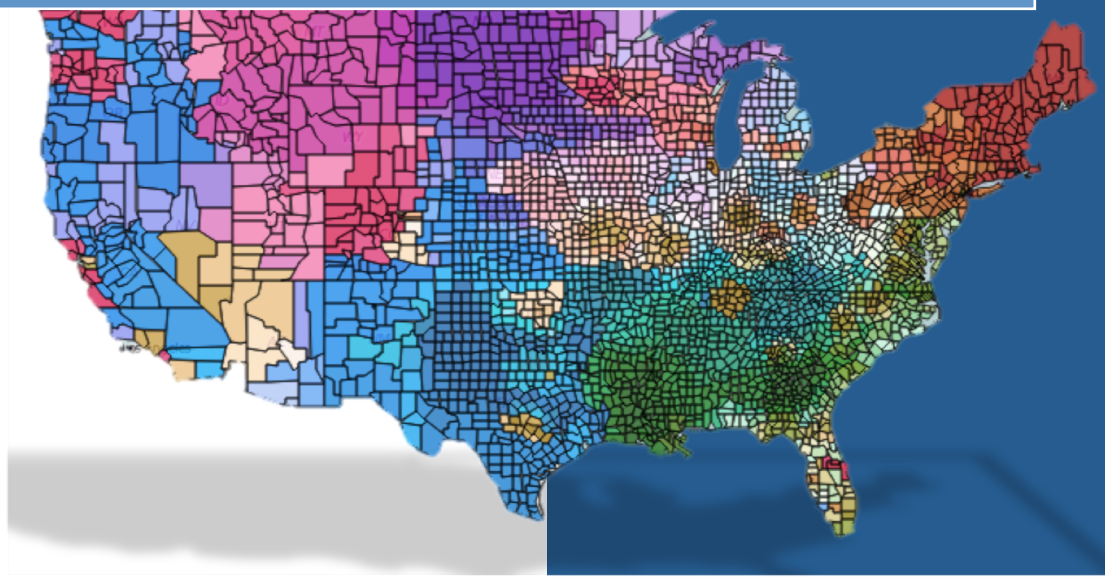
2016

Multivariate Mapping and Regionalization Online

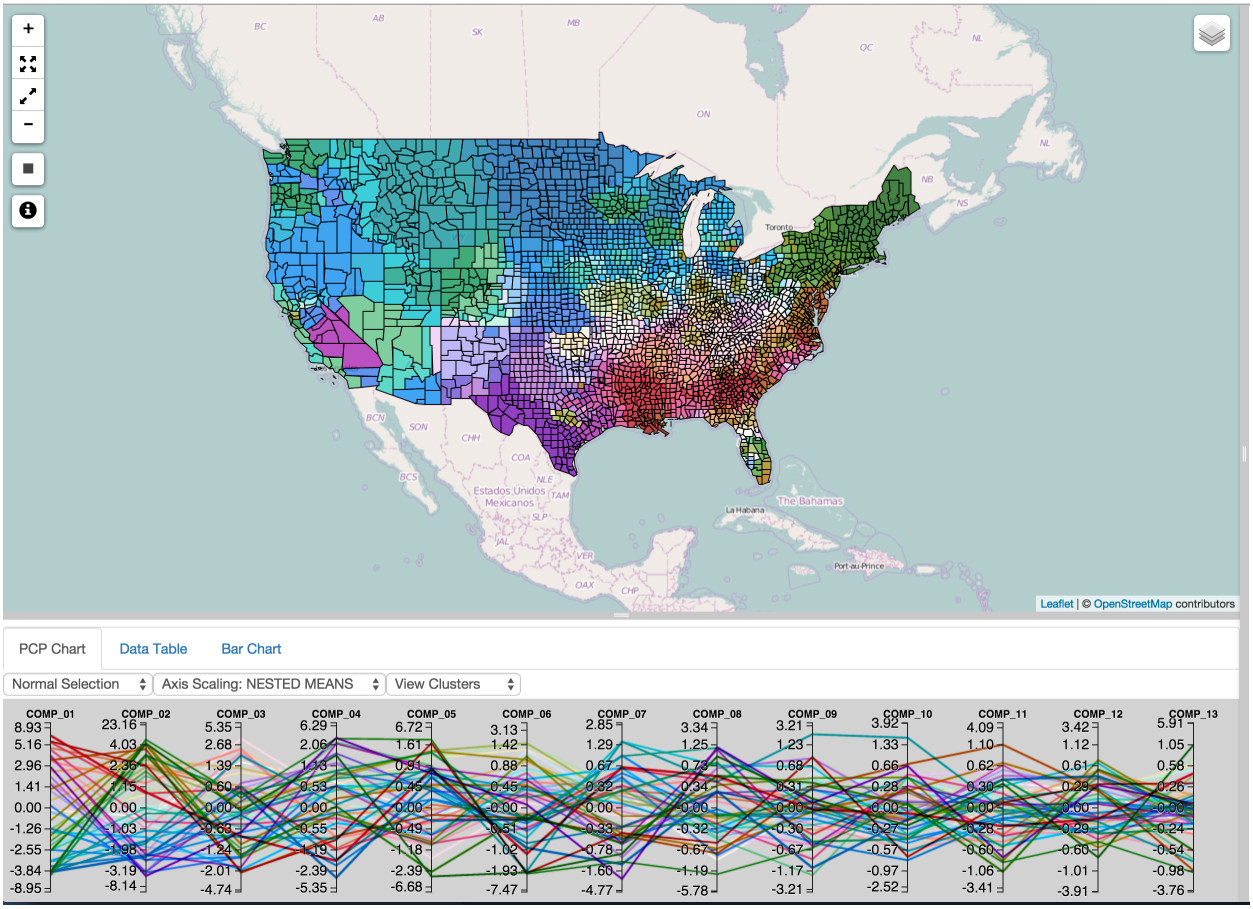


ZillionInfo

2/22/2016

Introduction

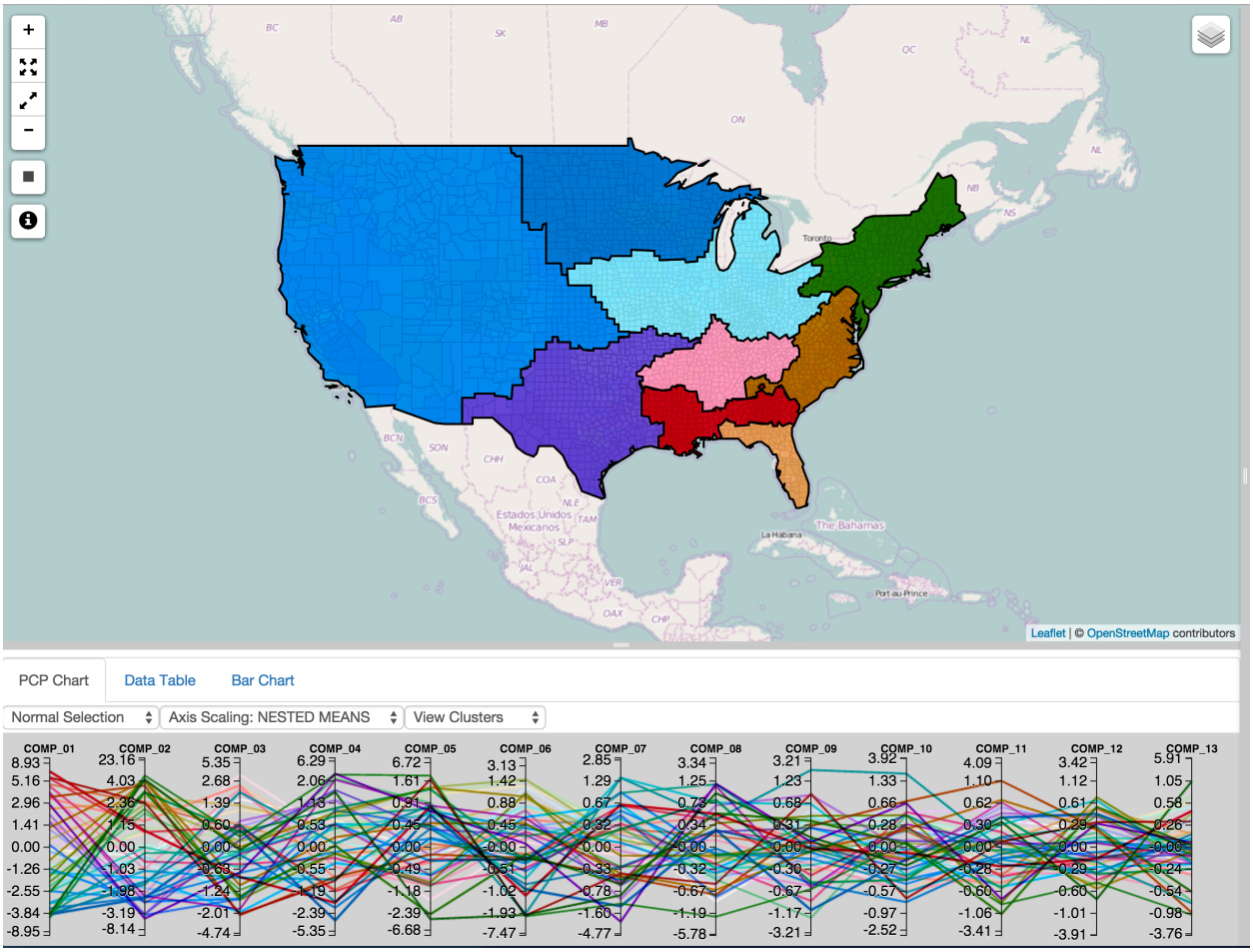
Multivariate Mapping allows for various sets of data to be simultaneously compared, facilitating an analysis that can contribute to a higher level of insight than can be provided by existing data analysis approaches. It creates an overview of multivariate data that quickly and easily reveals complex patterns. With a user-friendly interface, Multivariate Mapping is designed to support human interactions to explore and examine these patterns. Computational and visual methods, once combined together, can mitigate each other's weakness and collaboratively discover complex patterns in large geographic datasets in a simple way.



Regionalization is an online service that can be used to aggregate large set of spatial objects into a number of spatially contiguous regions while optimizing an objective function and satisfying a set of constraints (such as spatial contiguity, minimum region size in terms of population or other attribute values). The objective function can be a homogeneity (or heterogeneity) measure or other statistics of derived regions. Regionalization is needed to generalize large and highly detailed data into a set of appropriate-sized regions, which preserve original patterns while reducing random effects and protecting privacy. In the era of Big Data, regionalization can be used for data processing and analysis in a wide range of research and application domains, for example, detecting natural neighborhoods, data dissemination, market segmentation, housing market area delineation, insurance rate zoning, climatic

zoning, eco-region analysis, hazards and disasters management, agriculture data mapping, map generalization, location optimization, health data analysis, etc.

Regionalization outperforms existing manual and automated approaches in a number of important aspects, including (1) its superior optimization power in maximizing objective functions and meeting constraints; (2) it is efficient in processing large data; (3) its results are completely reproducible (i.e., different from many existing approaches that rely on random initialization and searching); and (4) it is easy to use with integrated user interface, data processing, and visualization techniques, allowing easy configuration, comparison, and interpretation



