**Cluster Analysis**

1. Why and what it is
	1. Organize entities (sampling units) in to discrete classes or groups
	2. Maximizes within-group similarity and minimizes between group similarity
	3. Prespecified, well-defined groups do not exist
	4. Summarizes data redundancy
	5. Identifies outliers because they get their own group(s)
	6. Reduces noise in the data set
	7. No dependent or independent variables
	8. In ecology, a lot of variability is concentrated on a few groups of entities
2. Tough
	1. Start with one big group and separate, or start with separate entities and group?
	2. How do we quantify similarity?
	3. “The ultimate criterion for determining what constitutes a meaningful cluster is the intuition of the user.” – McGarigal et al. 2000
	4. LOTS of different techniques
	5. most methods are biased towards finding spherical clusters
3. Data
	1. species – habitat relationships: group species with similar habitat requirements
	2. Genetic data
	3. Usually all categorical or all continuous data
4. Clustering technique properties
	1. Exclusive\* (hard, nonoverlapping; place each entity in to only one group) vs. nonexclusive (soft, overlapping)
	2. Sequential\* (recursive sequence of techniques) vs. simultaneous
	3. Hierarchical\* (group and arrange in to hierarchy that explains the relationships between groups) vs. nonhierarchical
	4. Agglomerative\* (entities to groups) vs. divisive (one group to more groups)
	5. Polythetic\* (consider all information) vs. monothetic
5. Metrics to calculate
	1. Dissimilarity (continuous data)
		1. Euclidean distance: $d\_{xy}= \sqrt{\sum\_{i=1}^{p}\left(x\_{i}-y\_{i}\right)^{2}}$
		2. Canberra: $d\_{xy}= \sum\_{i=1}^{p}\left[\frac{\left|x\_{i}-y\_{i}\right|}{\left(x\_{i}+y\_{i}\right)}\right]$, nonnegative variables only but insensitive to skewness and outliers
		3. City-block, Minkowski, Czekanowski, Gower’s, Bray-Curtis, …
	2. Similarity (categorical data)
		1. Jaccard’s similarity coefficient, Czekanowski’s similarity coefficient
6. Hierarchical (linkage) clustering
	1. Dendrograms
	2. All members of a child cluster belong to a parent cluster
	3. Compute the distances between entities based on the data characteristics (variables)
	4. Not robust to outliers
	5. Complexity makes it too slow for large datasets
7. Centroid-based clustering (*k*-means clustering)
	1. Find the centers of *k* clusters and assign the entities to the nearest cluster center by minimizing squared distances.
	2. Most often, *k* must be defined in advance
	3. Prefers clusters of equal size
	4. Optimizes cluster centers, not cluster borders
8. Distribution-based clustering
	1. Determine which entities belong to the same distribution
	2. Overfitting (too much complexity) is common
	3. Constrained by a need to define a distribution
9. Density-based clustering
	1. Connect entities that are within a certain distance threshold
	2. DBSCAN
		1. minimum number of points per cluster (*minPts*)
			1. = 1, each entity is its own cluster
			2. = 2, hierarchical
			3. ≥ 3, more density based
		2. Distance cut-off (*Ɛ*)
	3. Advantages
		1. Do not need to predetermine the number of clusters
		2. Cluster shape is arbitrary
		3. Robust to outliers
		4. Two parameters: minimum number of points per cluster (*minPts*) and
		5. Insensitive to entity order
	4. Disadvantages
		1. Not completely deterministic
		2. Differences in densities are difficult
10. Comparing method results (cluster evaluation, internal validation)
	1. Davies-Bouldin index
		1. $DB= \frac{1}{n}\sum\_{i=1}^{n}\max\_{j \ne i}\left(\frac{σ\_{i}+σ\_{j}}{d\left(c\_{i},c\_{j}\right)}\right)$
		2. *n* = number of clusters
		3. *cx* = centroid of cluster *x*
		4. *σx* = average distance of all entities in a cluster to its centroid
		5. *d*(*cx,cy*) = distance between two centroids
		6. lowest index = better method
	2. Dunn index: highest index = best method
11. Nonexclusive methods: Fuzzy C-means, Soft K-means,…