Data Analysis, Statistics, Machine Learning

Leland Wilkinson

Adjunct Professor
UIC Computer Science
Chief Scientist
H2O.ai

leland.wilkinson@gmail.com

Grouping

- We can create groups of variables or groups of cases
- These methods involve what we call Cluster Analysis
  - Hierarchical methods make trees of nested clusters
  - Non-hierarchical methods group cases into *k* clusters
    - These *k* clusters may be discrete or overlapping
- Two considerations are especially important for hierarchical
  - Distance/Dissimilarity measure
  - Agglomeration or splitting rule
- The collection of clustering methods is huge
  - Early applications were for numerical taxonomy in biology

Grouping

Non-hierarchical clustering

k-means

$$SSW = \sum_{j=1}^{k} \sum_{i=1}^{n_j} (\mathbf{y_{ij}} - \mathbf{m_j})^{\mathbf{2}}$$

sum of squares within clusters

We seek to minimize *SSW*

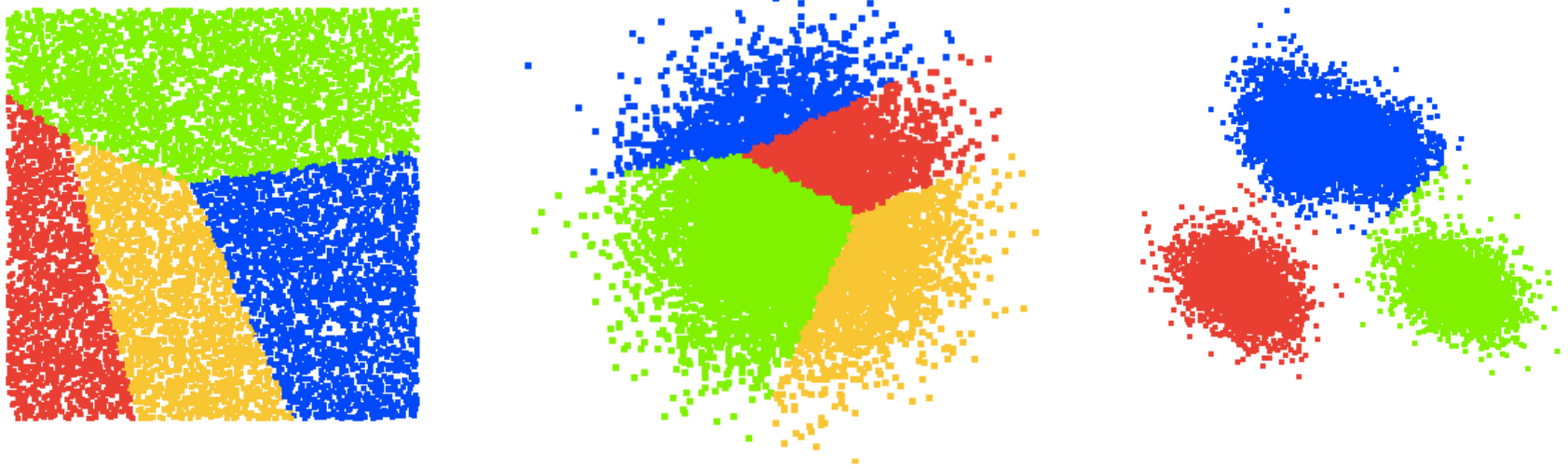
1. Choose *k*
2. Initialize *k* centroids
3. Assign every point y to nearest centroid (squared Euclidean distance)
4. Compute *SSW*
5. Repeat 3 and 4 until *SSW* does not improve (get smaller)

Notice similarity to ANOVA/MANOVA

k-means is a family of algorithms – many variations

Grouping

- Non-hierarchical clustering
  - Geometry of k-means
    - Voronoi Tessellation
    - Clusters are convex



Grouping

Aspects of the algorithm

Quantization

Even if there are no clusters, *k*-means can be used to partition

Detection of outliers

Compute outliers relative to cluster centroids

Clusters are convex

If a case is a weighted average of cases in the cluster, the case also lies in the cluster

k-means demo

https://www.cs.uic.edu/~wilkinson/Applets/cluster.html

Grouping

Non-hierarchical clustering

Picking *k*

Run *k*-means from 1 to k_{max}

Plot *SSW* against index and look for elbow

Or, (Hartigan, 1975) SSW is distributed approximately as chi-square

So are differences between SSW_{k-1} and SSW_k

Examine chi-square statistics for these differences

(Hamerly and Elkan did something like this in 2003)

They didn't seem to be aware of Hartigan

Or, compute a measure of clumpiness from minimum spanning tree (MST)

Hartigan RUNT statistic or Stuetzle & Nugent

Grouping

- Non-hierarchical clustering
  - Picking starting centroids
    - Random centroids (ugh!)
    - Hartigan
      - run *k*-means from 1 to *k*
      - split on variable with largest variance from previous solution
  - Evaluating *k*-means solutions
    - Can do F-tests on each variable (forget about *p* values)
    - Can do MANOVA on groups (you're kidding, right?)
    - Use visualization (we'll talk about that later)

Grouping

Spectral clustering

1. Compute k-nearest-neighbor adjacency matrix A
2. Choose bandwidth *t*
2. Compute heat kernel D matrix, $d_{i,j} = \exp(-a_{i,j} * a_{i,j}/t)$
3. Compute Laplacian matrix $\mathbf{L} = \mathbf{D} - \mathbf{A}$
4. Compute eigendecomposition of L
5. Cluster last two (or more) eigenvectors of L

This algorithm is due to Belkin and Niyogi

There are variations

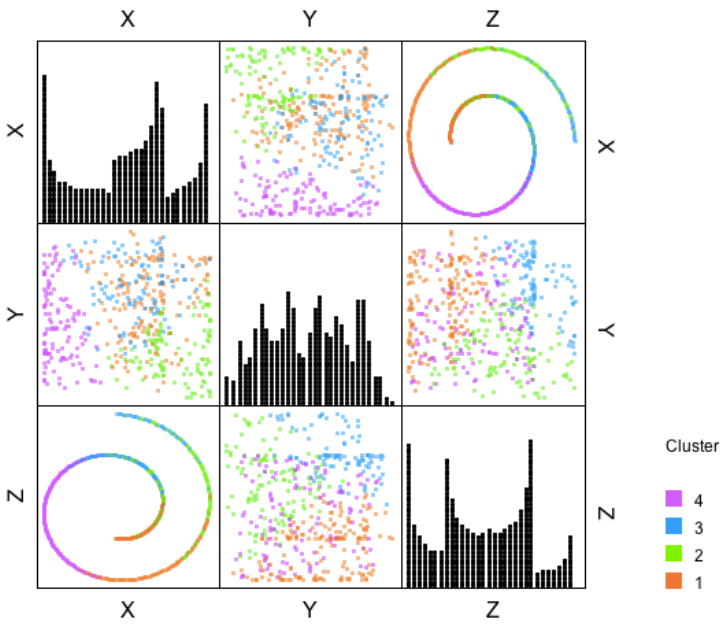
Spectral clustering works on nonlinear projection of points

