- The group numbers are arbitrary. Remember that you can rotate dendrograms around any node and not change the meaning. So, the order of the clusters is not meaningful.
- Taking a subset of the data changes the analysis quite a bit. Everyone's results could be quite different…
- Standardizing with daisy

## Bioclimatic and physical characterization of the world's islands

Patrick Weigelt[a,1], Walter Jetz[b], and Holger Kreft[a,1]

[a]Biodiversity, Macroecology and Conservation Biogeography Group, University of Göttingen, D-37077 Göttingen, Germany; and [b]Department of Ecology and Evolutionary Biology, Yale University, New Haven, CT 06520

Edited by Daniel S. Simberloff, The University of Tennessee, Knoxville, TN, and approved July 24, 2013 (received for review April 12, 2013)

Here, we aim to provide a comprehensive environmental synopsis and classification of the world's islands. We (*i*) provide a comprehensive multivariate characterization and a standardized dataset of island bioclimatic and physical conditions; (*ii*) compare island and mainland environments; (*iii*) explore multivariate approaches for delineating environmental island ecoregions; (*iv*) provide general perspectives how this unique multivariate characterization may be used in island research and management; and (*v*) implement an example application by making environment-based predictions of vascular plant species richness on islands worldwide.

- Dataset:
  - 85,122 islands
  - 19,392 > 1km$^2$
  - 17,883 with data
  - Data: size, shape, location, climate, seasonality, distance to mainland etc.
  - Island-mainland comparison dataset

## Analyses

- Analyses (especially clustering) focused on sets of variables
  - 10 climate and physical variables. Some analyses excluded area and elevation (8 variables)
- PCA summarized variables but did not classify

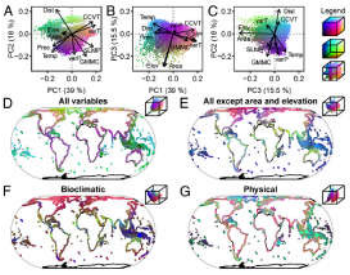



Fig. 2. Principal component analyses (PCAs) based on bioclimatic and physical variables for 17,883 marine islands >1 km$^2$ worldwide. (*A*–*C*) Biplots of the first three PCA axes when all 10 variables are included. (*D*–*G*) Maps of ordination site scores: (*D*) all 10 variables, (*E*) all variables but *Area* and *Elev*, (*F*) contemporary bioclimatic variables only (*Temp*, *var7*, *Prec*, *varP*), and (*G*) physical variables only (*Area*, *Dist*, *SLMP*, *GMMC*, *Elev*). Colors refer to a red-green-blue (RGB) color space (cubes in legend) projected onto the respective 3D PCA space. Hence, in *A*–*D* each island consistently has the same color. Cubes in *D*–*G* show PCA results in a 3D space. In *A*–*C*, points are plotted in decreasing order of the respective component not shown to give an impression of three-dimensionality, in *D*–*G* in decreasing order of island area, plotting the rare large islands on top if points overlap. Abbreviations follow Fig. 1.

## Analyses

- Classification was done via clustering (PAM and UPGMA) to provide objective ecoregion framework

To delimit island regions of similar bioclimatic and physical conditions, we performed cluster analyses based on the 10 environmental variables and the variable subsets mentioned above. We used agglomerative hierarchical (UPGMA) and nonhierarchical clustering methods (PAM). UPGMA produces a cluster dendrogram representing the relatedness of the delimited regions. From the dendrogram, a preferred number of clusters can be inferred (56). PAM requires a specified number of clusters in advance and does not provide relationships among regions. However, PAM tends to delineate clusters of similar size and upper limits of within-group variance, preventing the creation of regions that greatly differ in within-region variance (38). Due to

- Why 8 clusters?
- Package clusterCrit provides clustering indices such as those used in the paper:

axes (57). We chose a number of clusters small enough for presentation and discussion based on the Caliński and Harabasz index (56).

56. Milligan G, Cooper M (1985) An examination of procedures for determining the number of clusters in a data set. *Psychometrika* 50(2):159–179.

