**Overall approach to methods**

* Analysis of community compositional data
  + Sites (rows) x Species (cols)
* Data from NOAA’s Nearshore Fish Atlas Database
  + Subset for collections by beach seines
  + Lots of different projects producing similar data
    - But not quite the same (eg, research questions, sampling design, gear type, etc)
  + Able to be linked to ShoreZone habitat data
* Complications
  + Lots of zeros in data
    - Leads to double zero paradox
    - due to sampling inefficiency or random species occurrences across long env gradients
  + Varied sampling intensities
    - Samples range in number of seines sets from 1 to 8
    - Gear differences in mesh size
    - Local conditions (habitat) targeted differently depending on project goal
  + Likelihood of missing species
    - Each replicate is unlikely to contain all possible taxa
    - More sets will be closer to theoretical richness
      * But abundances will favor species most susceptible to capture method
* Solutions:
  + Deal with excessive zeros!
    - Beals smoothing
      * see vegan::beals() for explanation of uses
      * Use vegan::stepaccross() – similar to Swan’s version of Beal’s smoothing
        + For use when all else fails – when beta diversity is very high and a large proportion of dissimilarities are at upper limit
      * Not sure how to implement with abundance data – see ‘type’ explanation in beals()
    - Should only consider asymmetrical resemblance indices
  + Use of transformation formulas on raw data
    - Allows for input of data into linear methods, eg, PCA RDA, k-means
    - See figs 7.7 and 7.8 in numerical ecology
  + Also use of distance matrix (distance coefficients) to compare sites
    - Hellinger distance, Chi square distance, robust Aitchison distance, Bray-Curtis distance

**Choice of coefficient (table 7.4 in numerical ecology):**

* Using distance coefficients
  + Purpose is to find divergence of large groupings in the data (ordination and/or clustering)
    - Coefficient of variation
      * Ratio of non-linearity to resolution
      * Small ratios are good resemblance coefficients
      * High linearity is good for ordination, high resolution is good for clustering
  + Q-mode vs R-mode
    - Q-mode investigates relationships among site observations (rows)
    - R-mode investigates relationships among species (columns)
    - Will focus on Q-mode, but consider R-mode for further species associations
* Using quantitative data (taxa counts)
  + because that is nature of data in the NFAdb
  + As opposed to P/A or semi-quantitative (ordered)
    - Could re-code to ordered (0 = not there, 1 = rare, 5 = most abundant)
    - But not different than normalization (eg, rel abun) except possibly loss of info
* Needs to take into account the double-zero paradox
  + So must be *asymmetrical*
* Able to be calculated as distances for clustering/ordination
  + So *metric* is preferred but not wholly necessary
    - ie, metric coefficients do not violate the triangle inequality theorem

**R mode analysis (species comparisons)**

* Correlation coefficients: Pearson r or Spearman r
* To find species associations, should transform data,
  + eg, log, hellinger, χ2, see decostand() in vegan
  + Controls for difference in total abundances among sites
  + To linearize relationships for Pearson r
  + Or, to make more monotonic for Spearman r
* Should also transform to distance coefficient for clustering or ordination
  + Basic index: D = 1 – r or D = sqrt( 1 - r )
    - Species with identical distributions will have r = 1 (D = 0)
    - Negative correlations will produce D > 1
      * This is no problem in clustering or ordination
  + To induce clustering/ordination of species that are either pos or neg correlated
    - i.e., strongly associated and strongly disassociated species will be grouped
    - D = (1-r2)0.5
    - Idea – could use this to subset species list to only species that have strong associations
      * Maybe this reduces ‘messy’ clusters compared to full dataset

**Q mode analyses (site comparisons)**

Coefficients that can be used for raw or normalized quantitative data:

Bray-Curtis distance

* D14 - semi-metric and not Euclidean
  + However, sqrt( D14 ) is both metric and euclidean
* aka ‘percentage difference’ or Sorenson index or Steinhaus index
* based off of Steinhaus similarity which aims to lessen the effect of large differences in species abundances, and so is fitting for raw abundance data
* 0 to 1 scale

Hellinger distance

* D17 – metric and Euclidean
* “hellinger” in vegan
* related to chord distance as a geometric type standardization
  + in that row points are moved to surface of a hypersphere and distances are chords
  + except, data are standardized to row total (relative abundance) then sqrt transformed
* has upper limit of sqrt(2) = 1.414
* Along with Chord distance (D3), hellinger is recommended for general usage

χ2 distance

* Need to see if vegan “chisq” uses D15 or D16
* D15 is χ2 metric – uses *absolute* frequency of species (whole col sum) to weight calculation
* D16 is χ2 distance – uses *relative* frequency of species (whole col sum / sum of all col sums) to weight, essentially taking into account whether species as a whole is relatively abundant and weighting rarer species moreso.
  + D16 is the transposed version of D15 where rows are species then χ2 dist applied
* Takes the *species profile dist* (D18, which is Euclidean dist weighted by relative row abundance) then weights by the frequency of species
  + Essentially, data is weighted by rows then by columns
* Should try to use D15 since absolute dist is less
  + however, the two are monotonic in that D16 is just D15 multiplied by a constant (sqrt of total table sum).
  + Eg, if two sites in comparison, it doesn’t matter as much if there are lots/few sand lance caught, matters more that the other site did not catch them
* Overall, it gives more weight to rare species, ie, abundant species contribute less to the difference in distances btwn sites.
* Ordination of χ2 distance is same as in canonical correspondence analysis (CCA)

Other coefficients not outlined in Table 7.4 of Numerical Ecology

Chao distance

* D = 1 – S8, modified abundance-based Jaccard similarity index
* Unsure of metric or Euclidean status – will need to check this
* “chao” in vegan
* Less biased than Jaccard index when a substantial portion of species are missing from sample data
  + 1 – U\*V/(U+V-U\*V)
  + Aims to take into account the number of unseen species pairs
  + Many versions of chao dissimilarity but chose default (Jaccard) because it was the most simple to interpret and others give more weight to double presences (Sorenson)

Robust Aitchison distance

* Martino et al. 2019

\*Goodall distance

* Based on probabilities - Would like to learn more about!
* Usage with Beals Smoothing?

**Cluster Analysis**

Types of methods/algorithms

* Sequential vs simultaneous
* Agglomeration vs division
  + Agglomerative algorithms are most developed
  + In division, samples that could be ‘either-or’ grouped or not are hard to correct for after initial partition
    - Monothetic vs polythetic (for divisive clustering)
  + Maybe a division in ordination space could work:
    - Efficient algorithm for large dataset when a coarse division of objects is sought
    - Pg 380-381 in Numerical Ecology
* Hierarchical vs non-hierarchical
* Probabilistic vs non-probabilistic
* Constrained clustering
  + 13.3.2 space-constrained clustering

Hierarchical Agglomerative Clustering

* See table 8.9 for generalized model for all methods
  + Single linkage = contracting in A space
  + Complete linkage = dilating in A space
  + Others are ‘conserving’ in A space
* Use of weighted as opposed to unweighted
  + Gives equal weight to the two branches about to be fused
    - equivalent to differential weighting of original distances
    - ie, down-weighting the distances of the largest group...
  + WPGMA – weighted arithmetic average clustering
    - hclust(method = “mcquitty”)
  + WPGMC – weighted centroid clustering
    - hclust(method = “median”)
* Ward’s minimum variance
  + Ward.D2
    - cluster::agnes(method = “ward”)