Introduction to Spatial Statistics

Advanced Biostatistics

Dean Adams

Lecture 13

EEOB 590C

Point Patterns	
Example data	What is spatial association?
Mantel	Association between response and geography
Partial Mantel	Association between response and ecology, given geography
GLS	Estimating ecology effects with respect to geography
Comparisons	Multiple methods – which is best?
Summary	Much area for exploration2

- Biological organisms inhabit physical space
- Several interesting questions relate to geography/space

1: How are my observations distributed?

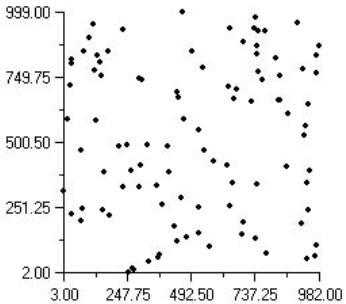
2: Are my data associated with geography?

3: Are my data spatially autocorrelated?

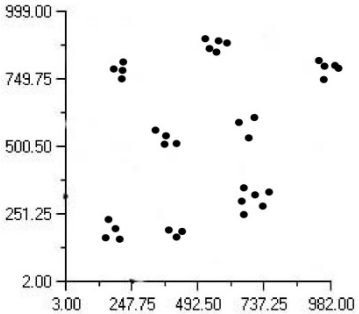
4: Can I account for spatial autocorrelation in my analysis?

- The distribution of objects in space can be of interest
- Possible distributions: the locations are **random,** they are more **clumped** than expected, they are more **regular** (dispersed) than expected

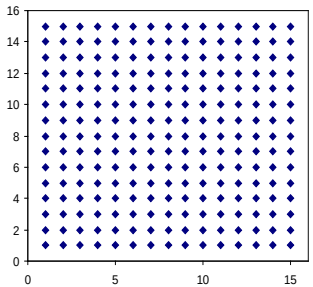
Random



Clumped

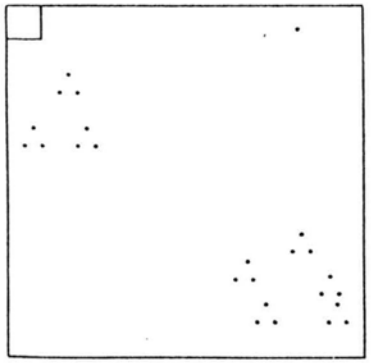


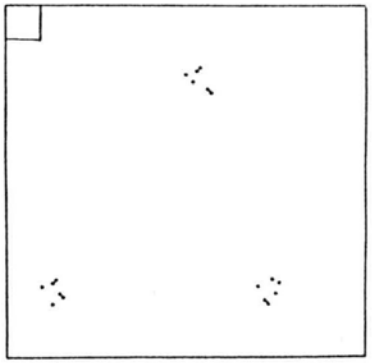
Regular (Dispersed)



- Other options exist

Clumps of Regular

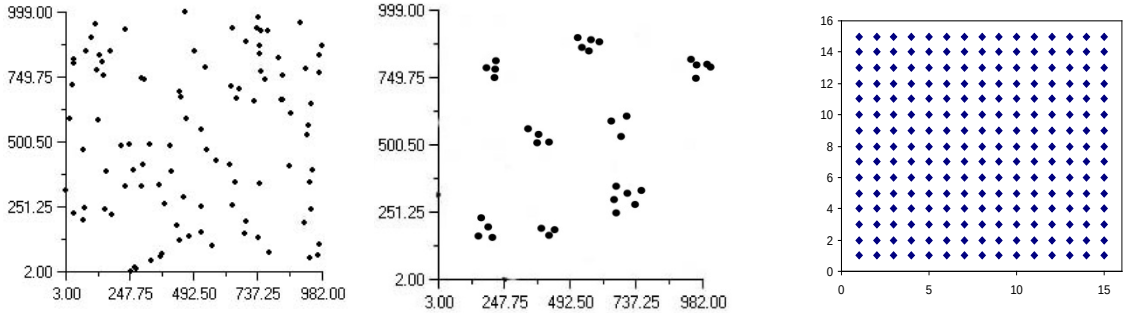




Regular Clumps

- How does one statistically assess these distributions?
  - **Area Partition Methods** and **Distance Methods**

- Are points random, clumped, or regular?



- 1st order nearest neighbor index (index of aggregation: R)
  - Calculate nearest neighbor distances (NND)
  - Determine average NND: $\overline{r}_{Obs}$
  - Determine ratio of observed vs. expected*: $R = \frac{\overline{r}_{Obs}}{\overline{r}_{Exp}}$
    - R < 1: clumped
    - R = 1: random
    - R > 1: regular (dispersed)
    - Test for significance (convert to Z-score)

*see Clark and Evans (1954: *Ecology*) for description of obtaining expected values (based on bivariate random data: expected value obtained analytically: see their Appendix)

