Using Representativeness to Guide Expansion of the AmeriFlux Network
William W. Hargrove and Forrest M. Hoffman, Oak Ridge National Laboratory

Multivariate Geographic Clustering (MGC)

☞ Multivariate clustering is the division or classification of objects into groups or classification of objects into groups or categories based on the similarity of their properties.

☞ Non–hierarchical clustering produces a single level of division of objects into some predetermined number of groups.

☞ Multivariate Geographic Clustering (MGC) applies non–hierarchical statistical clustering to the classification of geographic areas from maps in a Geographic Information System (GIS).
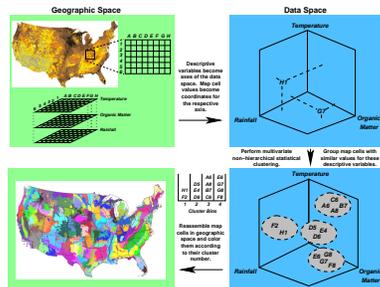
Selected Multivariate Characteristics

☞ Ecoregions will be created that have a similar mixture of the selected multi–variate conditions.

☞ Different multivariate properties can be selected to address specific problems.

☞ This set of 25 environmental properties were selected as generally important:

elevation	soil water capacity
maximum annual temperature	frost–free days
mean annual temperature	depth to water table
minimum annual temperature	soil bulk density
12 monthly mean precipitations	depth of mineral soil
soil nitrogen	solar aspect
soil organic matter	mean solar insolation

The 1000 Most–Different Ecoregions

☞ This map shows the 1000 most–different ecoregions in the U.S., based on the 25 environmental characteristics listed above.

☞ The map provides a high–resolution basis for subsequent analyses, and contains more ecoregions than usually produced by human experts.

☞ As the number of clusters or ecoregions is increased, each becomes more specialized and more tightly defined.
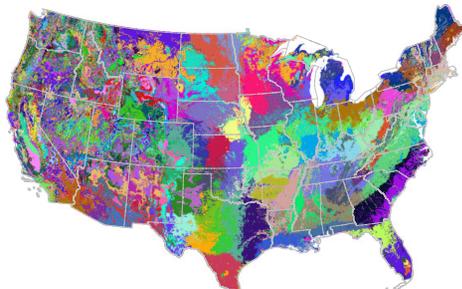
The Representativeness Concept

☞ Because the ecoregions are statistically derived, one can select a single ecoregion of particular interest, and then produce a sorted vector of the similarity of all other ecoregions to the selected one.

☞ The chosen ecoregion establishes an origin in data space. Using the Euclidean distance from this origin to each other ecoregion, a pairwise similarity measure can be calculated.

☞ Coding these pairwise similarity values as gray levels, a map can be drawn which cartographically shows the degree of similarity of all ecoregions in the map to the selected ecoregion of interest.

☞ Darker areas are high in similarity to the selected ecoregion, which is shown in red.

Quantifying "Smoky Mountains–ness"

☞ A map of "Smoky Mountains–ness" essentially rediscovers the entire Eastern Deciduous Forest Biome.

☞ One small spot in the Ozarks and one spot in the Monongahela National Forest of West Virginia consist of pure "Smokies–ness."

☞ The Adirondacks of New York are spatially disjunct, but are relatively high in multivariate "Smokies–ness."

Analysis of Sites in a Network

☞ The converse of quantitative similarity is anti–similarity or non–representativeness.

☞ Within the context of an existing network of sites or samples, this can form the basis for a tool which quantifies the degree of representativeness coverage for a particular established network.

☞ A network in this sense consists of a geographic constellation of existing installations, facilities, or locations where samples have been taken.

☞ Quantitative similarity is no longer based on a comparison with some single selected ecoregion, but on comparisons with multiple sites within an established network.

☞ The best place to locate an additional site is the place that is least well–represented by the network of existing sites.

☞ Instead of a one–to–one centroid comparison, this is a one–to–many centroid comparison in data space.

What Network Analysis Does NOT Consider

☞ Network Analysis is ecoregion–based, and operates at the scale of the entire sampling network.

☞ Does NOT consider specific local conditions, land uses, disturbance history, or anthropogenic treatments, i.e., clear–cutting, forest plantations, and agriculture.

☞ Fine–scale land conditions could be included as inputs if deemed important (and data are available as continuous variables).

☞ Results are with respect to the selected input variables only.

☞ Results depend on the intelligent selection of the environmental variables being considered, and on the quality of the input data.

☞ A custom ecoregionalization that is based on variables chosen specifically for a particular problem should be used for best results.

William W. Hargrove
hnw@fire.esd.ornl.gov

Forrest M. Hoffman
forrest@climate.ornl.gov

http://research.esd.ornl.gov/~hnw/networks2/

Quantifying Coverage of the AmeriFlux Network

☞ There were 52 AmeriFlux tower sites in the conterminous United States as of September 2002.

☞ Southern Texas, the Sonoran Desert, and the Pacific Northwest were poorly represented by existing AmeriFlux towers.
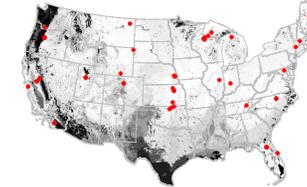
Coverage of the Upgraded AmeriFlux Network

☞ By June 2003, AmeriFlux had added 4 new sites, removed 2 existing sites, and moved 1 existing site (corrected the position info).

☞ This map shows the coverage of the new upgraded AmeriFlux network.

Improvement in AmeriFlux Representation

☞ Subtracting the last two network representation maps reveals the incremental improvement in network representation.

☞ Blue sites are new additions. Red sites are no longer in the network.

☞ Red shows representation losses. Gray shows representation gains.

☞ The undirected addition of four sites did NOT produce dramatic gains in representation for the upgraded AmeriFlux network.
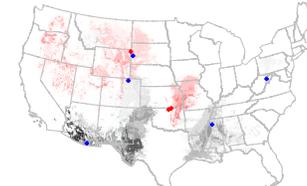
Directed Improvement in AmeriFlux Representation

☞ We can do about as well with the directed addition of a SINGLE new site as AmeriFlux has done with the undirected addition of four new sites.

☞ Single new theoretical site added halfway between San Antonio and Laredo, TX.

☞ Same two sites lost and same site moved as before.

☞ A "constrained clustering" algorithm could find the single best additional site which maximizes the incremental gain in network representation.