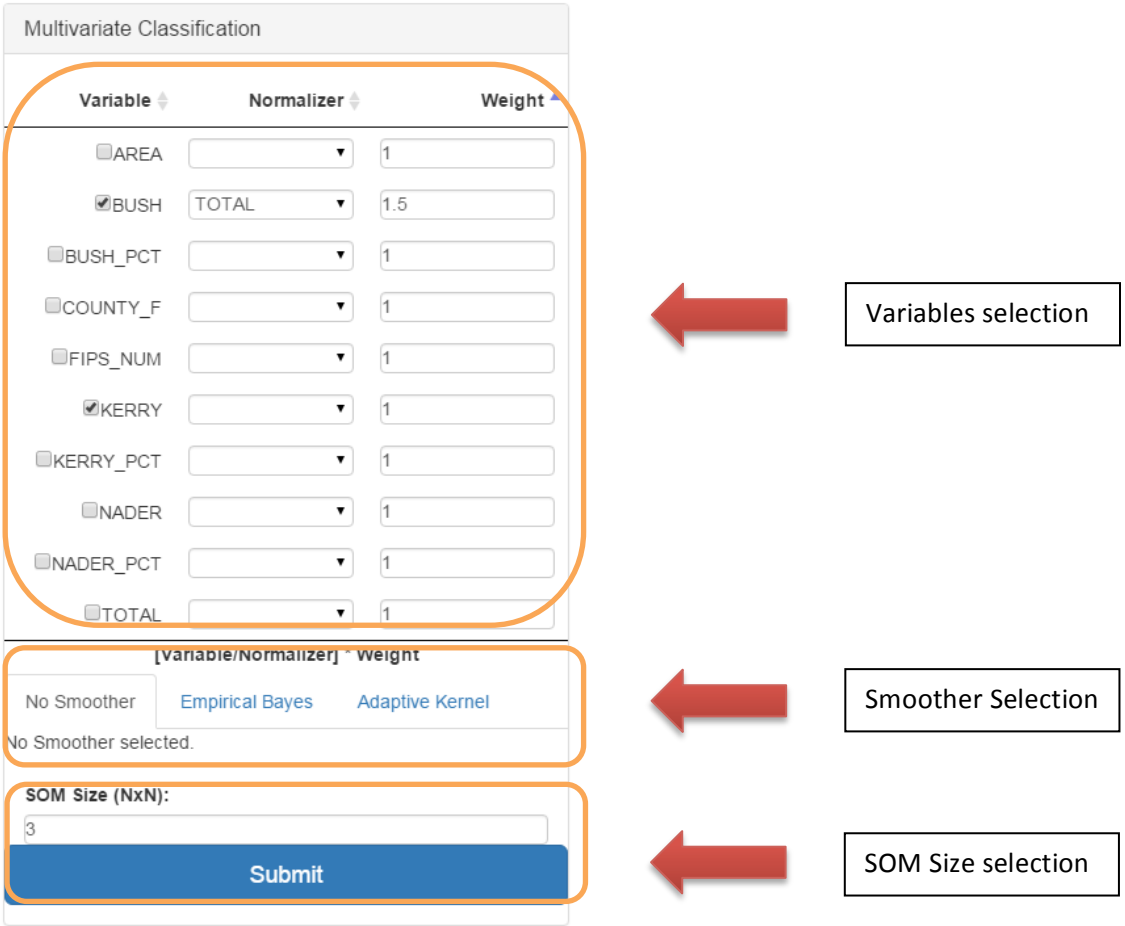
Multivariate Mapping

1. *Choose Variables*

- Click the checkbox to select variables
- Multiple variables can be selected.
- All of the variables can be normalized.*
- User can assign weights for selected variables. **

***Normalizers are used to reduce variables originally measured on different scales to a standard scale using a common factor to make them easily comparable. Select a variable that contains a scale you wish to normalize with such as total population, total votes, etc. Any variable can be used as long as there are no negative or zero values.**

****The weight is the importance factor in comparison to other factors. The default weight is 1. Any number between 1 and 100 can be used.**



2. *Choose a Smoother*

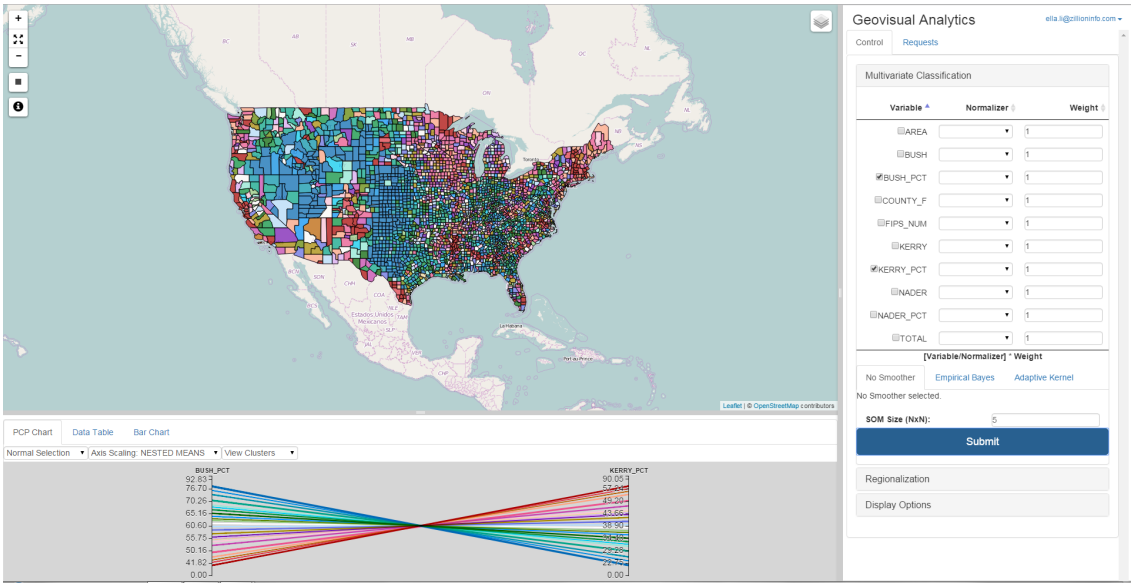
Smoothing data is required to help analysts detect trends and patterns in widely distributed data. Smoothing eliminates outliers from the majority of data items essentially "smoothing" the data to fit into a scale it might otherwise not have fit into. To smooth a data set is to create an approximating function that attempts to capture important patterns in the data, while leaving out noise or other fine-scale structures/rapid phenomena. **This is entirely optional. You do not have to use a smoother to get optimized results.** Determining which smoother to use depends on the range of your data set. See smoother options in Appendix I.

3. *Choose SOM Size*

Enter the number of SOM size. The SOM size determines how many clusters will be calculated. For example, because the SOM dimensions are N x N, if the number 3 is chosen 9 total clusters will be calculated.

4. *Submit*

Click "**Submit**" and wait for the classification to be calculated and be shown in the map and in the PCP.



2.2 Regionalization

Regionalization is to divide a large set of spatial objects into a number of spatially contiguous regions while optimizing an objective function, which is normally a homogeneity (or heterogeneity) measure of the derived regions.

This service uses a variety of clustering (grouping) methods designed to determine the distance, or differences, between data points based off of a chosen set of constraints for optimal clustering of data into regions.

Steps for Configure Regionalization Algorithm

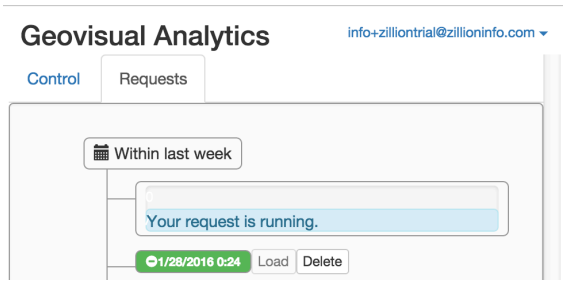
1. ***Use smoothed or original rate:***
   - Smoother= Empirical/Adaptive Kernel
   - If you did not choose a smoother in the Multivariate Classification panel then this is not needed
   - **"Smoother" + Original:** use smoothed rate to construct the hierarchical tree and original rate to partition the tree
   - **"Smoother" + "Smoother**": use smoothed rate to construct and partition the tree
   - **Original + Original**: use original rate to construct and partition the tree

2. ***Choose a regionalization method:***
   There are several regionalization methods provided in this service (See Appendix II). **ALK** is recommended.