## clusterCrit

- Calinksi Harabasz index at various K (PAM)

```
Grps<-cutree(UPGMA_object,K)
intCriteria(PCA_matrix,grps,"Calinski_Harabasz")
```

- K=6 3396.48
- K=7 3642.95
- K=8 4129.99
- K=9 3615.56
- K=10 3407.42

**1.2.4 The Calinski-Harabasz index**

Using the notations of equations (16) and (23), the Calinski-Harabasz index is defined like this:

$$\mathcal{C} = \frac{BGSS/(K-1)}{WGSS/(N-K)} = \frac{N-K}{K-1}\frac{BGSS}{WGSS} \quad (33)$$

## Pick correct proportions from each group…

- 2325 out of 2500 are in corresponding groups

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|---|
| 1 | **457** | 21 | 0 | 0 | 0 | 0 | 0 |
| 2 | 0 | 0 | **474** | 0 | 0 | 0 | 0 |
| 3 | 29 | 111 | 0 | **851** | 0 | 0 | 0 |
| 4 | 2 | **362** | 0 | 0 | 11 | 0 | 0 |
| 5 | 0 | 0 | 0 | 0 | 0 | **2** | 0 |
| 6 | 0 | 0 | 0 | 0 | **2** | 0 | 0 |
| 7 | 0 | 0 | 1 | 0 | 0 | 0 | **177** |

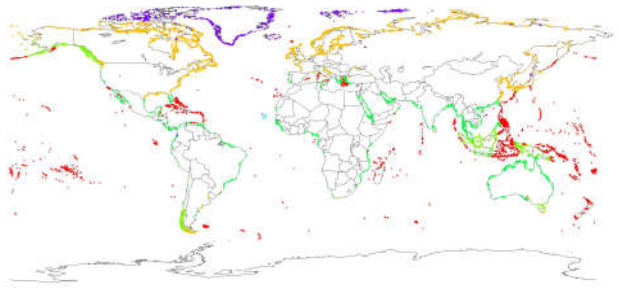
- Same set of islands, data standardized

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|---|
| 1 | 412 | 63 | **3** | 0 | 0 | 0 | 0 |
| 2 | 452 | **9** | 13 | 0 | 0 | 0 | 0 |
| 3 | **982** | 0 | 9 | 0 | 0 | 0 | 0 |
| 4 | 333 | 0 | 41 | **1** | 0 | 0 | 0 |
| 5 | 0 | 0 | 0 | 0 | **2** | 0 | 0 |
| 6 | 0 | 0 | 0 | 0 | 0 | **2** | 0 |
| 7 | 177 | 0 | 0 | 0 | 0 | 0 | **1** |

## Perfect Classification (?)

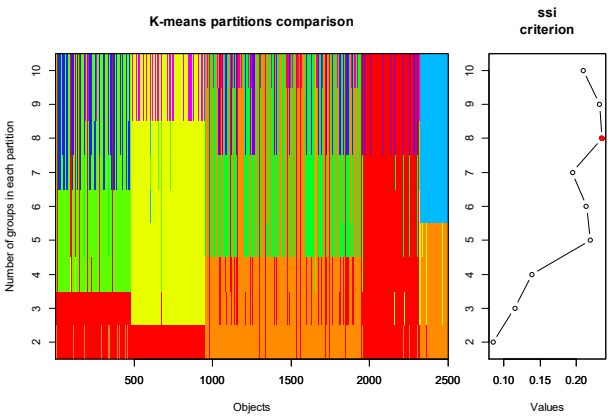
```
UPGMA:
table(grps,dat$UPGMAnoAE)

grps    1    2    3    4    5    6    7    8
   1    0 3391    0    0    0    0    0    0
   2    0    0 7092    0    0    0    0    0
   3    0    0    0 2680    0    0    0    0
   4 3419    0    0    0    0    0    0    0
   5    0    0    0    0   15    0    0    0
   6    0    0    0    0    0   13    0    0
   7    0    0    0    0    0    0 1272    0
   8    0    0    0    0    0    0    0    1

PAM:
  1    2    3    4    5    6    7    8
  1    0 2204    0    0    0    0    0    0
  2    0    0 3166    0    0    0    0    0
  3    0    0    0 1691    0    0    0    0
  4    0    0    0    0 2338    0    0    0
  5 2358    0    0    0    0    0    0    0
  6    0    0    0    0    0 2236    1    0
  7    0    0    0    0    0    1 2702    0
  8    0    0    0    0    0    0    0 1186
```

```
map("worldHires")
clus_colors<-rainbow(8)
points(dat$Long,dat$Lat,pch=16,cex=1,col=clus_colors[grps])
```