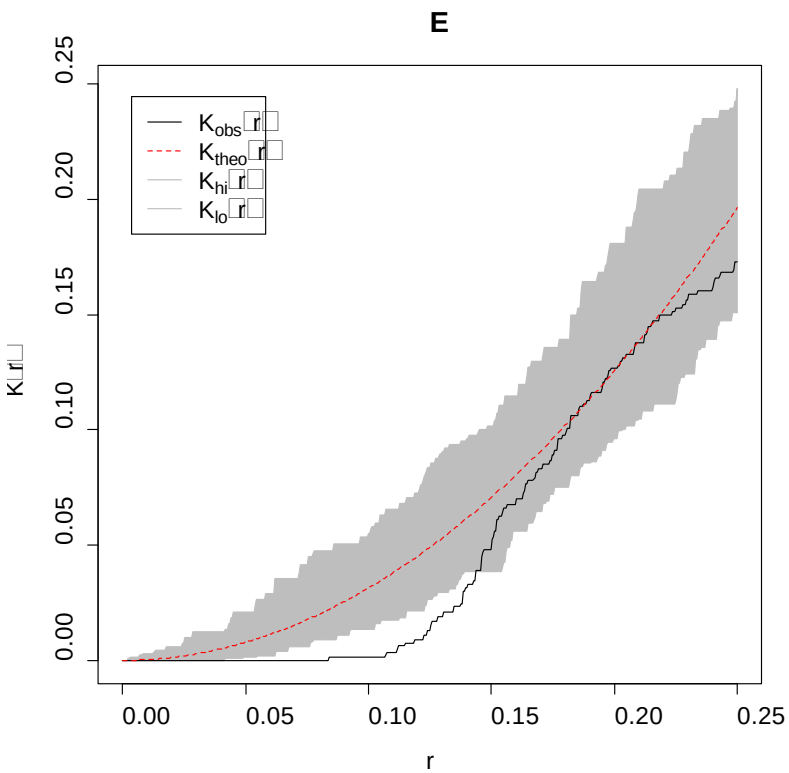
- Models the point pattern relative to some process

$$K = \frac{E(pts)}{\lambda}$$

E(pts) = expected # points in some area; λ = density of points

- *K* describes # points in area relative to expectation
- Use *K* to test hypotheses of clustering or regular dispersion

  Assessed statistically via stochastic simulations

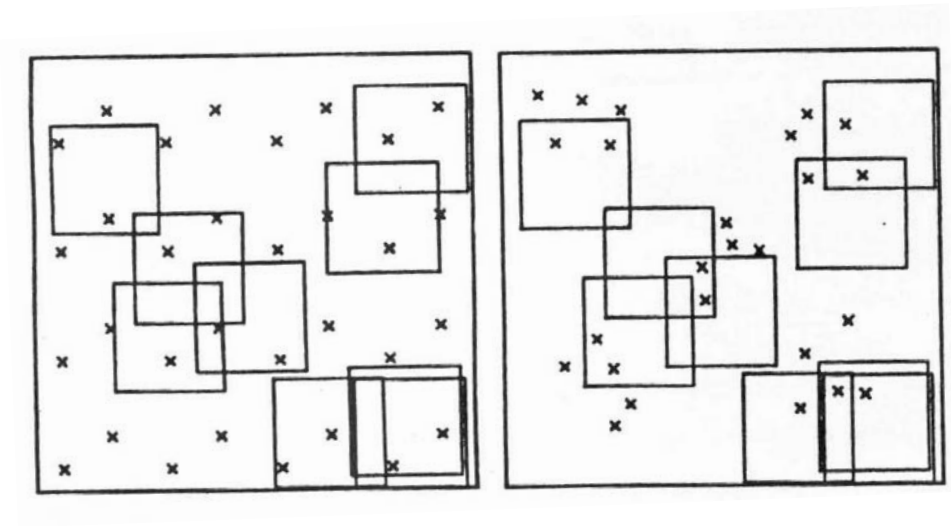


•Quadrat Method:

- •Break into quadrats and count objects (X) in each of the *n* quadrats
- •Calculate **Index of Dispersion** (variance – mean ratio):

$$ID = \frac{\sigma_X^2}{\bar{X}}$$

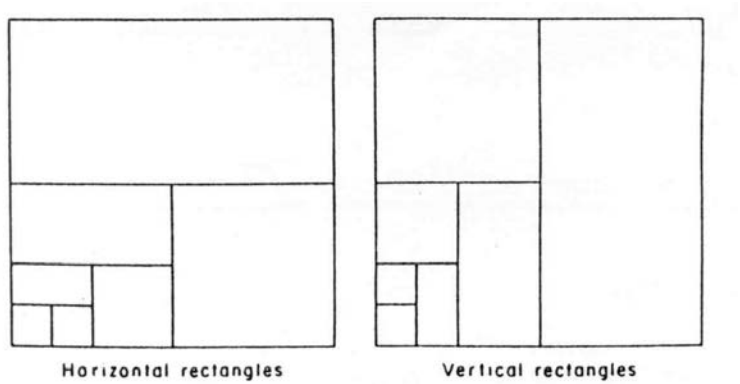
ID*(n-1) follows a X^2 with (*n*-1) df



- •ID Can estimate whether points differ from a random dispersion, but not HOW they differ (i.e., clumped vs. regular)

•Contiguous Quadrat Method

- •Break area into successively smaller quadrats nested one within the other, such that each quadrat is half the size of its 'parent'
- •Count objects in each quadrat
- •For each quadrat size (r) calculate: , $G_r = 2T_r - T_{2r}$ where Tr is the sum of squares for the rth quadrat size
- •Test each quadrat size using ratio: $\frac{G_r}{G_1}$ (Compare to F with N/2r and N/2 df)