So it can deal with cluster shapes that are intertwined

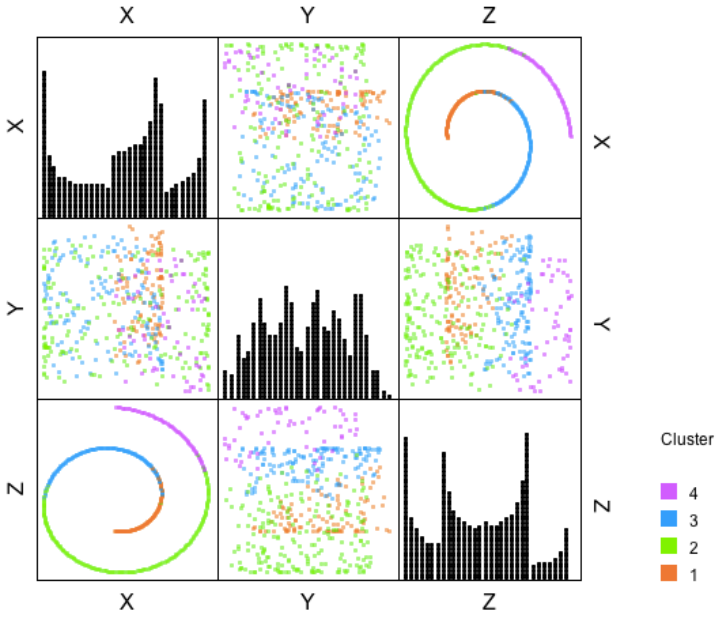
Grouping

Spectral clustering

Regular k-means



Spectral



Grouping

Hierarchical clustering (agglomerative)

1. Start with each point in a cluster of its own
2. Until there is only one cluster
   (a) Find the closest pair of clusters
   (b) Merge them

Grouping

Hierarchical clustering

Two essential components

1. **Distance** method for computing distance between points
2. **Linkage** (amalgamation) method for computing distance between clusters

There are many versions of each, so the number of possible algorithms is enormous.

Strengths

- You can use any distance method
- Works well with non-convex clusters
- Hierarchical, so clusters do not have to be disjoint

Weaknesses

- A pig in both space and time
- There are workarounds, but they are not pretty

Grouping

Hierarchical clustering

Distance methods

Euclidean $\sqrt{\sum_i (x_j - x_k)^2}$

City block $\sum_i |x_j - x_k|$

Chebyshev $\max |x_j - x_k|$

Cosine $1 - \frac{\mathbf{x}_j \cdot \mathbf{x}_k}{\|\mathbf{x}_j\| \|\mathbf{x}_k\|}$ (1 – Pearson R)

Jaccard $1 - \frac{|A \cap B|}{|A \cup B|}$

. . .

Grouping

Hierarchical clustering

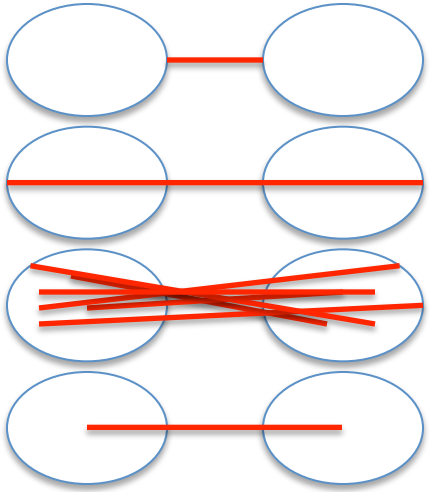
Linkage methods

Single $\min\{d_{a,b} : a \in A, b \in B\}$

Complete $\max\{d_{a,b} : a \in A, b \in B\}$

Average $\sum_{a \in A} \sum_{b \in B} d_{a,b}/(n_a n_b)$

Centroid $\|c_A - c_B\|$



. . .

Grouping

Hierarchical clustering

Linkage methods

Single	long, stringy clusters
Complete	convex clusters having comparable diameters
Average	clusters with small variance joined first
Centroid	slightly more robust to outliers

Boris Mirkin has developed a full continuum between Single and Complete

. . .

Grouping

Hierarchical clustering

Ward's method

The distance between two clusters, A and B, is how much the sum of squares within clusters will increase when we merge them

Choose minimum merging distance for next join

Similar to *k*-means in that it favors convex clusters

Euclidean distances with Ward's method yields solutions similar to *k*-means

Sensitive to outliers and departures from normality

Joe Ward has a high-school named after him

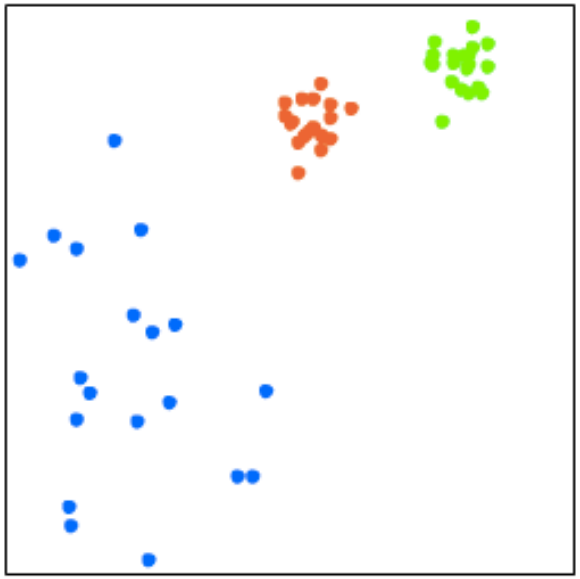
He is passionate about helping secondary students learn statistics

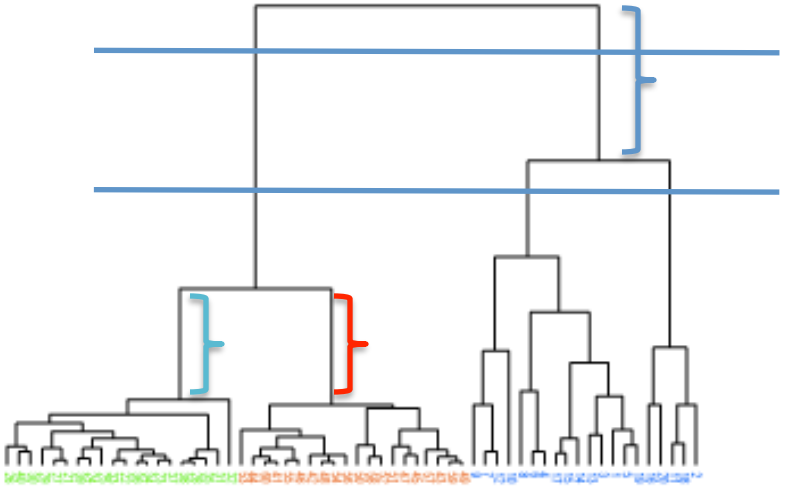
Grouping

Hierarchical clustering

Choosing the number of clusters

- Can't just cut the tree at some distance
- Must consider separation of density modes
  - Look at length of branch above merged node