3. ***Type a maximum number of regions***
   - A hierarchy of that many regions will be calculated. E.g. if 10 is set, a hierarchy of 10 regions will be calculated.

4. ***Chose a control population(Optional):***
   - Choose a control variable (e.g. # of features, population, or other variables in your dataset)
   - Set control threshold (e.g. # of features in a region >3, population > 5000, etc.)
   - If you do not need a constraint use # of features and type in "0"

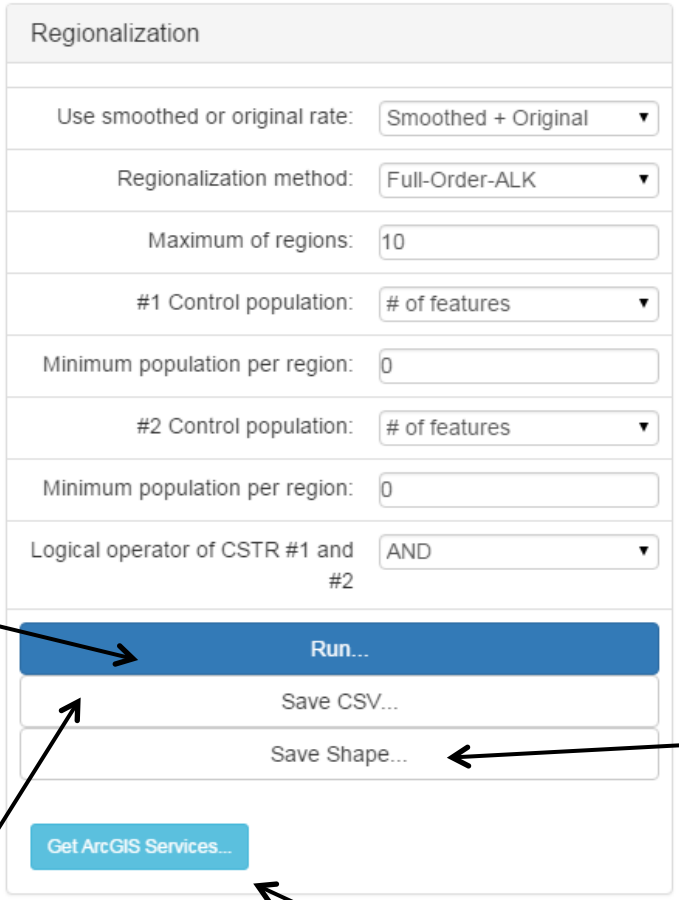
5. ***Click "Run"***
   A regionalization calculation is running on the server side. You can see the calculation status as shown below. At this point, you don't have to wait for the calculation result. You may do other work or even close the browser. The result will be shown in the Request list. You click "Load" to load the calculation result to the Map and PCP



a.

b. Save the results

- After it finishes save the regionalization result by clicking “Save CSV” (saves as a .csv file) or Clicking ”Save Shape…”(saves as a .shp file).
- The result can also be uploaded to ArcGIS Services to display within ArcGIS online



Compute regions

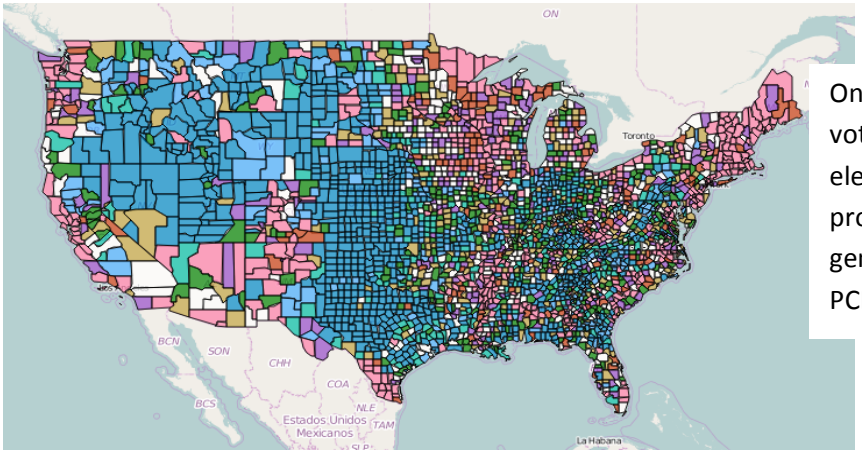
Save regions as shapefile

Export as .csv file. This will allow you reload a regionalization file later.

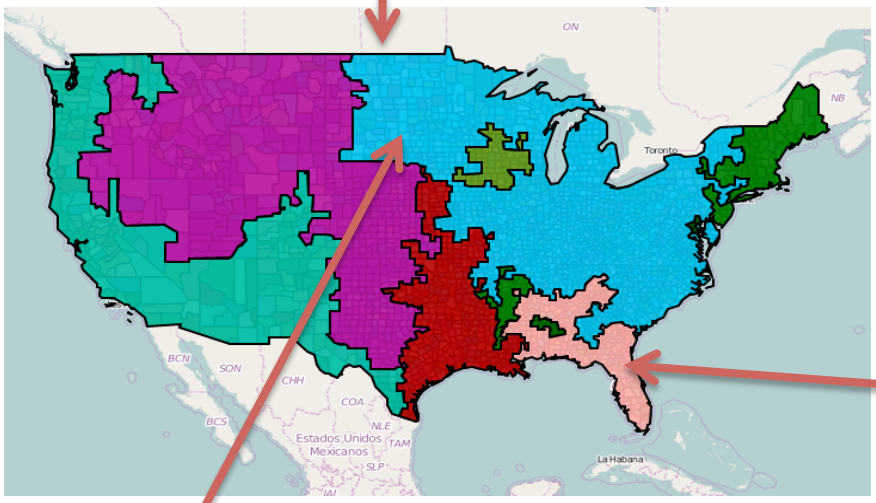
Connect with ArcGIS services

Regionalization Result

After the algorithm has been run, the number of regions (specified earlier) is computed into the map window. Regionalization groups together areas with similarities to make it easier to detect patterns. From here generalizations can be made by using the PCP as illustrated in Appendix III.