## cascadeKM

- Function cascadeKM does K means clustering over a range of K
- Here, doing clustering from K=2 to K=10
  - `caskmeans<-cascadeKM(weighted_variables,2,10,iter=100, criterion="ssi")`
- Function returns
  - group membership for each group size (similar to cuttree)
  - Number of samples in each group
  - SSE for each K
    - `2 groups  3 groups  4 groups  5 groups  6 groups  7 groups  8 groups`
    - `SSE  889.22760 676.20559 568.21645 505.22622 449.86885 408.17714 379.02992 iii 67.37003  61.06284  54.84252  49.16689  46.36945  43.97377  41.34622`



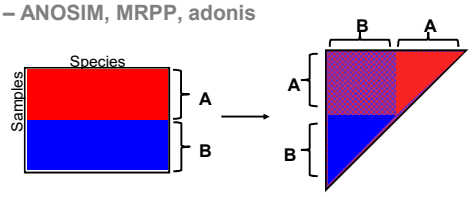

## Fuzzy Clustering

- Specify k ahead of time, not hierarchical
- Observations may belong to multiple clusters (fuzzy)
- Functions
  - fanny(cluster package)
  - cmeans(e1071 package), fuzzy version of kmeans
- For the island dataset we are working with:
  - `fanny(weighted_variables,k=8,memb.exp=1.4)`

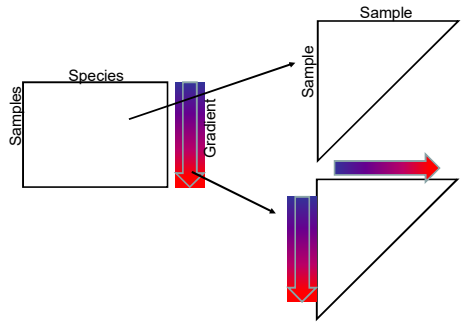
```
Fuzzy Clustering object of class 'fanny' :
m.ship.expon.      1.4
objective     271.2391
tolerance        1e-15
iterations         257
converged            1
maxit              500
n                 2500
Membership coefficients (in %, rounded):
      [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8]
[1,]   29   27   11   16    8    3    4    2
[2,]   29   33   15    9    8    2    2    2
[3,]   11    6   62   11    2    2    3    3
[4,]   15    9   56   12    2    1    2    1
[5,]    8    4   64   13    1    2    4    4
[6,]   34   34   14   10    3    1    2    1
Etc
```

For each island, Scores for assignment to each group.

## Eariler – ANOSIM, MRPP, adonis

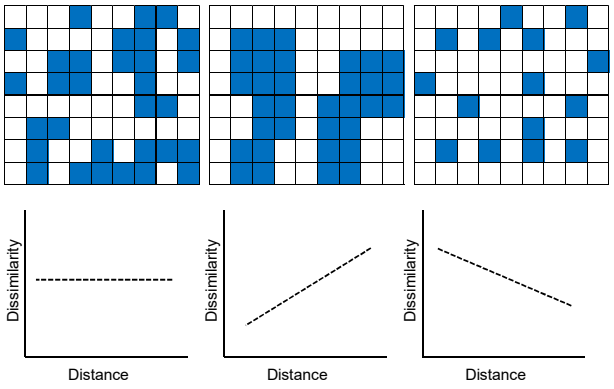



- We had distinct groups of samples and wanted to know if groups A and B were different.
- Our question involved similarity among samples – Are samples within A more similar to each other than samples in B?
- What if we do not have distinct groups (e.g. samples taken along an ecological gradient)?