Grouping

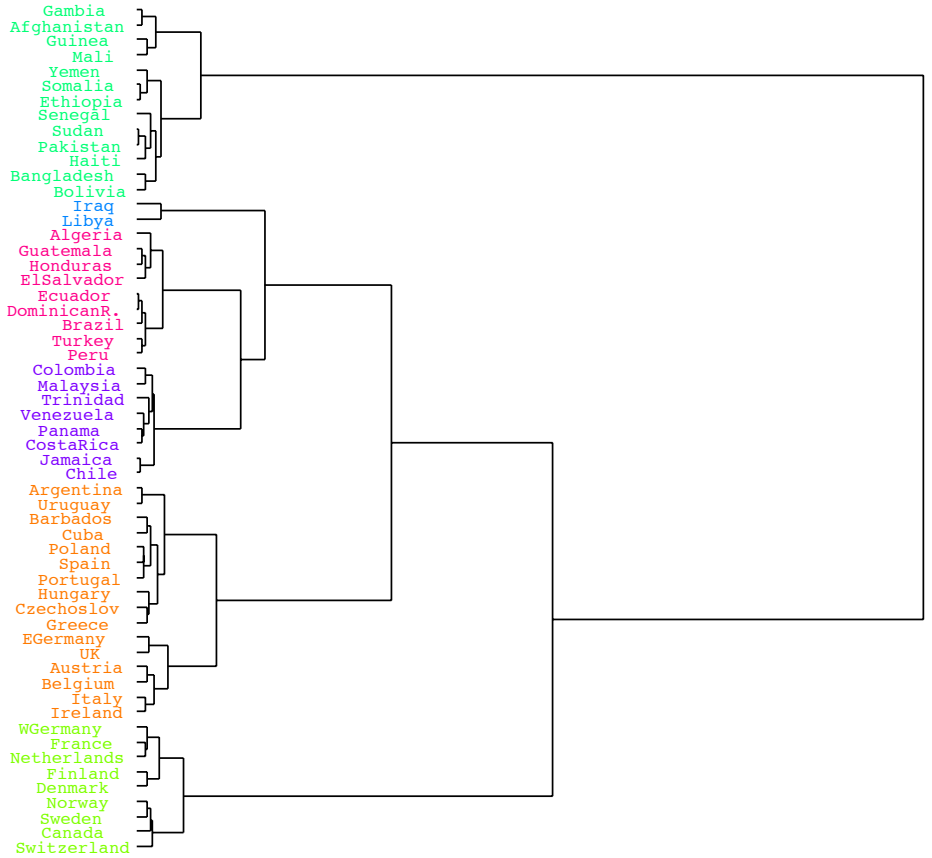
- Hierarchical clustering
  - Choosing the number of clusters
    - Can compute margin to other clusters
    - Silhouette method of Rousseeuw works as well
      - Includes tightness as well as separation
    - Methods based on Hartigan RUNT statistic work well too (Stuetzle)
    - Forget SS(Within) methods unless you have Gaussian clusters
    - Forget looking for elbow in SS plot
      - You never see them with real data

Grouping

Visualizing clustering

Trees

The distance between two leaves is proportional to the position of the closest node joining them.