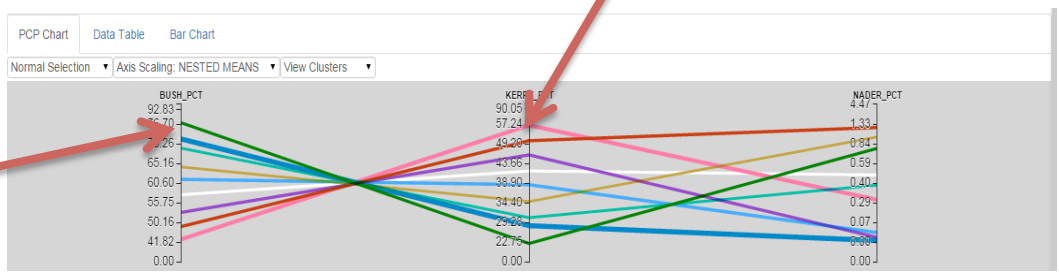


One can get an overall view of the voting patterns during the 2004 elections. The regionalized map provides evidence that the generalizations made using the PCP were true.



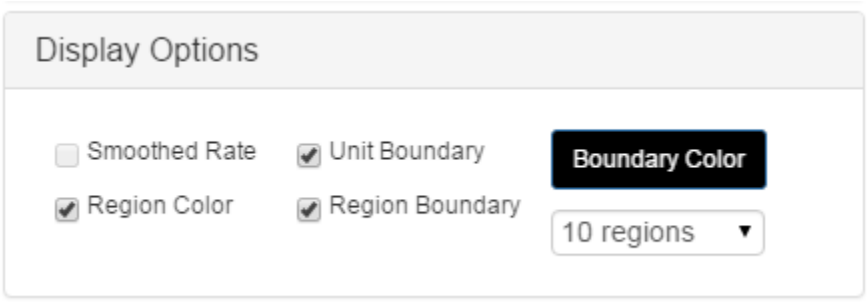
One can easily determine where Kerry dominated the polls

One can easily determine where Bush dominated the polls



2.3 Display Options

Click "Display Options" to adjust setting of map display.

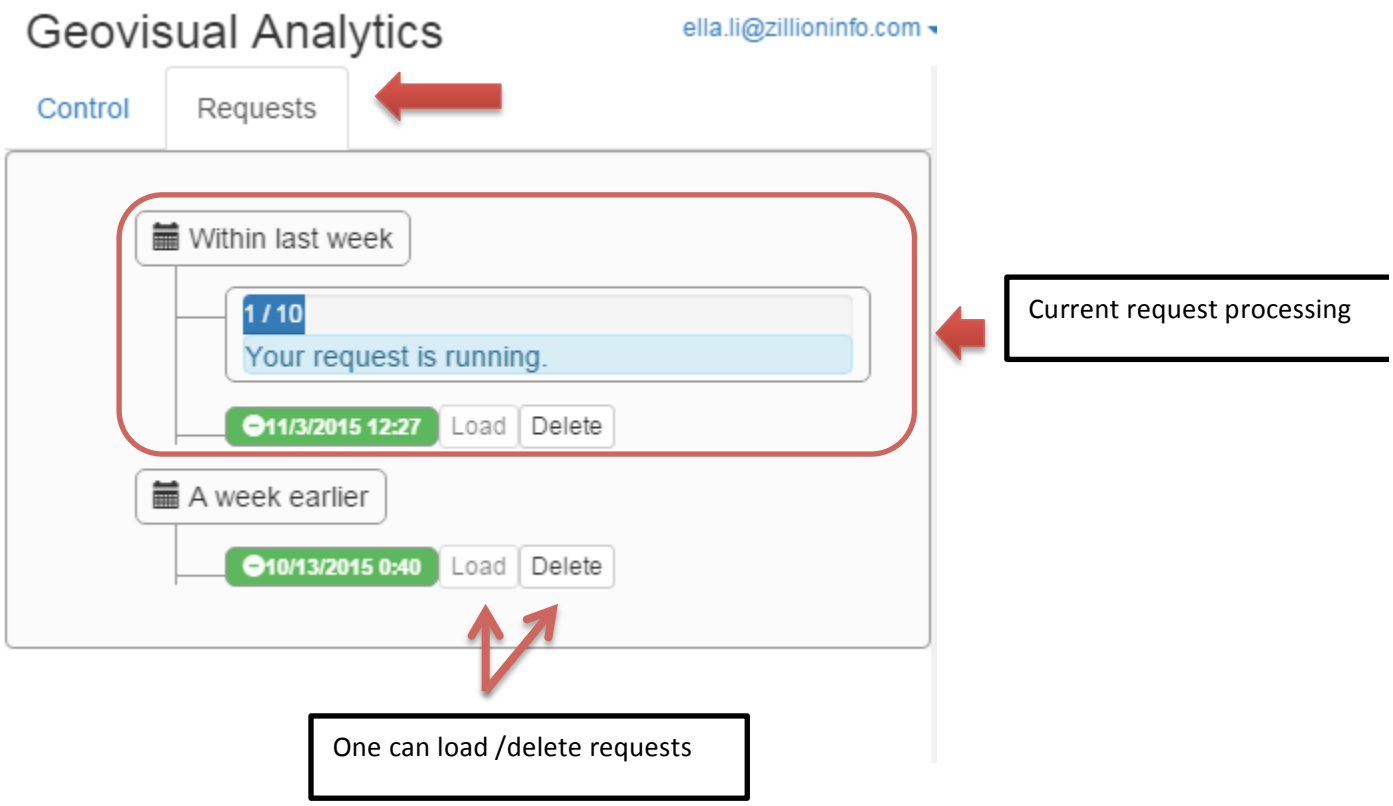


Result Option	Description
Smoothed Rate	Turn the smoothed rate on or off for each unit
Unit Boundary	Turn the unit boundary on or off to better view region boundaries
Region Color	Turn the region colors on or off
Region Boundary	Turn the region boundaries on or off
Boundary Color	Change the color of the region boundaries
Number of regions	Select the number of regions that are visible on the map.

2.4 Request

After clicking “Run” in the regionalization panel, the requests tabs will allow users to view the status of the regionalization process currently being run. To upload the regionalization result into the map display click “Load”.