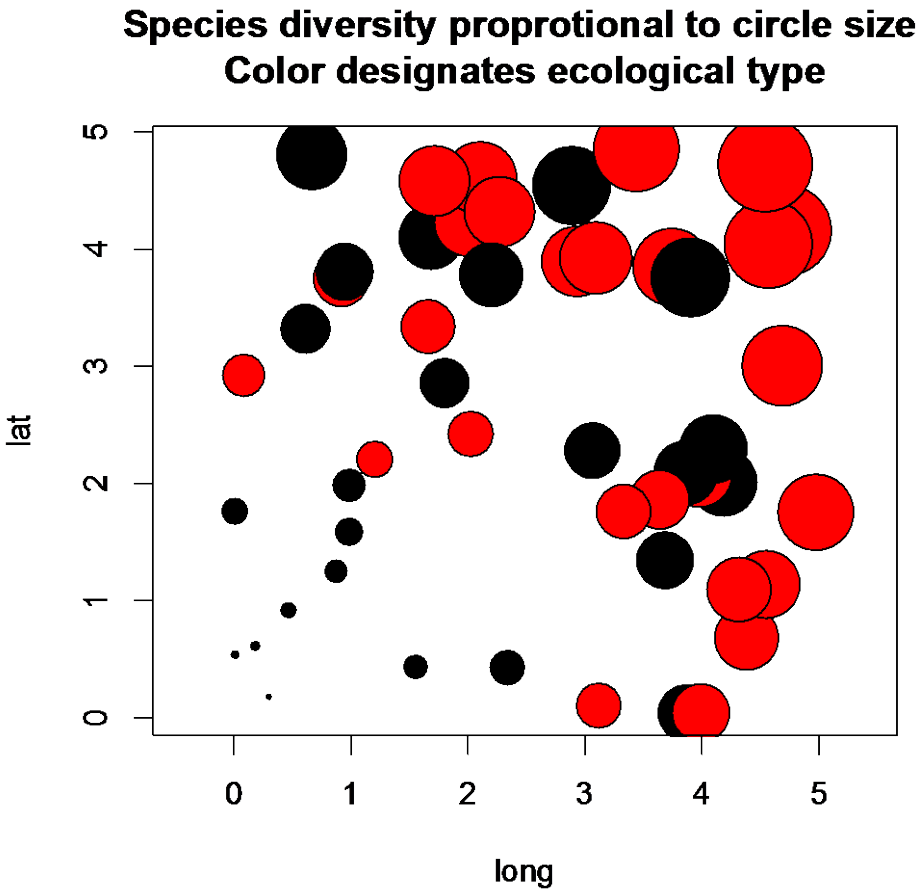


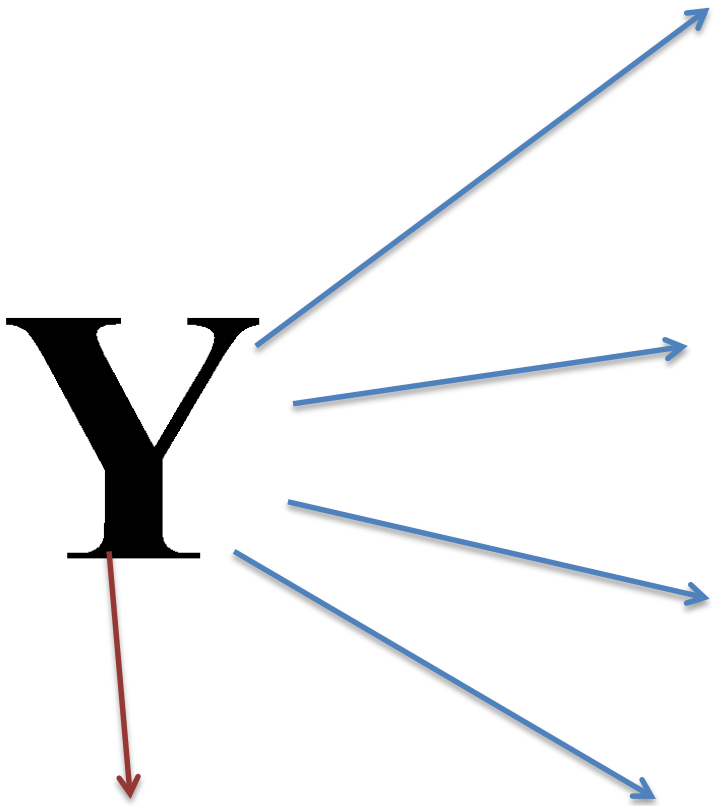
- •Random points: constant Gr with r; Clumped points: Gr peaks at the clump size (r); Regular points: Gr decreases and bottoms out at regular grain size (r)

- •Note: There are **many** other dispersion indices (see Dale, 1999; Rosenberg, 2003)

- Use distances among objects to determine whether they are farther apart, or closer together than expected by chance (randomness)
- **Plant-Plant Distance (W):** Calculate distance between objects and their 1st, 2nd, 3rd (etc.) nearest neighbors
- **Point-Plant Distance (X):** Calculate distance between randomly chosen locations and their 1st, 2nd, 3rd (etc.) nearest neighbor objects
- **Point-Plant-Plant Distance (Y):** choose random location, find the nearest object to it, and calculate the distance from that object to its 1st, 2nd, 3rd (etc.) nearest neighbors
- Determining whether the distribution of objects is different from Poisson random is basically asking the probability that another object is within an area radius W (or X, Y)
- MANY different tests for randomness have been proposed based on these measurements

- One can also compare observed data to randomly generated point patterns


Species diversity proprotional to circle size
Color designates ecological type
lat
long
0
1
2
3
4
5



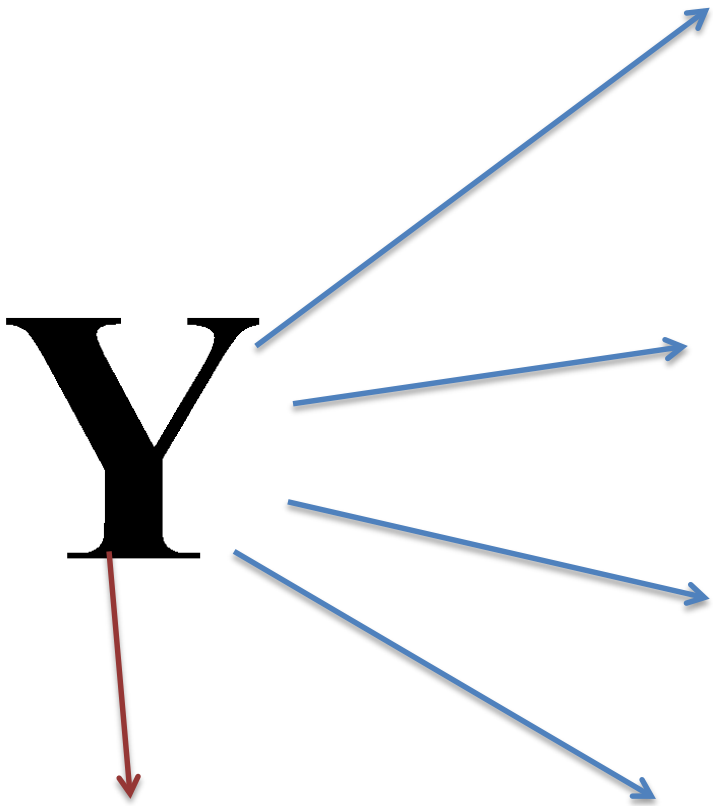
Are the data associated with geography ? (More precisely, are changes found among subjects associated with changes in geography?)