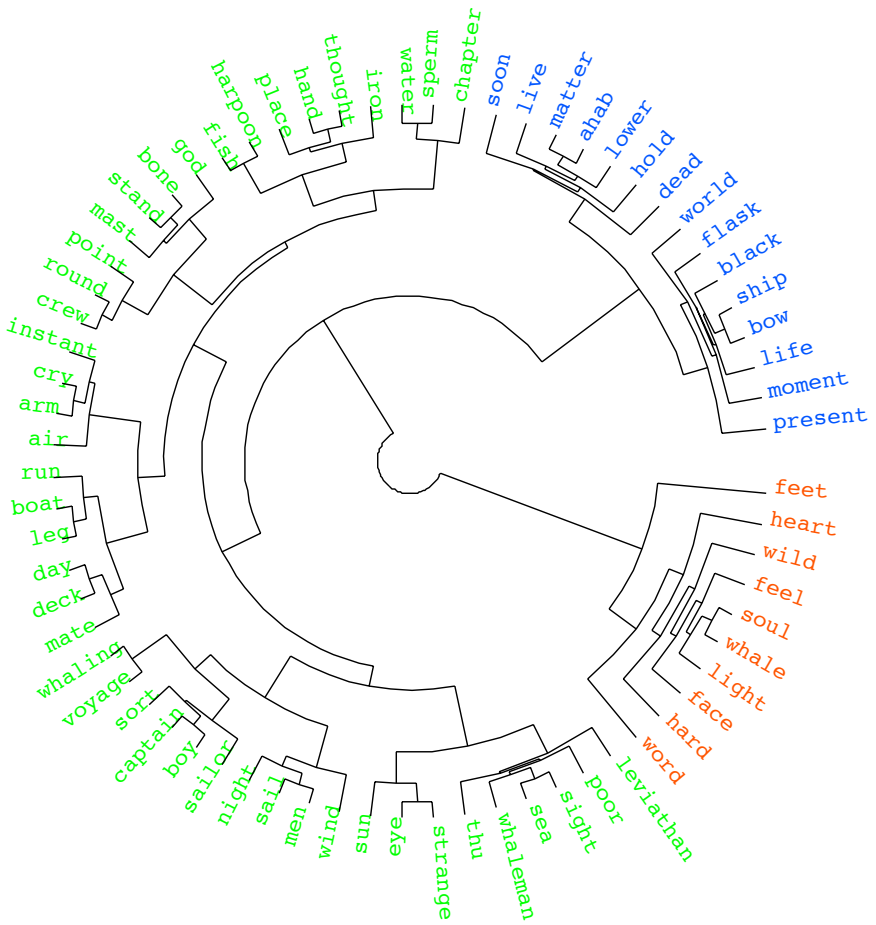
⬆ 0 distance … greatest distance ⬆

Grouping

Visualizing clustering

Polar Trees

Useful when there are numerous leaves

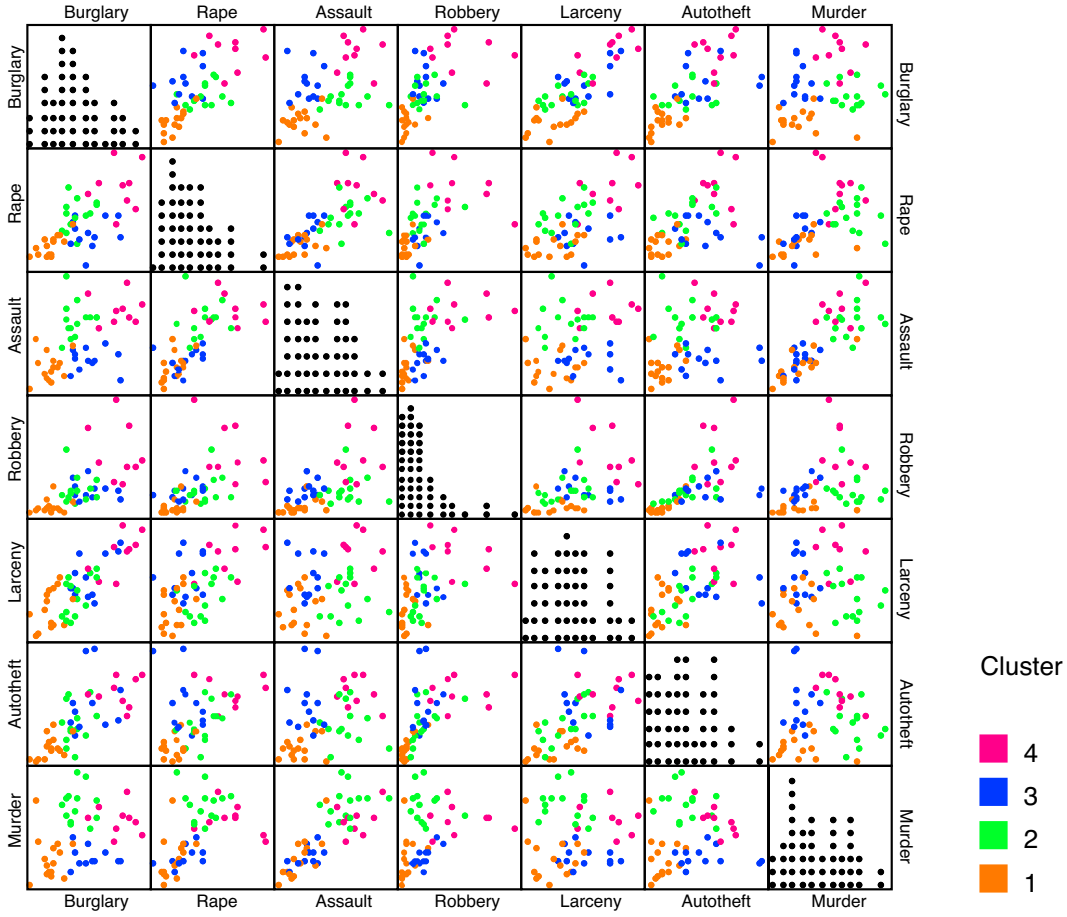


Grouping

Visualizing clustering

SPLOM

Coloring can reveal which projections have greatest cluster separation

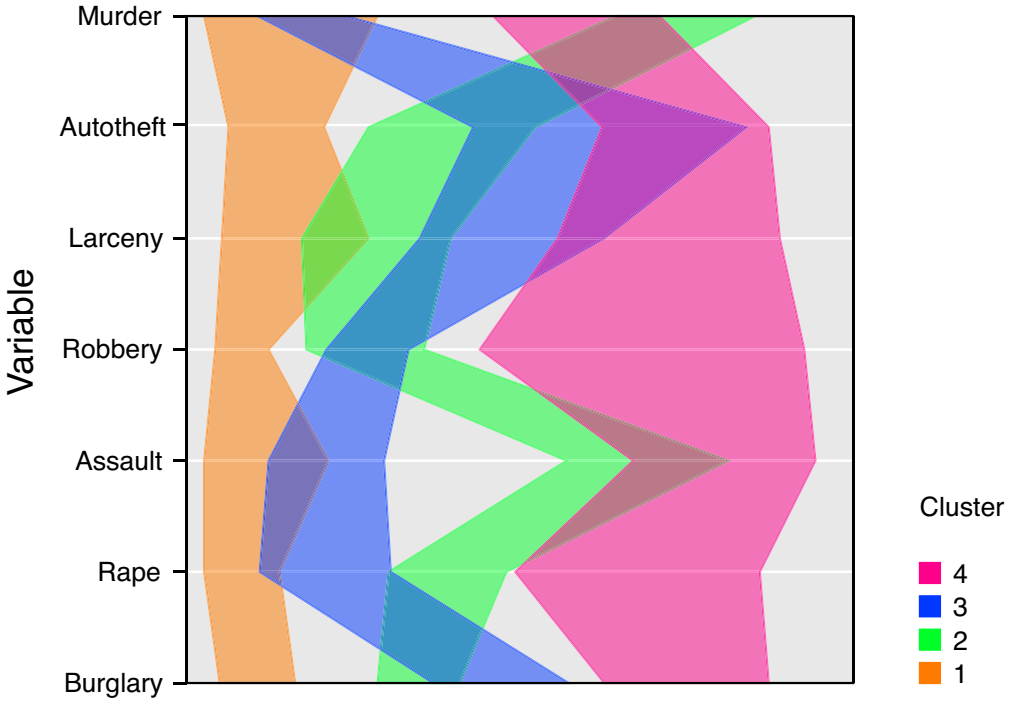


Grouping

Visualizing clustering

Profiles

The bands represent confidence intervals on each variable in each group. More informative than parallel coordinates.