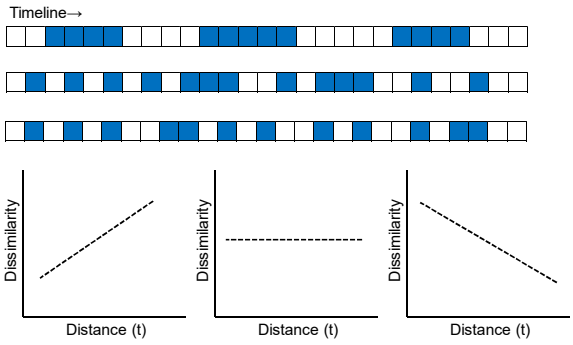



- Question: are samples taken close together along the gradient more similar than samples separated along the gradient?

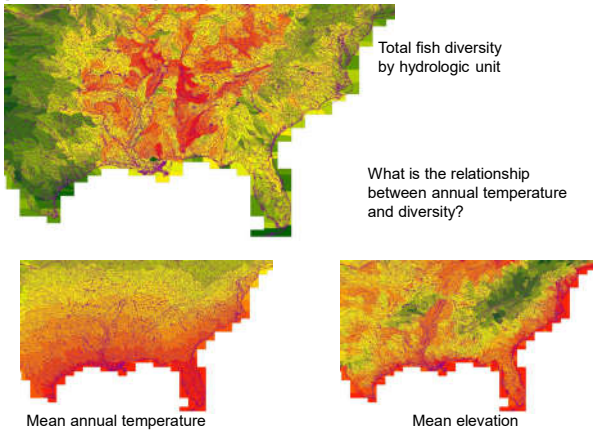
## Spatial Autocorrelation



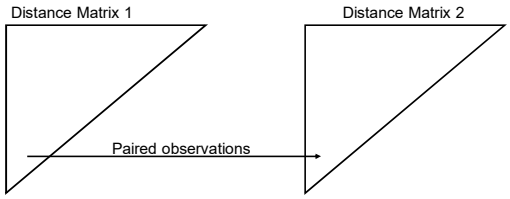

## Temporal autocorrelation




## Spatial (and temporal) Autocorrelation



## Mantel test – concordance among distance matrices



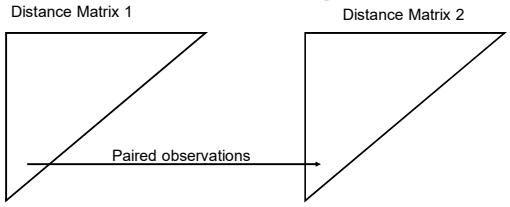

- Any two distance matrices can be used. Common types:
  - community similarity vs. time (temporal autocorrelation)
  - community similarity vs. distance (spatial autocorrelation)
  - genetic similarity vs. distance (isolation by distance)

## Mantel Statistic

- Where $X_{ij}$ and $Y_{ij}$ are paired elements of two distance matrices.
- Z increases with the concordance among two matrices increases
- Correlation coefficient (r) often used instead of Z, the two are linearly related

$$Z = \sum_{i<j}^{n} X_{ij} Y_{ij}$$

## Mantel test – concordance among distance matrices




- Calculate the observed Mantel statistic.
- Permute data x number of times (by convention, permute first matrix), recalculate to produce a distribution of Mantel statistics to assess significance.
- If null is true (no concordance among matrices), observed will fall somewhere in the middle of the permuted distribution.

## Code

- In R (vegan package):

```
community_distance<-vegdist(community_log, method='bray')
environmental_distance<-vegdist(environmental, method='bray')
mantel<-mantel(community_distance, environmental_distance, method="pearson",
permutations=1000)
```

- Options
  - Permutations
  - Strata
  - Spearman, Pearson or Kendall correlation
- Output
  - Mantel statistic, vector of permuted values

## Partial Mantel

- Compare two distance matrices, controlling for influence of a third.
- Uses a partial correlation. The correlation is between the residuals of the first and second matrix when correlated with the third.
- Code is similar, call different function and pass three matrices:
  - `mantel.partial(x, y, z, method="pearson", permutations=1000)`
- Only first matrix is permuted

## Burn regimes and community change

- Mantel tests well suited to testing for spatial and temporal autocorrelations.
- Konza example:
  - Plots with 1, 5, 20 year burn regimes, with and without buffalo
  - LTER – over 25 years of data available
  - Question – how do communities respond to burn regimes through time?