Are the data spatially autocorrelated? (Not the same as associated – more precise as a definition)

If my data have spatial autocorrelation or association with geography, how can I assess ecological hypotheses independent of geographical confounding?

Of the various methods to address the previous concerns, which are suited for what kind of circumstances?

How are my observations distributed in space? Is there a particular pattern that describes their dispersion? I.e., point patterns? (Left for a workshop or course on spatial analysis)



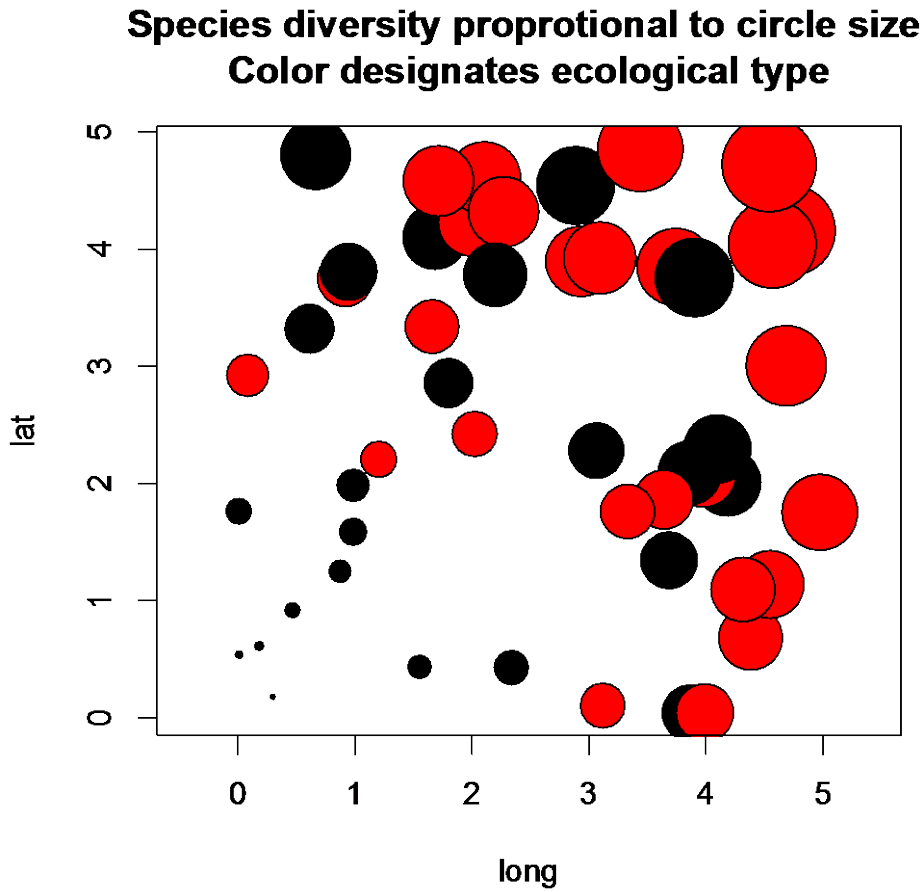
Are the data associated with geography ? (More precisely, are changes found among subjects associated with changes in geography?)

Are the data spatially autocorrelated? (Not the same as associated – more precise as a definition)

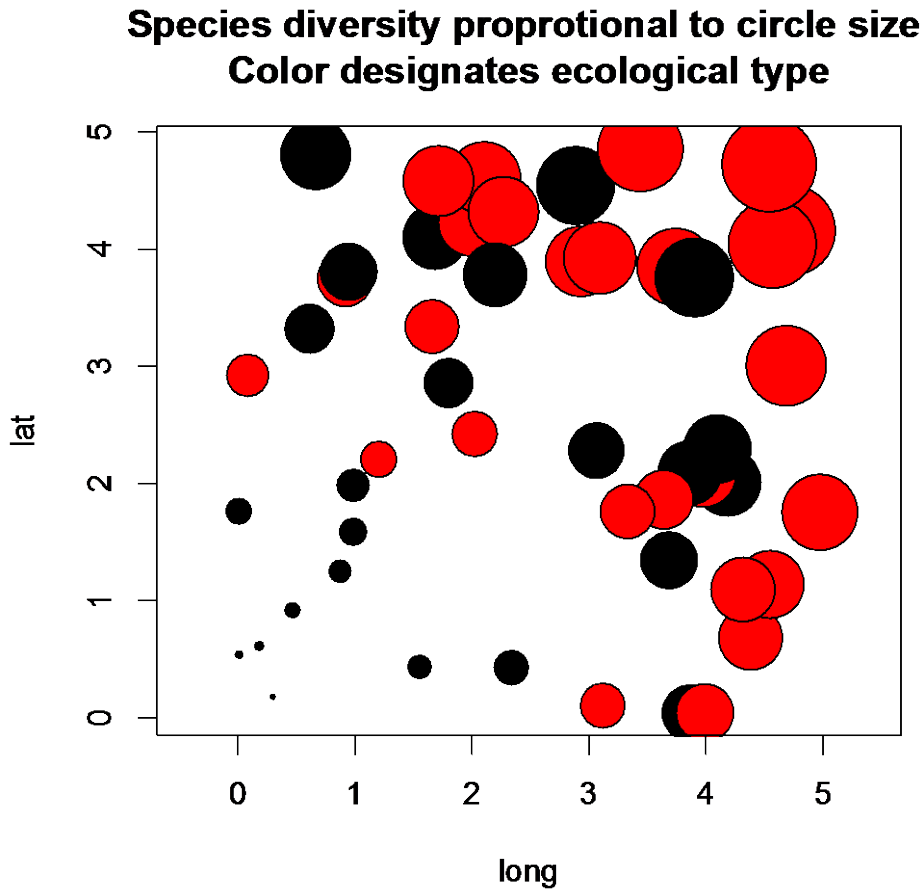
If my data have spatial autocorrelation or association with geography, how can I assess ecological hypotheses independent of geographical confounding?