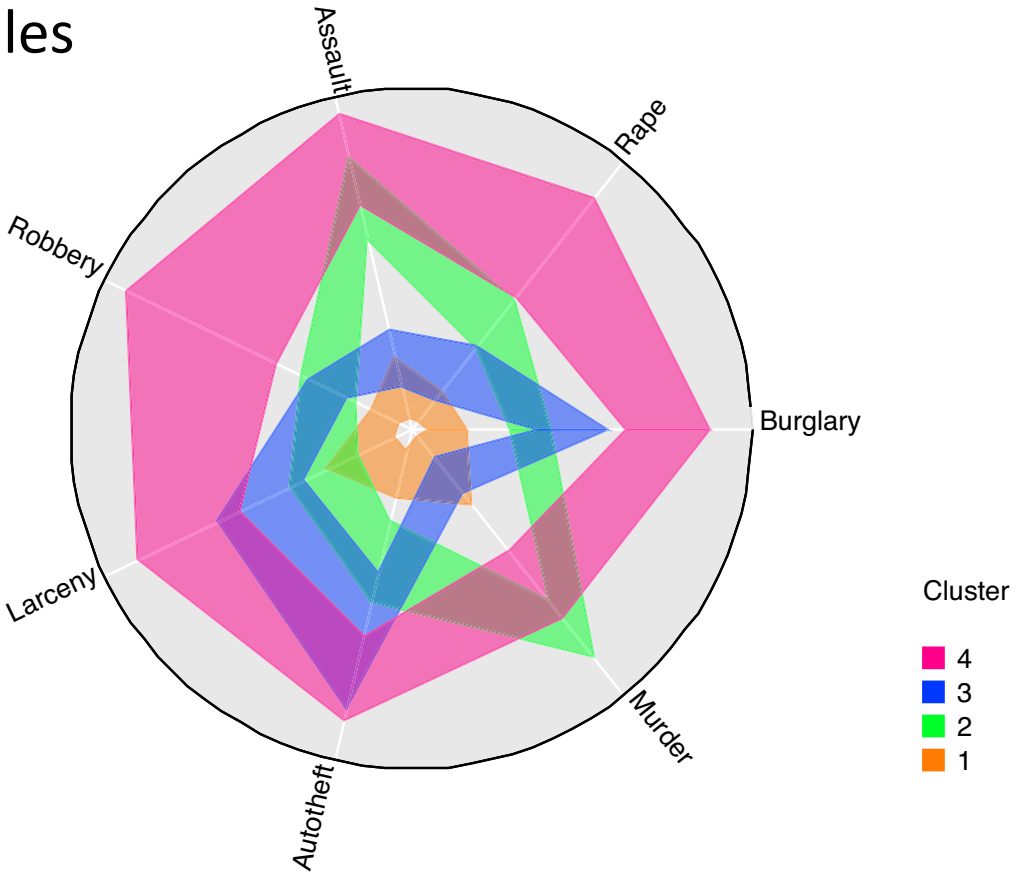


Grouping

Visualizing clustering

Polar Profiles

Good for many variables

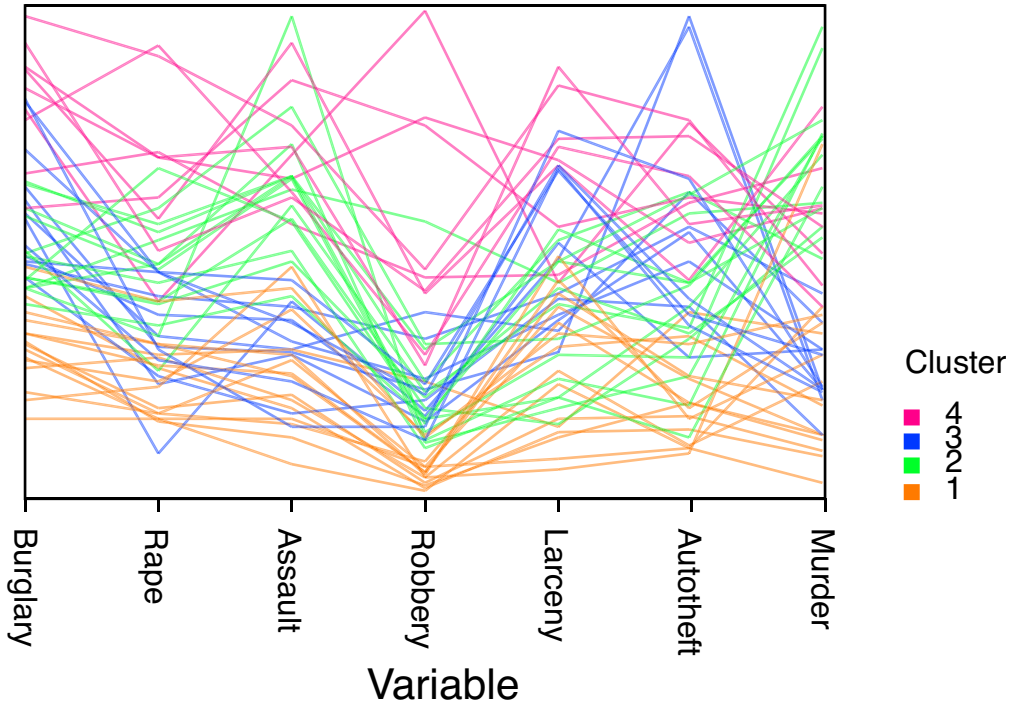


Grouping

Visualizing clustering

Parallel Coordinates

A poor cluster display method unless clusters are tight. Confidence bands are more useful.

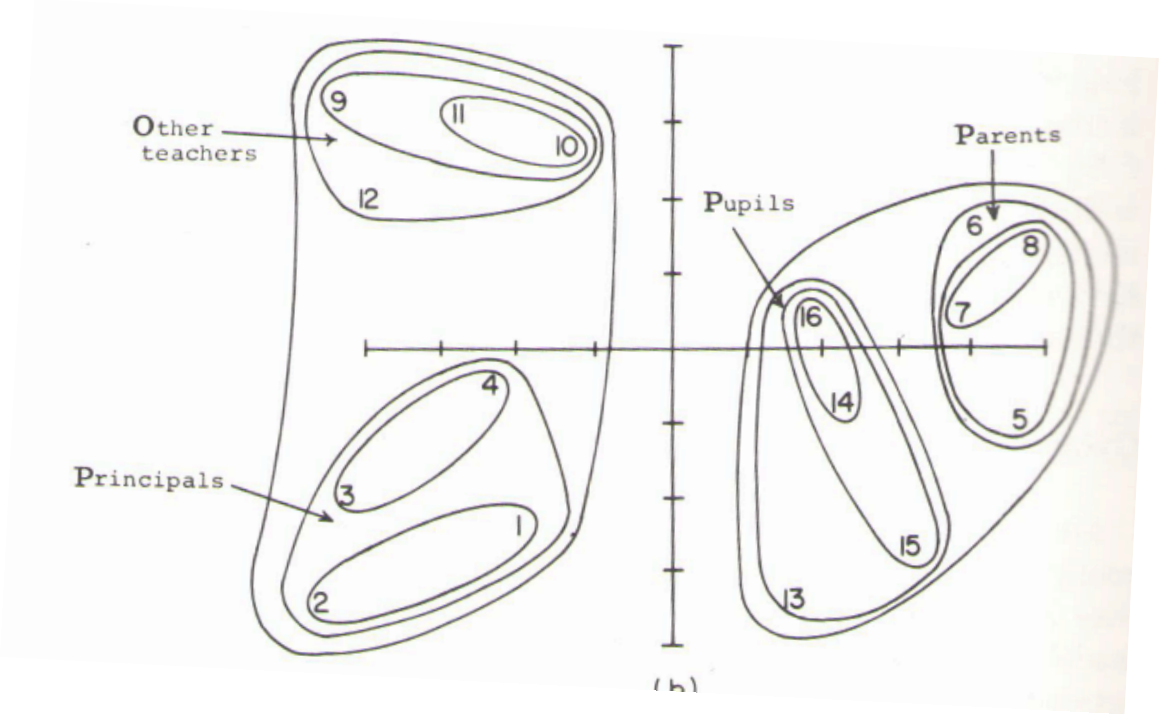


Grouping

Visualizing clustering

Wroclaw diagram (level curves onto MDS)

Effective for hierarchical clusterings. If contours are not tight, this reflects badness of fit in 2D.