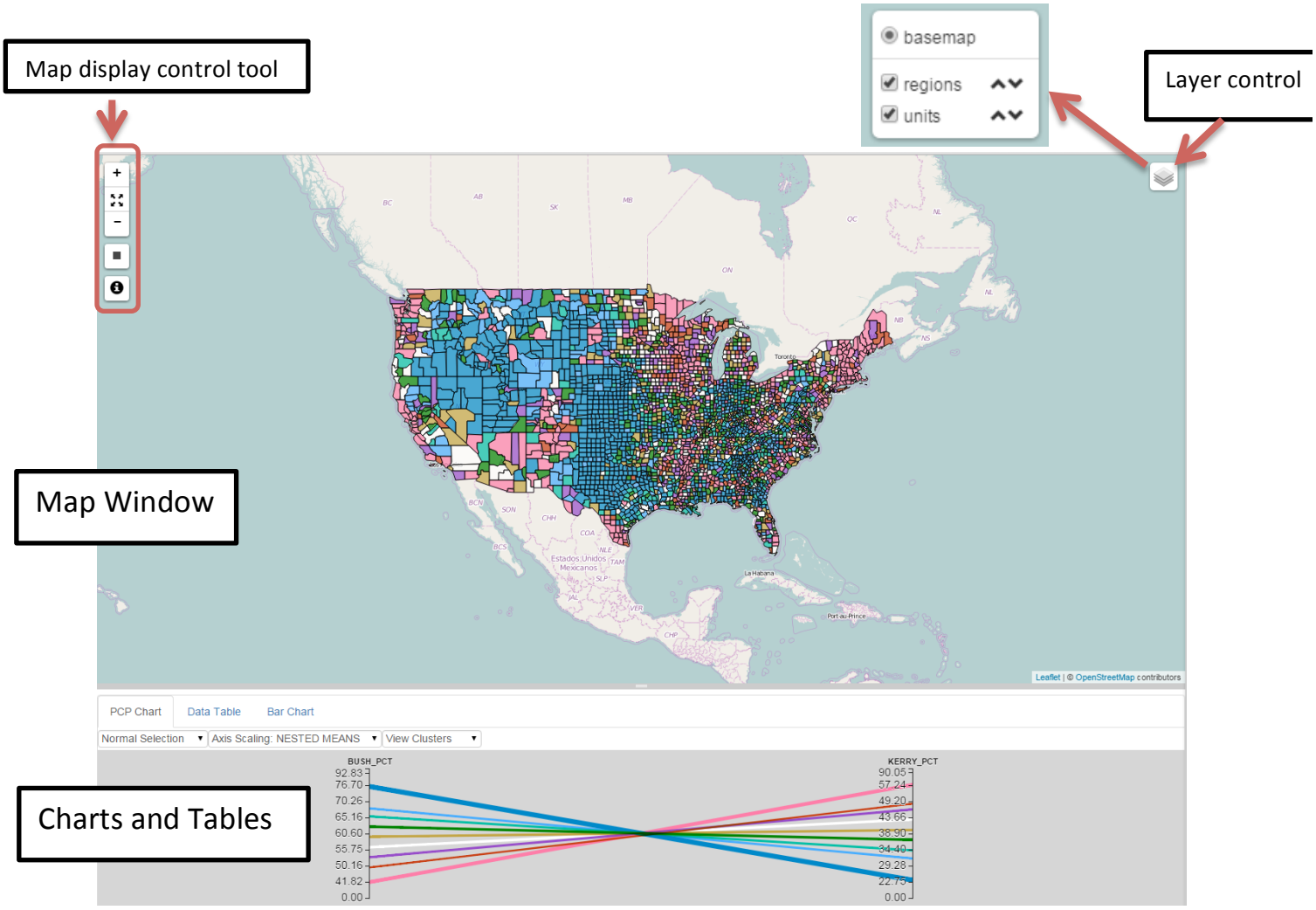
Past requests (regionalizations) can also be viewed below the current request. All requests can be loaded or deleted at any time.



2.5 Display

2.5.1 Map

The results of the classification and regionalization can be viewed in the map window. The map display can be used as a control tool to change the map display (see details below). Also, map layers can be adjusted by clicking the layer control button . Use to adjust order of layers.



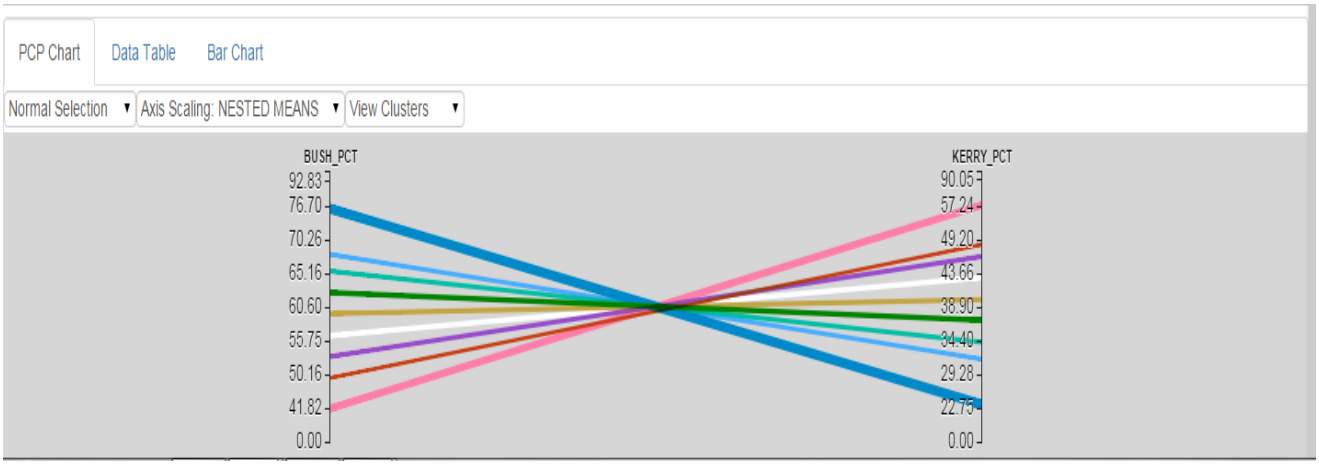
Function	Description
Zoom In	Zoom in on areas of the map – you can also use the mouse wheel to zoom in by scrolling up.
Default Extent	View the targeted area and get a full view of your dataset
Zoom out	Zoom out to areas of the map – you can also use the mouse wheel to zoom out by scrolling down.
Zoom to Selection	Zoom in on selection areas of the map. Click and drag rectangle to selection.
Identify	Identify all attributes for each polygons

2.5.2 Charts and tables

There are three charts and tables to visualize data and results**: PCP chart, Data Table, and Bar Chart**.

PCP Chart:

Click "PCP Chart" to check the Parallel Coordinate Plot (PCP). Please check Appendix III for more details.