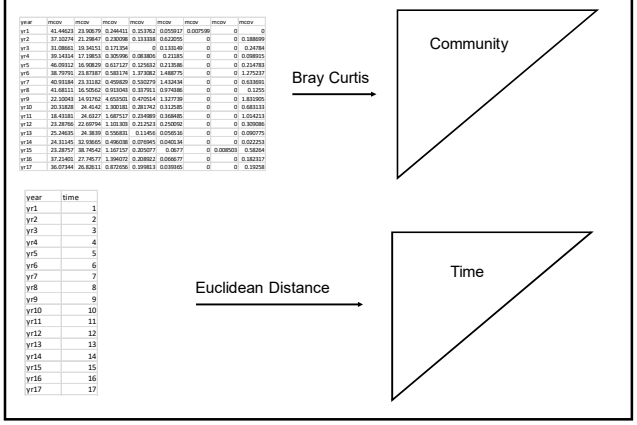
## Example Mantel Test

- Sample data (Konza LTER project) and script
  - Time file (time.csv)
  - Community data from one year (burn1.csv) and 20 year (burn20.csv) burn regimes.
  - Script to examine temporal autocorrelations (quantitative and qualitative similarity indices)

| year | mcov | mcov | mcov | mcov | mcov | mcov | mcov | mcov |
|---|---|---|---|---|---|---|---|---|
| yr1 | 41.44623 | 23.90679 | 0.244411 | 0.153762 | 0.055917 | 0.007599 | 0 | 0 |
| yr2 | 37.10274 | 21.29847 | 0.230098 | 0.133338 | 0.622055 | 0 | 0 | 0.188699 |
| yr3 | 31.08661 | 19.34151 | 0.171354 | 0 | 0.133149 | 0 | 0 | 0.24784 |
| yr4 | 39.14314 | 17.19853 | 0.305996 | 0.083806 | 0.21185 | 0 | 0 | 0.098915 |
| yr5 | 46.09312 | 16.90829 | 0.617127 | 0.125632 | 0.213586 | 0 | 0 | 0.214783 |
| yr6 | 38.79791 | 23.87387 | 0.583174 | 1.373082 | 1.488775 | 0 | 0 | 1.275237 |
| yr7 | 40.93184 | 23.31182 | 0.459829 | 0.530279 | 1.432434 | 0 | 0 | 0.633691 |
| yr8 | 41.68111 | 16.50562 | 0.913043 | 0.337911 | 0.974386 | 0 | 0 | 0.1255 |
| yr9 | 22.10043 | 14.91762 | 4.653501 | 0.470514 | 1.327739 | 0 | 0 | 1.831905 |
| yr10 | 20.31828 | 24.4142 | 1.300181 | 0.281742 | 0.312585 | 0 | 0 | 0.683133 |
| yr11 | 18.43181 | 24.6327 | 1.687517 | 0.234989 | 0.368485 | 0 | 0 | 1.014213 |
| yr12 | 23.28766 | 22.69794 | 1.101303 | 0.212523 | 0.250092 | 0 | 0 | 0.309086 |
| yr13 | 25.24635 | 24.3839 | 0.556831 | 0.11456 | 0.056516 | 0 | 0 | 0.090775 |
| yr14 | 24.31145 | 32.93665 | 0.496038 | 0.076945 | 0.040134 | 0 | 0 | 0.022253 |
| yr15 | 23.28757 | 38.74542 | 1.167157 | 0.205077 | 0.0677 | 0 | 0.008503 | 0.58264 |
| yr16 | 37.21401 | 27.74577 | 1.394072 | 0.208922 | 0.066677 | 0 | 0 | 0.182317 |
| yr17 | 36.07344 | 26.82611 | 0.872656 | 0.199813 | 0.039365 | 0 | 0 | 0.19258 |

| year | time |
|---|---|
| yr1 | 1 |
| yr2 | 2 |
| yr3 | 3 |
| yr4 | 4 |
| yr5 | 5 |
| yr6 | 6 |
| yr7 | 7 |
| yr8 | 8 |
| yr9 | 9 |
| yr10 | 10 |
| yr11 | 11 |
| yr12 | 12 |
| yr13 | 13 |
| yr14 | 14 |
| yr15 | 15 |
| yr16 | 16 |
| yr17 | 17 |