Napior (1972)

Grouping

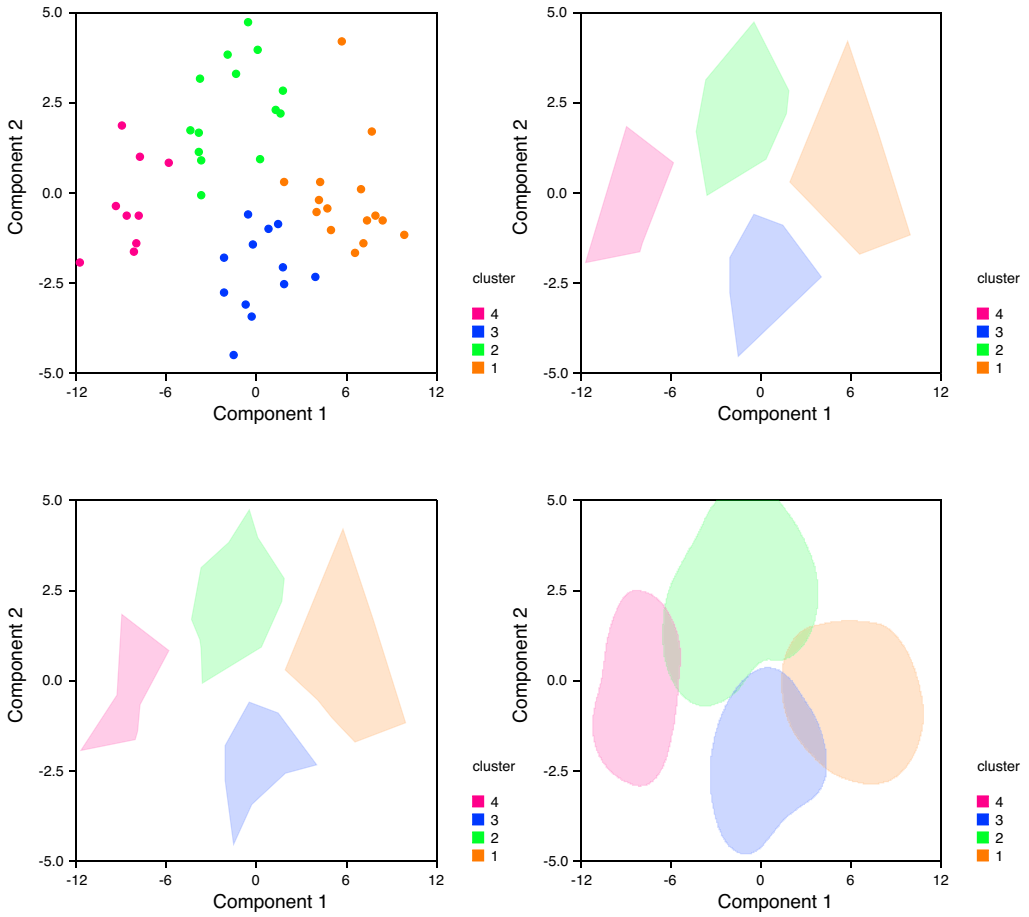
Visualizing clustering

Projections

Convex hulls (upper right)
Alpha shapes (lower left)
Kernels (lower right)

Alpha shapes are nonconvex hulls. They reveal shape of clusters better.

Best to project these on MDS, but principal components will do OK.

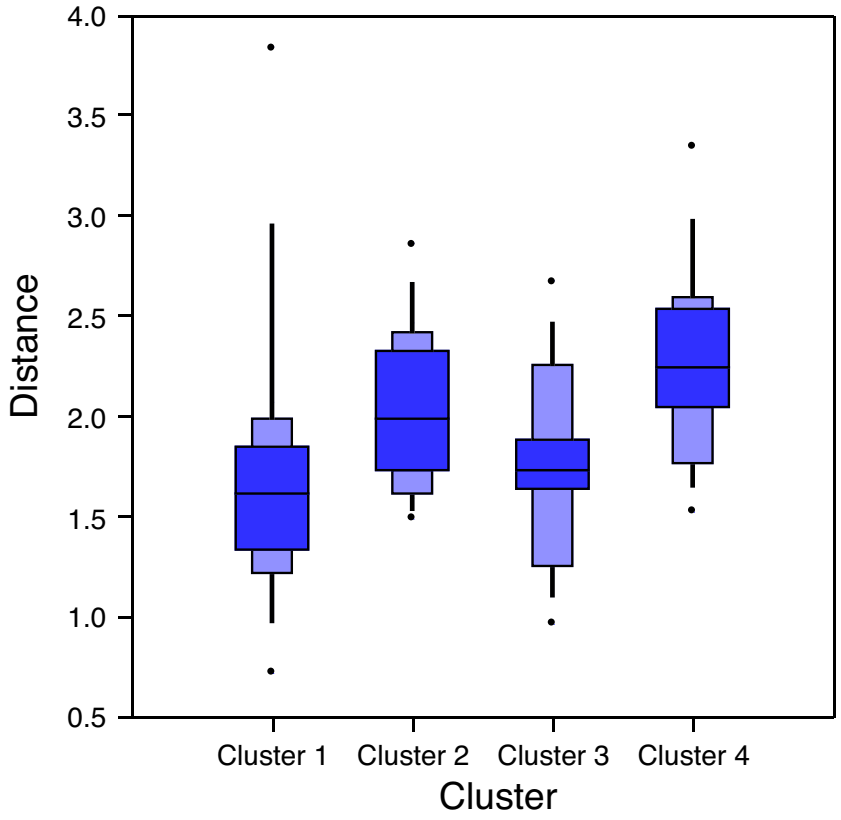


Grouping

Visualizing clustering

Box Plots

Hofman, Kafadar, Wickham letter value box plots reveal outliers better than ordinary box plots.