Data Table:

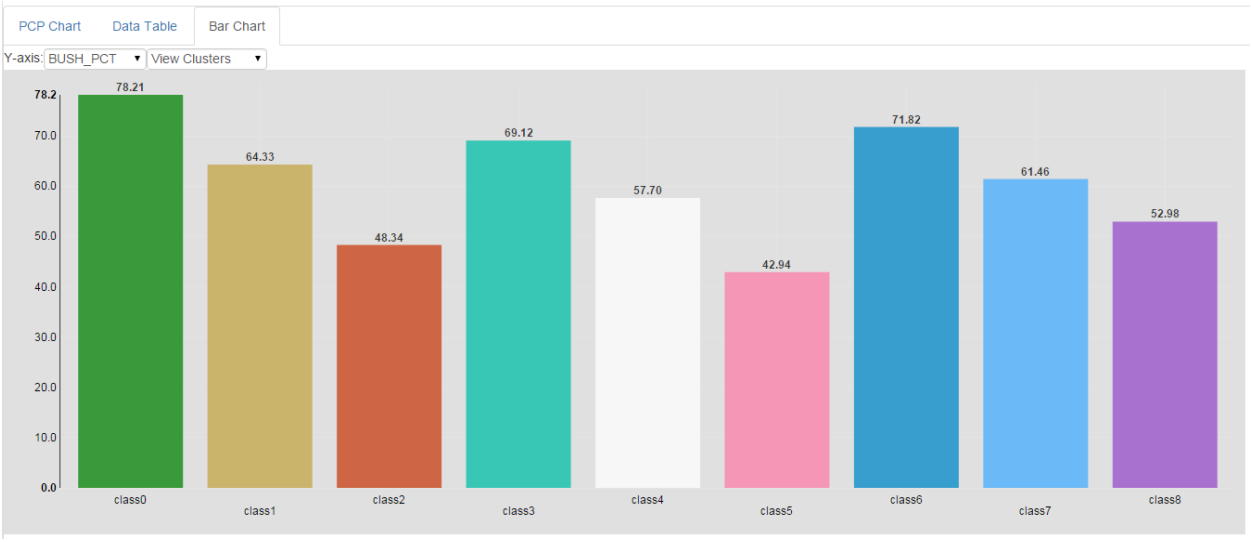
Click "Data table" to check details of classification by table. The mean value of each class will be displayed in the data table.

PCP Chart | Data Table | Bar Chart

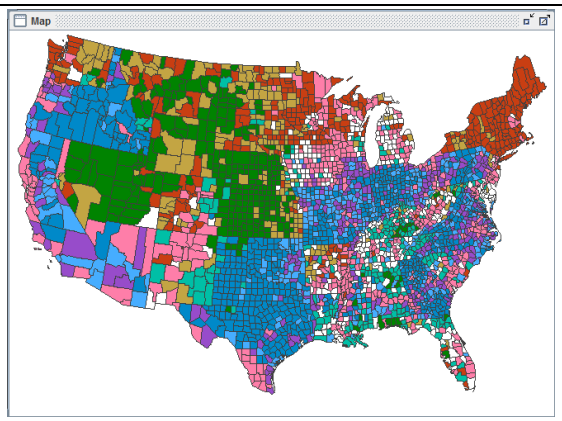
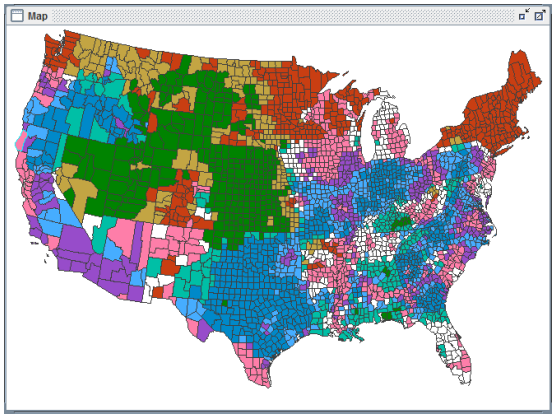
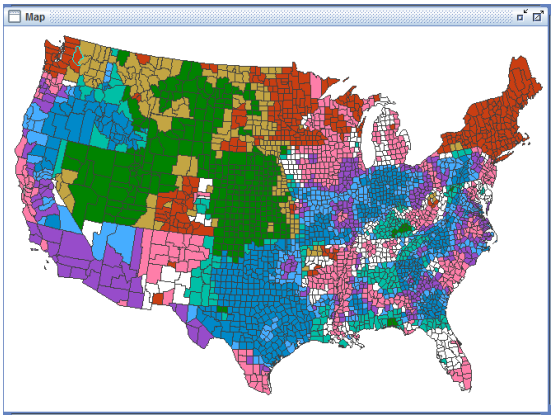
BUSH_PCT ▲	KERRY_PCT	NADER_PCT	class
42.94	56.75	0.31	class5
48.34	50.41	1.25	class2
52.98	46.04	0.02	class8
57.70	41.82	0.48	class4
61.46	38.51	0.03	class7
64.33	34.66	1.01	class1
69.12	30.49	0.39	class3
71.82	28.17	0.01	class6
78.21	21.01	0.78	class0

Bar Chart:

Click “Bar Chart” to display details of classification by bar chart. Any attributes previously selected in the control panel can be represented on the Y-axis by using the dropdown menu. Choose “View Clusters” to identify the attributes value in each class or choose “View Data Items” to identify the attributes value for each item. The mean value of each cluster or data item will be displayed.



Appendix I: Smoothing Methods

Smoother	Description	
No smoother	Data will remain in its original set. Outlier and widely distributed data will be treated as such. Selecting "No smoother" caused the clustering of data items to be slightly widespread. More often than not an area with a dominant cluster(color) is speckled with other clusters	
Empirical Bayes	The minimum population could be used to control the "neighbors" which are used to smooth data. The default value of minimum population is 0, which means **only first-order neighbors will be used to smooth data.** Now the map has much more contiguous clusters and significantly less speckling of clusters within those areas	
Adaptive Kernel	Two parameters could be set. One is the number of nearest neighbors used to smooth data. Another is model type. Selecting Adaptive Kernel led to much more contiguous clusters than empirical bayes or not smoothing data at all.	

Appendix II: Regionalization Methods

Regionalization Method	Description
WARD	the distance between two clusters is how much the sum of squares will increase when the two clusters merge together into a region
Full-Order ALK	Average Linkage: the distance between two clusters is the average dissimilarity between cross-cluster pairs of data points Depends on numerical scale on which the differences are measured
Full-Order CLK	Complete Linkage: the distance between two clusters is the dissimilarities between the furthest pair of data points Clusters are considered similar only if all the observations in the two clusters are similar to each other
Full-Order SLK	Single Linkage: uses distance between two clusters as the dissimilarity between the closest pair of data points from each cluster Groups points linked by a series of close intermediate points Full-Order: considers all edges in the clustering process therefore is dynamic
First-Order SLK	Single Linkage: uses distance between two clusters as the dissimilarity between the closest pair of data points from each cluster Groups points linked by a series of close intermediate points First-Order: only considers edges from beginning; stays static.