## Example Mantel Test

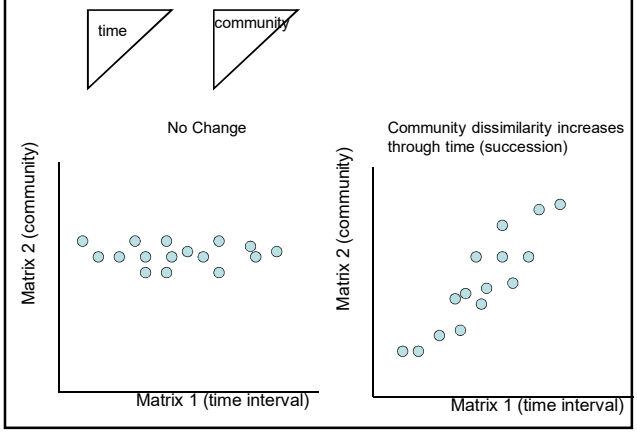


- `time_distance<-vegdist(time, method='euclidean')`

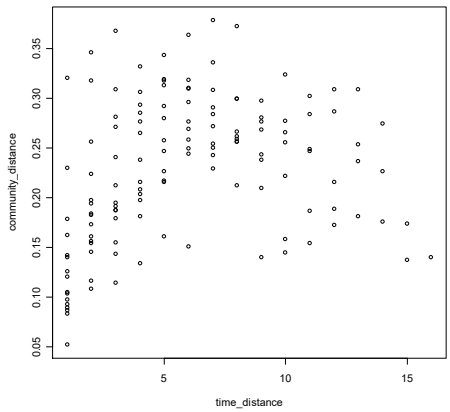
| samples | year |
|---|---|
| yr1 | 1 |
| yr2 | 2 |
| yr3 | 3 |
| yr4 | 4 |
| yr5 | 5 |
| yr6 | 6 |
| yr7 | 7 |
| yr8 | 8 |
| yr9 | 9 |
| yr10 | 10 |
| yr11 | 11 |
| yr12 | 12 |
| yr13 | 13 |
| yr14 | 14 |
| yr15 | 15 |
| yr16 | 16 |
| yr17 | 17 |

| | yr1 | yr2 | yr3 | yr4 | yr5 | yr6 | yr7 | yr8 | yr9 | yr10 |
|---|---|---|---|---|---|---|---|---|---|---|
| yr2 | 1 | | | | | | | | | |
| yr3 | 2 | 1 | | | | | | | | |
| yr4 | 3 | 2 | 1 | | | | | | | |
| yr5 | 4 | 3 | 2 | 1 | | | | | | |
| yr6 | 5 | 4 | 3 | 2 | 1 | | | | | |
| yr7 | 6 | 5 | 4 | 3 | 2 | 1 | | | | |
| yr8 | 7 | 6 | 5 | 4 | 3 | 2 | 1 | | | |
| yr9 | 8 | 7 | 6 | 5 | 4 | 3 | 2 | 1 | | |
| yr10 | 9 | 8 | 7 | 6 | 5 | 4 | 3 | 2 | 1 | |
| yr11 | 10 | 9 | 8 | 7 | 6 | 5 | 4 | 3 | 2 | 1 |
| yr12 | 11 | 10 | 9 | 8 | 7 | 6 | 5 | 4 | 3 | 2 | 1 |
| yr13 | 12 | 11 | 10 | 9 | 8 | 7 | 6 | 5 | 4 | 3 | 2 | 1 |
| yr14 | 13 | 12 | 11 | 10 | 9 | 8 | 7 | 6 | 5 | 4 | 3 | 2 | 1 |
| yr15 | 14 | 13 | 12 | 11 | 10 | 9 | 8 | 7 | 6 | 5 | 4 | 3 | 2 | 1 |