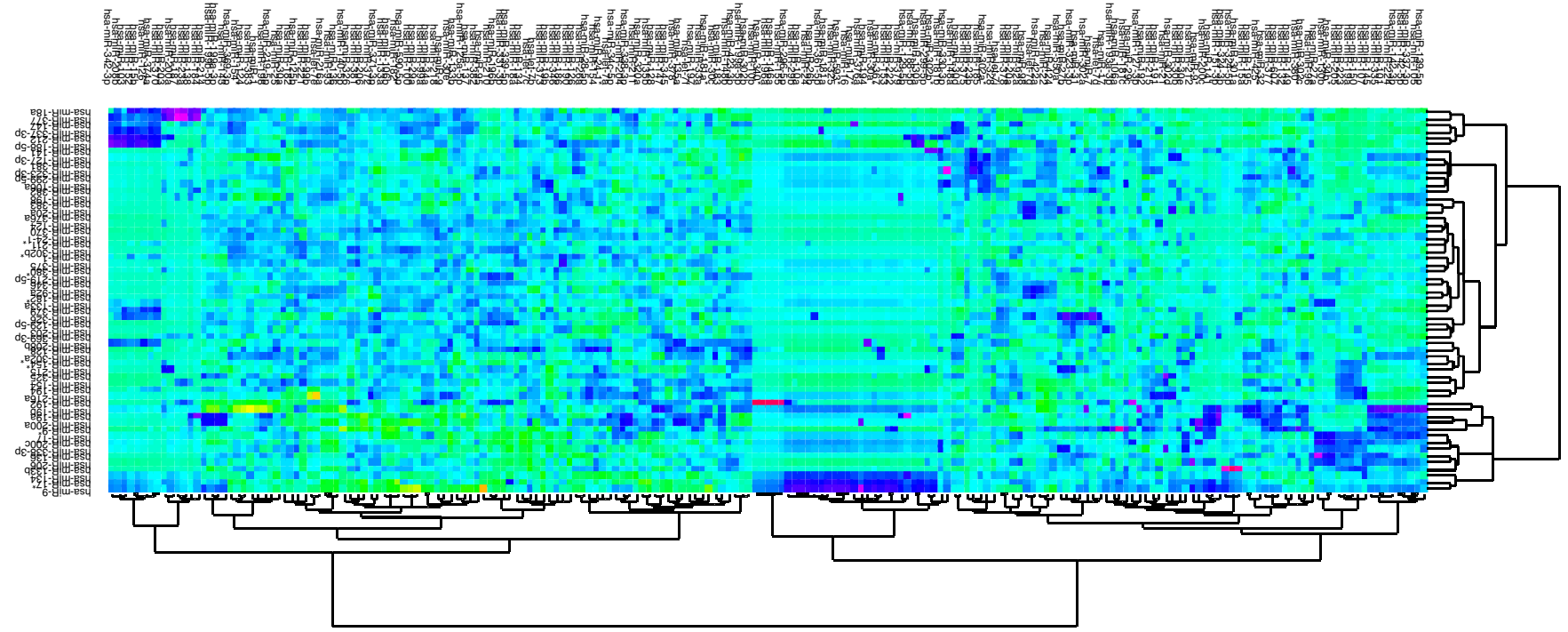


Grouping

Visualizing clustering

Cluster Heatmap

http://www.datavis.ca/papers/HeatmapHistory-tas.2009.pdf

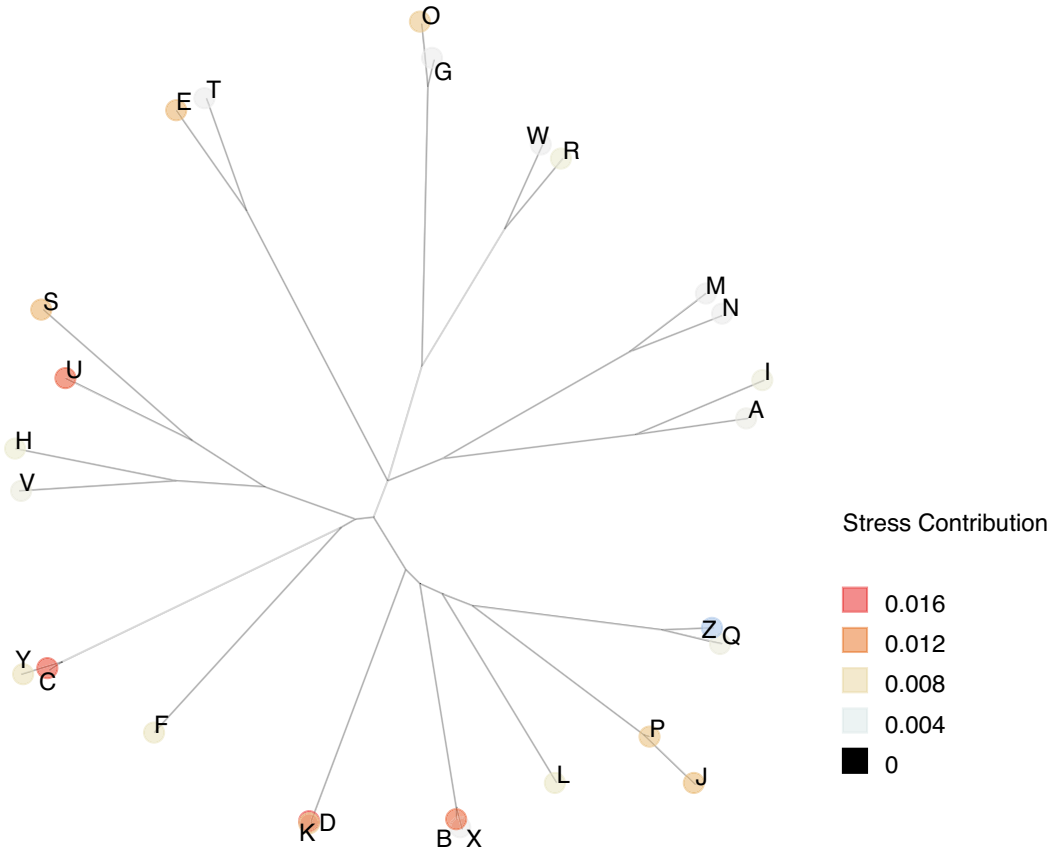


Grouping

Visualizing clustering

Additive Trees

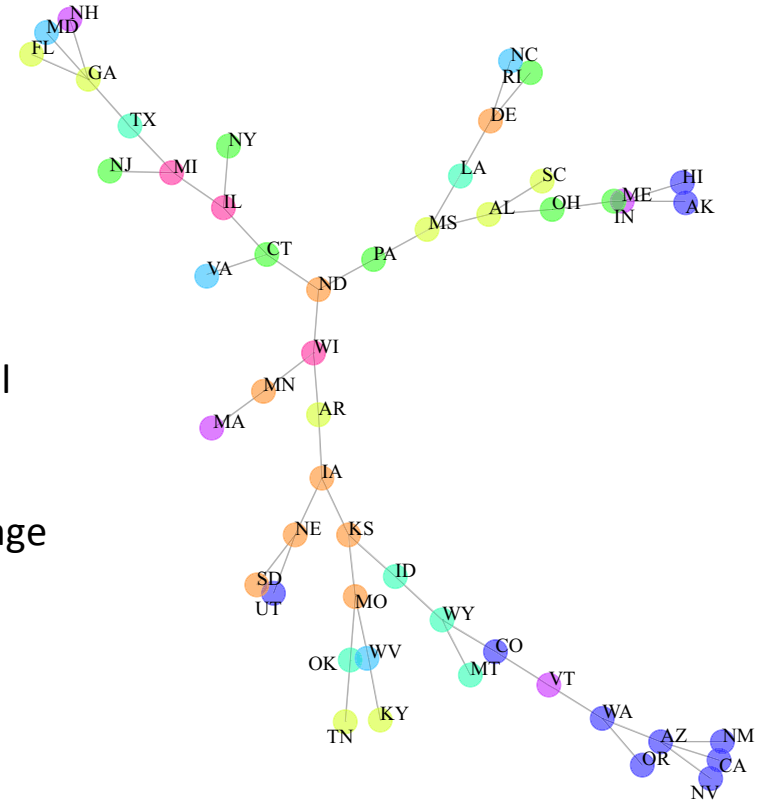
The distance between two leaves is proportional to the shortest path joining them.



Grouping

Visualizing clustering

Minimum Spanning Tree (MST)



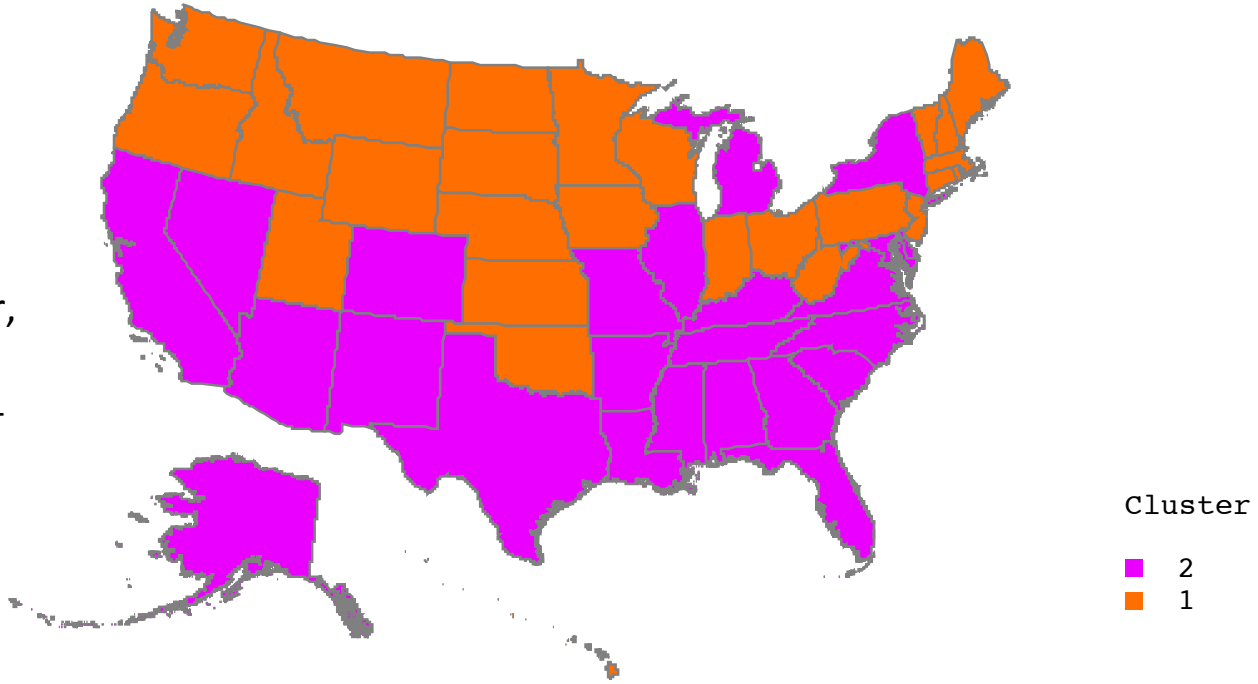
This is shortest (or one of shortest) tree spanning all the objects. Successively deleting longest edges in MST defines a single-linkage tree.

Grouping

Visualizing clustering

Maps

Drawing clusters on maps is a natural way to display geographic clusters. Clustering by state, however, can be misleading. The level of aggregation conceals city-rural differences.



Grouping

References

Fisher, L. and Van Ness, J.W. (1971). Admissible clustering procedures. *Biometrika, 58*, 91-104.

Gower, J.C. (1967). A comparison of some methods of cluster analysis. *Biometrics, 23*, 623-637.

Gruvaeus, G. and Wainer, H. (1972). Two additions to hierarchical cluster analysis. *The British Journal of Mathematical and Statistical Psychology, 25*, 200-206.

Hartigan, J.A. (1975). *Clustering algorithms*. New York: John Wiley & Sons.