Appendix III: Parallel Coordinate Plot (PCP)

The PCP supports two detail levels: Clusters vs. Data Items

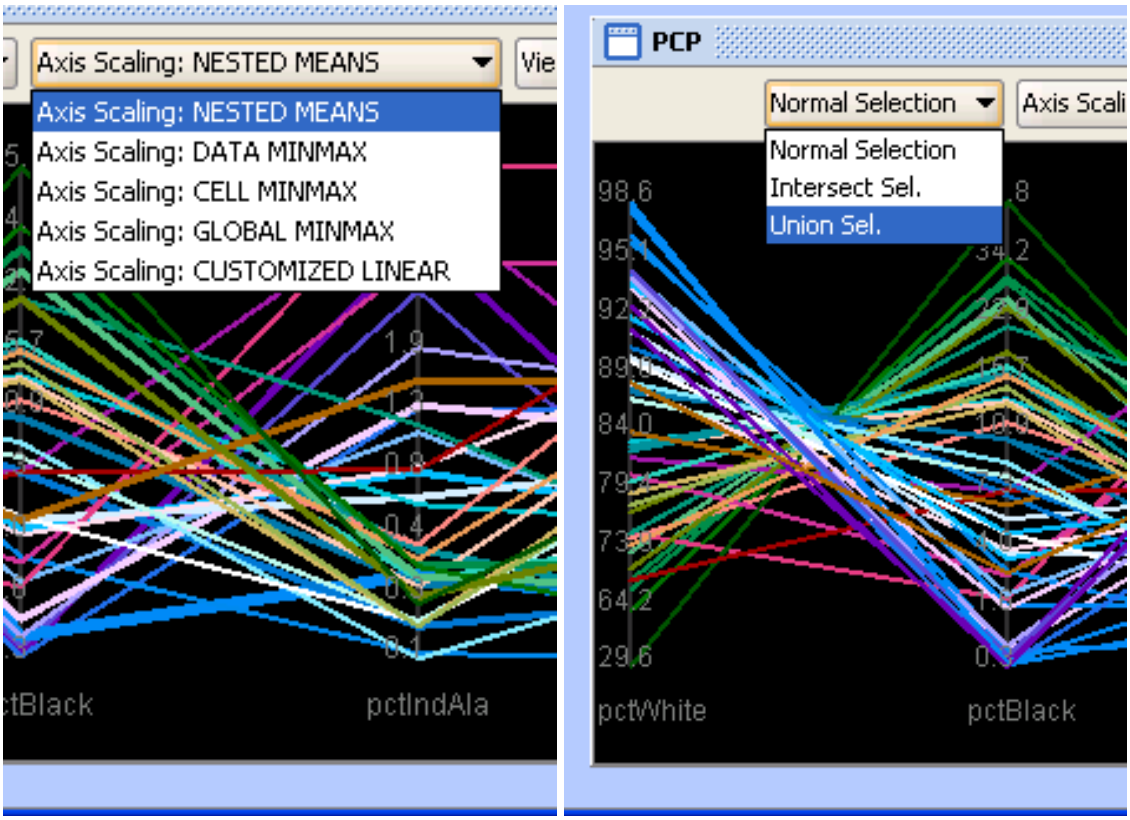
View Clusters: Each string represents a cluster with its mean vector. The thickness of each string represents the cluster size (i.e., the number of data items in the cluster).

View Data Items: Each string represents a data item with its multivariate vector.

For either choice, each cluster (or data item) has the same color as it does in the SOM view.

The PCP supports five axis-scaling methods:

- **Nested-Means:** scaling on each axis using nested means and thus adjust the spacing of intervals according to data distribution. This method can alleviate the overlapping problem in PCP for skewed data distribution. Specifically, nested-means is a *non-linear scaling* method that recursively calculates a number of mean values (and sub-means) and uses these values as break points to divide each axis into equal-length segments. Therefore, nested-means scaling always puts the mean value at the center of each axis and thus makes axes defined by different units and data ranges comparable.
- **Data Min-Max:** each axis is linearly scaled using its min and max values.
- **Cluster Min-Max** (or Cell Min-Max): each axis is linearly scaled using cluster centroid min and max values.
- **Global Min-Max:** this option is only useful when all the variables are comparable to each other, for example percentage values. Axes will be scaled linearly using the global min and max values (for all variables).
- **Customized Linear:** this option is only useful when all the variables are comparable to each other, for example percentage values. The user will define the min and max (same for all variables) to linearly scale each axis. In future versions, the user may be able to define the min/max differently for each axis.



The PCP can *optimally* order dimensions:

Optimal Ordering of Axes: dimensions are ordered using an optimal hierarchical ordering method based on the mutual correlations among dimensions

Original Ordering of Axes: dimensions are in their original order as in the data file.

The PCP supports different types of selection at different levels:

The PCP at the cluster level presents a global view of the overall patterns. A user can select one or more clusters in the PCP (or in the SOM), then switch to the data item level (instead of the cluster level), and examine all the data items in the cluster(s).

Selection can also be made at the data item level. For example, one can show data at the item level in the PCP and then select a single data item to read its exact variable values. One can also switch back to the cluster level and see which cluster the selected item belongs to. If that cluster also contains other items, its circle will become a wedge to show the partial selection.

By selection “Intersect Sel.” or “Union Sel.”, the user may also combine two different selections or select within a selection.

Interacting with the PCP

Function	How to:
Select Clusters	1. Use your mouse to **click and drag vertically** to select clusters of interest **2.** When selecting with mouse, **do not drag horizontally across vertical axes** 3. When features are selected they remain colored and all other clusters turn to gray(default)
Select within a selection	1. Use the dropdown menu in the top left-hand corner and click "Normal Selection" 2. **Click and drag vertically** to select clusters of interest 3. After selected, use the dropdown menu **and click "Intersect Sel."** 4. You can now click and drag vertically to select clusters within the selection already made.
Combine two different selections	**1.** Use the dropdown menu to **click "Normal Selection"** 2. **Click and drag vertically** to select clusters of interest **3.** After selected, use the dropdown menu and **click "Union Sel."** 4. Click and drag vertically over other clusters you did not previously select to add to the selection.
Combine two different selections **(Different Way)**	**1.** Use the dropdown menu to **click "Normal Selection"** 2. Click and drag vertically to select clusters of interest 3. After selected, **hold down shift key** AND 4. **Click and drag vertically** over other clusters you did not previously select to add to the selection.
Clear Selection	1. Click and drag vertically **over gray space** within PCP 2. All clusters will become unselected