Of the various methods to address the previous concerns, which are suited for what kind of circumstances?

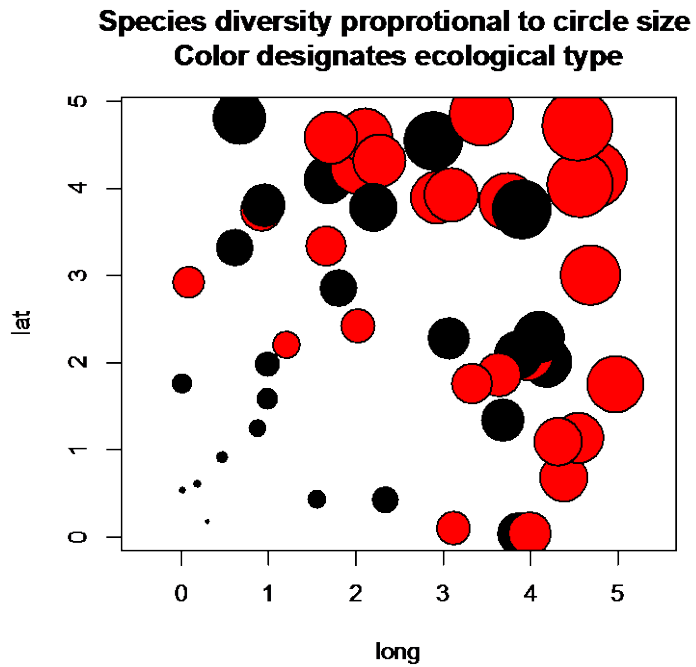
How are my observations distributed in space? Is there a particular pattern that describes their dispersion? I.e., point patterns? (Left for a workshop or course on spatial analysis)



With multivariate data, imagine vectors at each dot pointing in any direction, and with different lengths. Imagine that those near each other tend to be similar in species diversity (or whatever multivariate data are used) and point in similar directions



In this example, species diversity and geography are obviously associated. The Mantel test is the exact same approach for univariate and multivariate data, since it uses distance matrices. (The Euclidean distance between two values of one variable is the absolute difference between them.)



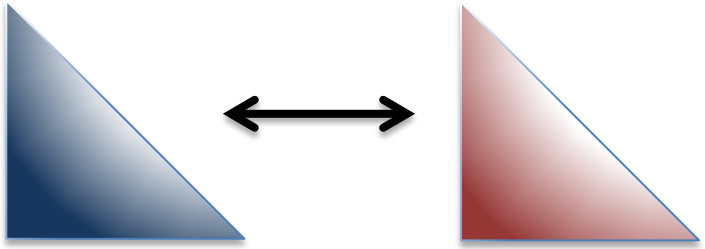

```
> # Mantel Test to measure and test association between geography and species
diversity
>
> library(vegan)
> mantel(dist(g), dist(y), permutations = 9999)

Mantel statistic based on Pearson's product-moment correlation

Call:
mantel(xdis = dist(g), ydis = dist(y), permutations = 9999)

Mantel statistic r: 0.6219
      Significance: 1e-04

Empirical upper confidence limits of r:
   90%    95%  97.5%    99%
0.0630 0.0824 0.1006 0.1252

Based on 9999 permutations
```

$$z_M = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} X_{ij} Y_{ij}$$

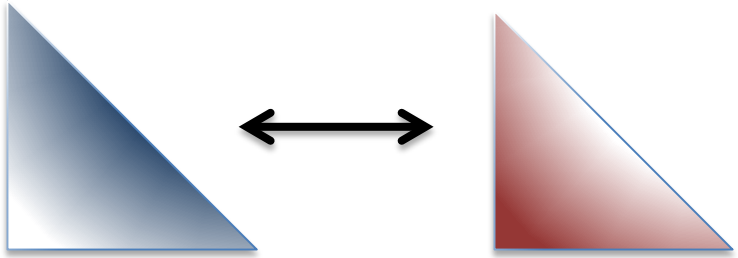
$$r_M = z_M \Big/ \left[n(n-1)/2 \right] - 1$$

In this example, species diversity and geography are obviously associated. The Mantel test is the exact same approach for univariate and multivariate data, since it uses distance matrices. (The Euclidean distance between two values of one variable is the absolute difference between them.)