## Burn regimes and community change

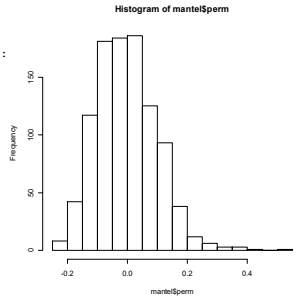



## 1 year burn regime




## 1 year burn regime

- Mantel statistic based on Pearson's product-moment correlation
- Call:
- mantel(xdis = time_distance, ydis = community_distance, method = "pearson", permutations = 1000)
- Mantel statistic r: 0.368
- Significance: 0.003
- Empirical upper confidence limits of r:
- 90% 95% 97.5% 99%
- 0.125 0.167 0.202 0.280
- Based on 1000 permutations



## 20 year burn regime

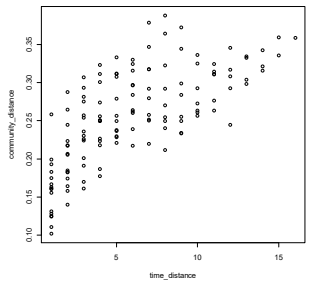
Mantel statistic based on Pearson's product-moment correlation

Call:
mantel(xdis = time_distance, ydis = community_distance, method = "pearson", permutations = 1000)

Mantel statistic r: 0.6988
Significance: 0.001

Empirical upper confidence limits of r:
90% 95% 97.5% 99%
0.130 0.176 0.208 0.256

Based on 1000 permutations




## Mantel Test

- In the example, there are Mantel tests using Bray Curtis and Jaccards indices for community similarity.
- Biologically, what is the difference between these two tests?

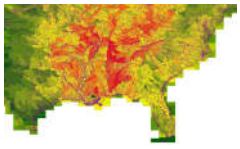
| year | mcov | mcov | mcov | mcov | mcov | mcov | mcov | mcov |
|---|---|---|---|---|---|---|---|---|
| yr1 | 41.44623 | 23.90679 | 0.244411 | 0.153762 | 0.055917 | 0.007599 | 0 | 0 |
| yr2 | 37.10274 | 21.29847 | 0.230098 | 0.133338 | 0.622055 | 0 | 0 | 0.188699 |
| yr3 | 31.08661 | 19.34151 | 0.171354 | 0 | 0.133149 | 0 | 0 | 0.24784 |
| yr4 | 39.14314 | 17.19853 | 0.305996 | 0.083806 | 0.21185 | 0 | 0 | 0.098915 |
| yr5 | 46.09312 | 16.90829 | 0.617127 | 0.125632 | 0.213586 | 0 | 0 | 0.214783 |
| yr6 | 38.79791 | 23.87387 | 0.583174 | 1.373082 | 1.488775 | 0 | 0 | 1.275237 |
| yr7 | 40.93184 | 23.31182 | 0.459829 | 0.530279 | 1.432434 | 0 | 0 | 0.633691 |
| yr8 | 41.68111 | 16.50562 | 0.913043 | 0.337911 | 0.974386 | 0 | 0 | 0.1255 |
| yr9 | 22.10043 | 14.91762 | 4.653501 | 0.470514 | 1.327739 | 0 | 0 | 1.831905 |
| yr10 | 20.31828 | 24.4142 | 1.300181 | 0.281742 | 0.312585 | 0 | 0 | 0.683133 |
| yr11 | 18.43181 | 24.6327 | 1.687517 | 0.234989 | 0.368485 | 0 | 0 | 1.014213 |
| yr12 | 23.28766 | 22.69794 | 1.101303 | 0.212523 | 0.250092 | 0 | 0 | 0.309086 |
| yr13 | 25.24635 | 24.3839 | 0.556831 | 0.11456 | 0.056516 | 0 | 0 | 0.090775 |
| yr14 | 24.31145 | 32.93665 | 0.496038 | 0.076945 | 0.040134 | 0 | 0 | 0.022253 |
| yr15 | 23.28757 | 38.74542 | 1.167157 | 0.205077 | 0.0677 | 0 | 0.008503 | 0.58264 |
| yr16 | 37.21401 | 27.74577 | 1.394072 | 0.208922 | 0.066677 | 0 | 0 | 0.182317 |
| yr17 | 36.07344 | 26.82611 | 0.872656 | 0.199813 | 0.039365 | 0 | 0 | 0.19258 |

