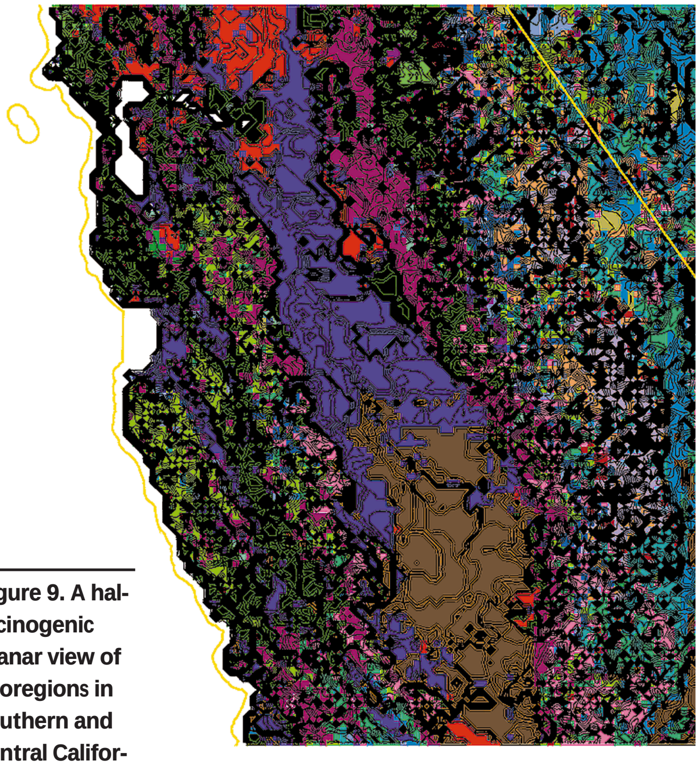

Figure 9. A hallucinogenic planar view of ecoregions in southern and central California, using sharpness contours and randomly selected ecoregion colors.

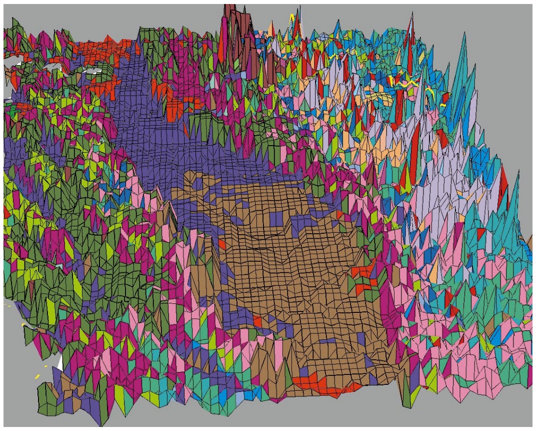

Figure 10. California ecoregions represented as a mesh draped over a representativeness topography, again using randomly selected colors.

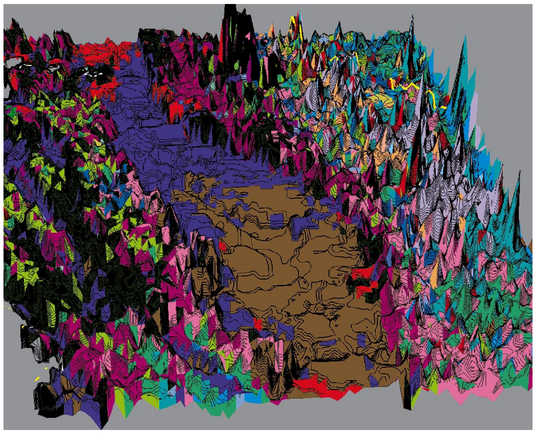

Figure 11. Sharpness contours show little difference between the northern part of the San Joaquin Valley (purple) and the southern part (gray).

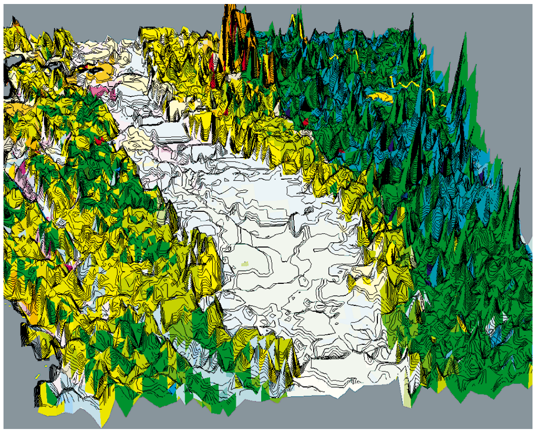

Figure 12. Visualization of California ecoregions using similarity colors based on nine environmental characteristics.

Even at the national scale, careful application of sharpness contours can reveal ecoregion patterns. Plains in the southern and northwestern US share a relatively gentle topography, although their environmental characteristics differ, as indicated by the different similarity colors in Figure 13. The transitions between such zones are represented by gradual ecopauses. Sharper, ecotone-type boundaries are shown in mountainous regions and much of the western US.

Characterizing ecoregion borders is important for more than just ecological understanding. Border movement at fuzzy edges can be the first detectable evidence of climate change. Characterizing borders can also facilitate comparisons among alternative ecoregionalizations. For example, differences in the location of sharp edges are more significant than different placements of fuzzy edges.

Visualizations such as those we show here can also provide a way to inspect the appropriateness of geographic clustering. For example, the appearance of multiple low areas within a single cluster ecoregion might suggest that you need more divisions, whereas borders passing through low areas might suggest you need fewer.