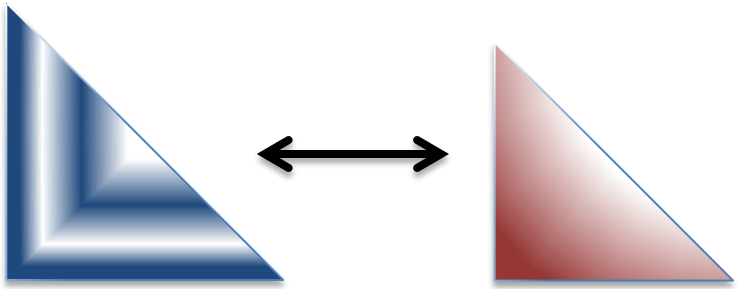
Basic design of Mantel method
Design matrices: **Y** = blue; geographic distance = red
Color intensity = greater difference



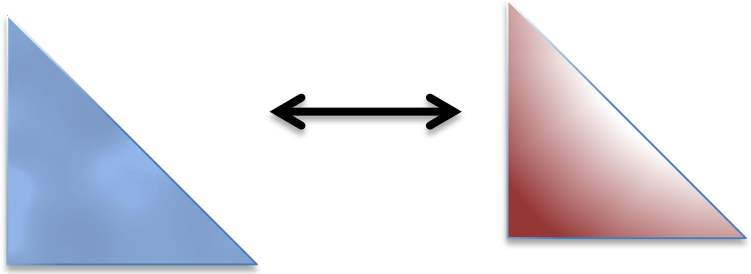
Positive association between data and geographic distance



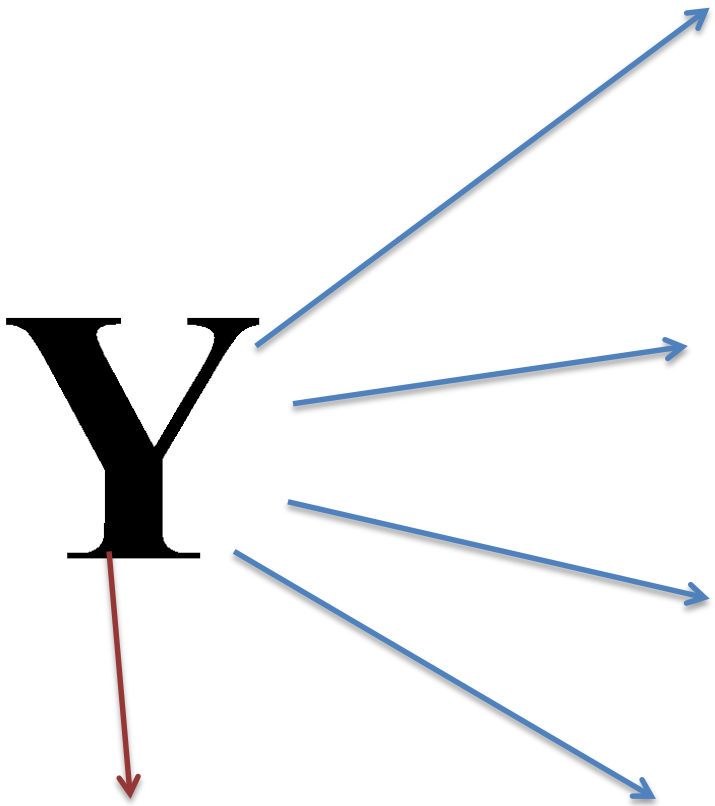
Negative association between data and geographic distance



Non-linear association between data and geographic distance (no association)



No association between data and geographic distance (no association)



Are the data associated with geography ? (More precisely, are changes found among subjects associated with changes in geography?)

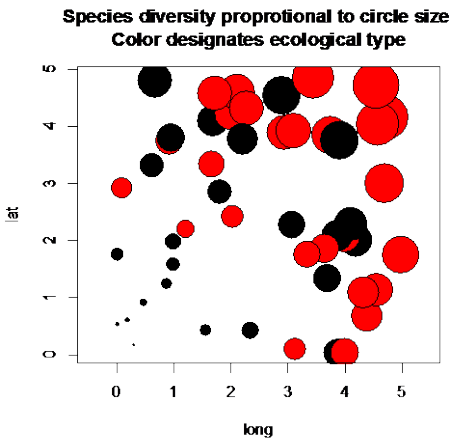
Are the data spatially autocorrelated? (Not the same as associated – more precise as a definition)

If my data have spatial autocorrelation or association with geography, how can I assess ecological hypotheses independent of geographical confounding?

Of the various methods to address the previous concerns, which are suited for what kind of circumstances?

How are my observations distributed in space? Is there a particular pattern that describes their dispersion? I.e., point patterns? (Left for a workshop or course on spatial analysis)

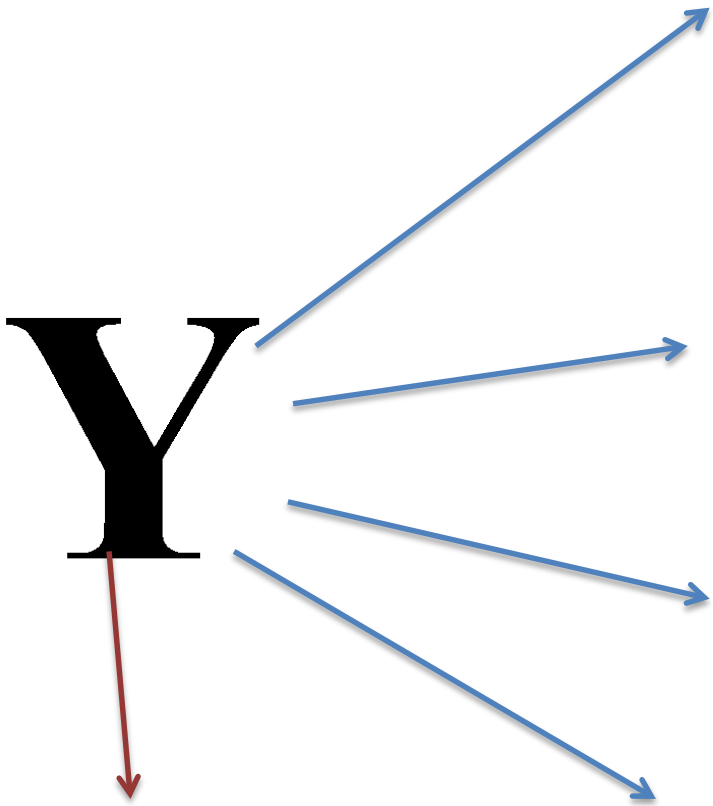
- If Y & X covary geographically, one can ‘partial out’ (account) effect of geography
- Conceptually: Y~X |Z
  - Assesses relationship of Y & X while holding Z constant




```
> # Mantel Test to test for ecological differences, after
accounting for geographical variation in species diversity
>
> mantel.partial(dist(y), dist(t), dist(g), permutations =
9999)

Partial Mantel statistic based on Pearson's product-
moment correlation

Call:
mantel.partial(xdis = dist(y), ydis = dist(t), zdis = dist(g),
permutations = 9999)

Mantel statistic r: 0.1064
      Significance: 3e-04

Empirical upper confidence limits of r:
   90%    95%  97.5%    99%
0.0231 0.0334 0.0436 0.0569

Based on 9999 permutations
```