Hartigan, J.A., and Mohanty, S. (1992), The RUNT Test for Multi-modality, *Journal of Classification, 9*, 63-70.

Johnson, S.C. (1967). Hierarchical clustering schemes. *Psychometrika, 32*, 241-254.

Lance, G.N., and Williams, W.T. (1967), A general theory of classificatory sorting strategies, I. Hierarchical Systems. *Computer Journal, 9*, 373-380.

Milligan, G.W. (1980). An examination of the effects of six types of error perturbation on fifteen clustering algorithms. *Psychometrika, 45*, 325-342.

Milligan, G.W. and Cooper, M. C. (1985). An examination of procedures for determining number of clusters in a data set. *Psychometrika, 50*, 159-179.

Milligan, G.W. and Cooper, M.C. (1988). A study of standardization of variables in cluster analysis. *Journal of Classification, 45,* 181-204.

Sattath, S., & Tversky, A. (1977). Additive similarity trees. *Psychometrika, 42*, 319-345.

Sokal, R.R. and Sneath, P.H.A. (1963). *Principles of numerical taxonomy*. San Francisco: W. H. Freeman and Company.

Stuetzle, W. and Nugent, R. (2010). A generalized single linkage algorithm for estimating the cluster tree of a density. *Journal of Computational and Graphical Statistics, 19,* 397-418.

Ward, J.H. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association, 58*, 236-244.