It is human nature to attempt to impose order

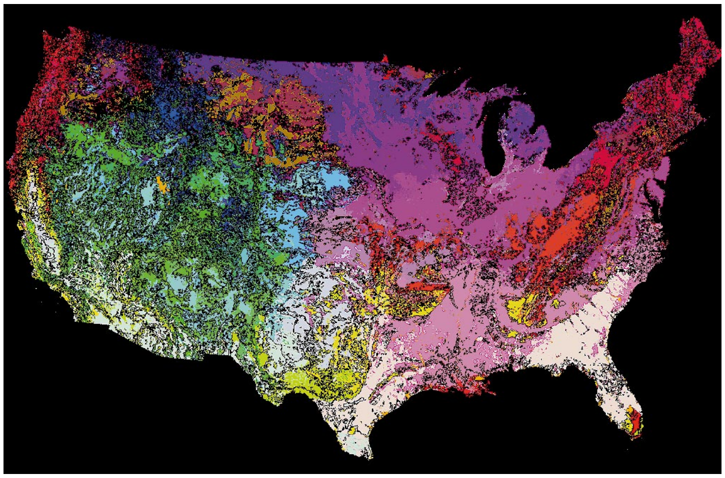

Figure 13. Ecoregions of the US based on nine environmental characteristics visualized using similarity colors and five sharpness contour levels.

by drawing lines to divide and categorize. The problem is that the world is full of Escher-like gradients. Our Multivariate Geographic Clustering technique helps not only to draw the lines, but also to characterize the sharpness of the borders that they represent.

Acknowledgments

Oak Ridge National Laboratory is managed by Lockheed Martin and supported by the US Dept. of Energy under contract DE-AC05-96OR22464.

References

1. J.M. Omernik, "Ecoregions of the Conterminous United States" (map), *Annals of the Assoc. of Am. Geographers*, Vol. 77, No. 1, 1987, pp. 118–125.
2. R.G. Bailey, "Delineation of Ecosystem Regions," *Environmental Management,* Vol. 7, 1983, pp. 365–373.
3. J.M. Omernik and R.G. Bailey, "Distinguishing between Watersheds and Ecoregions," *AWRA Water Resources Bulletin,* Vol. 33, No. 5, 1997, pp. 935–949.
4. J.M. Omernik, "Ecoregions: A Spatial Framework for Environmental Management," *Biological Assessment and Criteria: Tools for Water Resource Planning and Decision Making*, W.S. Davis and T.P. Simon, eds., Lewis Publishing, Boca Raton, Fla., 1995, pp. 49–62.
5. R.G. Bailey, *Ecosystem Geography*, Springer-Verlag, Berlin, 1996.
6. "Ecoregions and Subregions of the United States" (map), R.G. Bailey et al., eds., *US Geological Survey*, Washington, DC, 1994; accompanied by table of map unit descriptions compiled and edited by W.H. McNab and R.G. Bailey, prepared for the US Forest Service.
7. R.G. Bailey, *Description of the Ecoregions of the United States*, 2nd ed., Misc. Pub. No. 1391, US Forest Service, Washington, DC, 1995.
8. R.G. Bailey, *Ecoregions Map of North America: Explanatory Note*, Misc. Pub. 1548, US Forest Service, 1998.
9. Commission for Environmental Cooperation, *Ecological Regions of North America: Toward a Common Perspective*, Montreal, Canada, 1997.
10. W.W. Hargrove, P.M. Schwartz, and F.M. Hoffman, "The Fractal Landscape Realizer," 1997; www.esd.ornl.gov/projects/realizer/ (current June 1999).
11. F.M. Hoffman and W.W. Hargrove, "Cluster Computing: Linux Taken to the Extreme," *Linux Magazine,* Vol. 1, No. 1, Mar. 1999, pp. 56–59.
12. W.W. Hargrove and R.J. Luxmoore, "A Spatial Clustering Technique for the Identification of Customizable Ecoregions," 1997; www.esri.com/library/userconf/proc97/PROC97/TO250/PAP226/P226.htm (current June 1999).
13. W.W. Hargrove and R.J. Luxmoore, "A New High-Resolution National Map of Vegetation Ecoregions Produced Empirically Using Multivariate Spatial Clustering," 1998; www.esri.com/library/userconf/proc98/PROCEED/TO350/PAP333/P333.htm (current June 1999).
14. W.W. Hargrove and F.M. Hoffman, "National Clustering," 1998; www.esd.ornl.gov/projects/clustering/ (current June 1999).

William W. Hargrove is a member of the research faculty at the University of Tennessee's Energy, Environment, and Resources Center, serving on contract to the Oak Ridge National Laboratory's Geographic Information and Spatial Technologies Group. His areas of expertise include computer algorithms, Geographic Information Systems, landscape ecology, and simulation modeling. Contact him at the Univ. of Tennessee Energy, Environment, and Resources Center, Systems Development Inst., 10521 Research Dr., Ste. 100, Knoxville, TN 37932; hnw@fire.esd.ornl.gov.

Forrest M. Hoffman is a computer specialist in the Environmental Sciences Division at Oak Ridge National Laboratory in Oak Ridge, Tennessee, where he develops parallel (and serial) environmental models and tools for scientific visualization, large data-set management and administration, and Internet technologies. In his spare time, he builds parallel computers. Contact him at the Oak Ridge Nat'l Laboratory, Environmental Sciences Div., P.O. Box 2008, M.S. 6036, Oak Ridge, TN 37831-6036; forrest@esd.ornl.gov.