- Recall however, that Mantel test can have high type I error, bias, and may not account for such autocorrelation when present (see Guillot and Rousset 2013. *Methods Ecol. Evol.*)

- Use other methods (below)



Are the data associated with geography ? (More precisely, are changes found among subjects associated with changes in geography?)

Are the data spatially autocorrelated? (Not the same as associated – more precise as a definition)

If my data have spatial autocorrelation or association with geography, how can I assess ecological hypotheses independent of geographical confounding?

Of the various methods to address the previous concerns, which are suited for what kind of circumstances?

How are my observations distributed in space? Is there a particular pattern that describes their dispersion? I.e., point patterns? (Left for a workshop or course on spatial analysis)

What produces SAC?

1. Biological processes (e.g., assortative mating)
2. Ecological processes/Natural Selection/Adaptation/Phenotypic plasticity
3. Non-linear relationships between species and environment modeled as linear
4. An undetected environmental covariate, which is, itself, SAC.

Spatial dependency, but not SAC (Legendre et al. 2002. Ecography 25: 601-615)

$$\mathbf{Y} = \mathbf{X}\beta + \mathbf{e}$$ i.i.d.

Independent identically distributed

Assumptions that cannot be met when there is spatial-autocorrelation (SAC) – the tendency of observations to be similar to or expressly different than neighbor observations.
Similar to = positive SAC; expressly different than = negative SAC

What produces SAC?

1. Biological processes (e.g., assortative mating)
2. Ecological processes/Natural Selection/Adaptation/Phenotypic plasticity
3. Non-linear relationships between species and environment modeled as linear
4. An undetected environmental covariate, which is, itself, SAC.

$$\mathbf{Y} = \mathbf{X}\beta + \boxed{\mathbf{e}\ \text{i.i.d.}}$$

Two steps

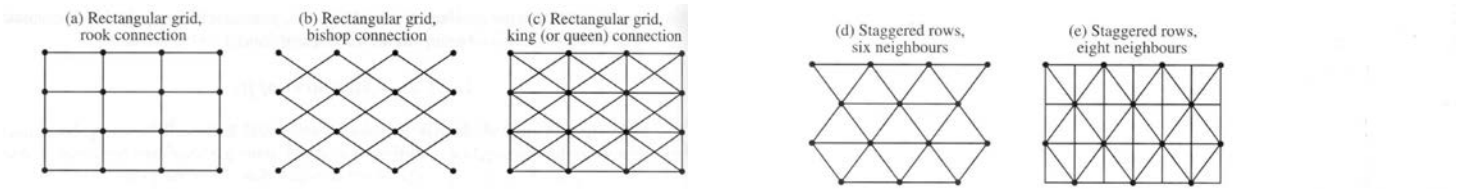
1. Determine if geography could influence variation in **Y**
2. Describe the association and/or account for it to estimate **β** properly (i.e., model must change)

- Spatial Autocorrelation: lack of independence of values based on spatial properties (self-correlation)
- Can test whether observed values at one locality depend (at least in part) on those at neighboring localities.
- Knowing the spatial auto-correlation ‘structure’ informs one how to model ecological factors

- Spatial Autocorrelation: lack of independence of values based on spatial properties (self-correlation)
- Can test whether observed values at one locality depend (at least in part) on those at neighboring localities.
- Knowing the spatial auto-correlation 'structure' informs one how to model ecological factors

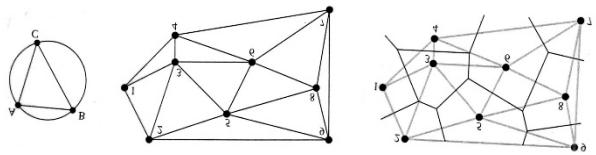
- Procedure:
  - Obtain data for each locality (Y) (can be continuous or categorical)
  - Determine which localities are 'neighbors' (or distance classes)
    - Neighbors represented in a CONNECTIVITY MATRIX (or **weight matrix)**
  - Estimate spatial autocorrelation for neighbors/distance class

Regularly-spaced data: W from ‘chessboard’ methods:

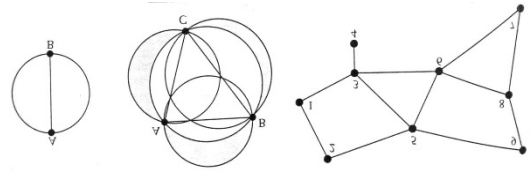


Irregularly-spaced data: Spatially-defined neighbors (many methods)

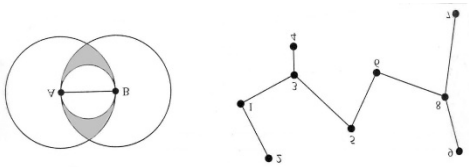
Delauney tesselation/ theisson polygons



Gabriel network



Relative neighborhood



Divide distances into distance-classes: equal interval or equal frequency

For categorical data (A,B)