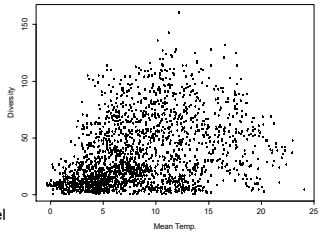
## Assignment

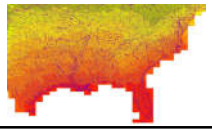
- Reading in text (pages 234,235 and 237 in the spatial analysis chapter)
- Riginos, C., and M.W. Nachman. 2001. Population subdivision in marine environments: the contributions of biogeography, geographical distance and discontinuous habitat to genetic differentiation in a blennioid fish, *Axoclinus nigricaudus.* Molecular Ecology 10: 1439-1453.

## Actual data to analyze

- North American Fish diversity by hydrologic unit





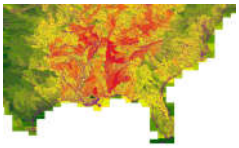



1. Calculate the raw correlation between 1) temperature and diversity, and 2) area and diversity.
2. Test both of these with a Mantel test (you will need three Euclidean matrices).
3. For any significiant differences in 2, test with a Mantel test partialing out geographic proximity (you will need a geographic distance matrix).

## Dataset

| HUC_8 | diversity | mean_ter | mean_pre | elevation | area | lat | long |
|---|---|---|---|---|---|---|---|
| 01010001 | 27 | 0.133147 | 13.39436 | 125.671 | 0.657723 | -69.4271 | 46.83062 |
| 01010002 | 22 | 0.423606 | 13.54572 | 111.576 | 0.374111 | -69.2963 | 46.65681 |
| 01010003 | 24 | 0.19675 | 13.52068 | 110.727 | 0.273418 | -68.576 | 47.01407 |
| 01010004 | 27 | 0.793202 | 13.84615 | 124.358 | 0.730823 | -68.4177 | 46.61989 |
| 01010005 | 18 | 1.414006 | 14.24819 | 158.898 | 0.191612 | -67.9182 | 46.29275 |
| 01020001 | 29 | 1.410458 | 13.97125 | 116.725 | 0.64102 | -69.4474 | 45.91178 |
| 01020002 | 30 | 1.319476 | 14.03739 | 112.736 | 0.335988 | -68.9005 | 46.11152 |
| 01020003 | 17 | 2.019296 | 14.53922 | 133.657 | 0.452414 | -68.1568 | 45.78355 |
| 01020004 | 22 | 2.271052 | 14.42044 | 105.166 | 0.428189 | -69.2263 | 45.34653 |
| 01020005 | 29 | 2.848104 | 14.85065 | 129.052 | 0.700102 | -68.6693 | 45.05461 |
| 01030001 | 25 | 1.677835 | 13.96975 | 117.216 | 0.474825 | -69.9541 | 45.60227 |
| 01030002 | 16 | 2.11347 | 14.01606 | 98.3316 | 0.261736 | -70.4132 | 45.25812 |
| 01030003 | 33 | 3.162897 | 14.72402 | 118.491 | 1.016256 | -69.7963 | 44.73991 |

Spatial – Centroid of each HUC



All North American 8 digit Hydrologic Units (HUC). For each HUC:

- It's 8 digit code name
- diversity – S
- mean annual temperature, and precipitation
- elevation and area
- latitude and longitude

Large dataset (~2000 HUCs). Distance matrices will have ~2 million values.
On a fast computer, 1000 permutations will take ~15 minutes.