Join-Count statistic: sums the number of connections of a given type:

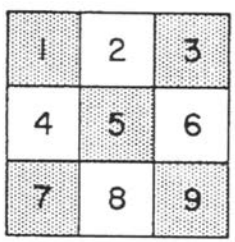
Convert to Z-score and assess vs. normal distribution

$$r_{AA} = \frac{1}{2}\sum_{ij} w_{ij}(AA)_{ij}$$

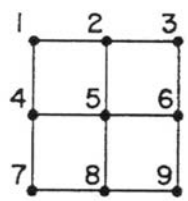
w_{ij} is the weight (0 = not connected, 1 = connected), $(AA)_{ij}$ is A-A connection (divide by ½ because upper and lower diagonal *W* are identical)

Note: negative or positive depends on how you make the connections (e.g., bishop vs. rook yield complementary -/+ values)

A



B



Localities i	j	Weight w_{ij}	$(BB)_{ij}$	$(BW)_{ij}$
1	2	1	0	1
1	3	0	1	0
1	4	1	0	1
1	5	0	1	0
1	6	0	0	1
1	7	0	1	0
1	8	0	0	1
1	9	0	1	0
2	1	1	0	1
2	3	1	0	1
⋮	⋮	⋮	⋮	⋮
9	7	0	1	0
9	8	1	0	1
			$\sum_{ij} w_{ij}(BB)_{ij} = 0$	$\sum_{ij} w_{ij}(BW)_{ij} = 24$

$r_{BB} = 0$

$r_{BW} = 12$

z-scores are: BB = -3.4156 WW = -2.245 BW = 3.456

Conclusion: significant negative spatial autocorrelation

- For continuous data, statistic is still Σ (weights * data)
- **Moran's *I***: range –1 to +1 (- are negative autocorrelation, + are positive autocorrelation)

$$I = \frac{n\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}\left(y_i - \hat{y}\right)\left(y_j - \hat{y}\right)}{W\sum_{i=1}^{n}\left(y_i - \hat{y}\right)^2}$$

w_{ij} is the weight given site *j*'s influence over site *i*, y_i is data value ($\hat{y}$ estimate for that distance class), $W = \Sigma\ (w_{ij})$

- **Geary's *c*:** range 0 to positive values (0 is positive autocorrelation, + values are negative)

$$c = \frac{(n-1)\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}\left(y_i - y_j\right)^2}{2W\sum_{i=1}^{n}\left(y_i - \hat{y}\right)^2}$$

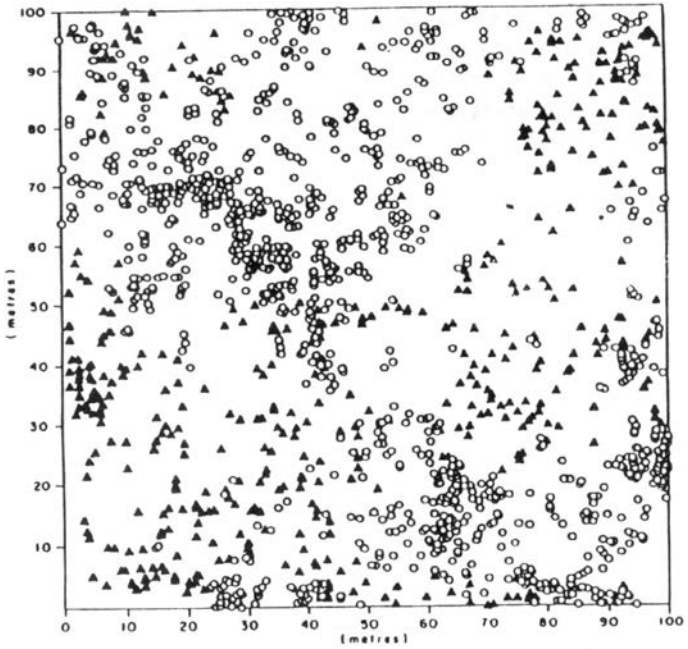
- Significance based on normal approximation (see Legendre and Legendre [1998] for details), or a permutation method

$$z_I = \frac{I_{obs} - E(I)}{\sigma_I} \qquad E(I) = -\frac{1}{n-1} \qquad E(c) = 1$$

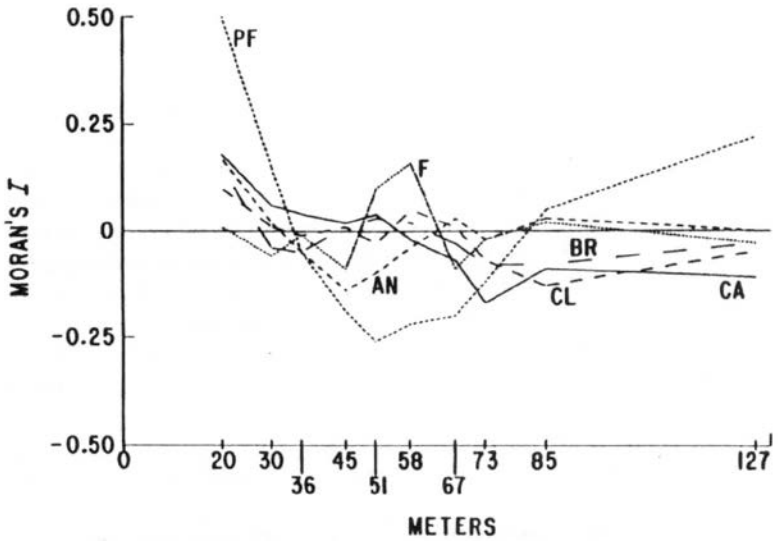
Note: **Semivariograms** based on unstandardized Geary's c (Geary's c = semivariogram standardized by variance).

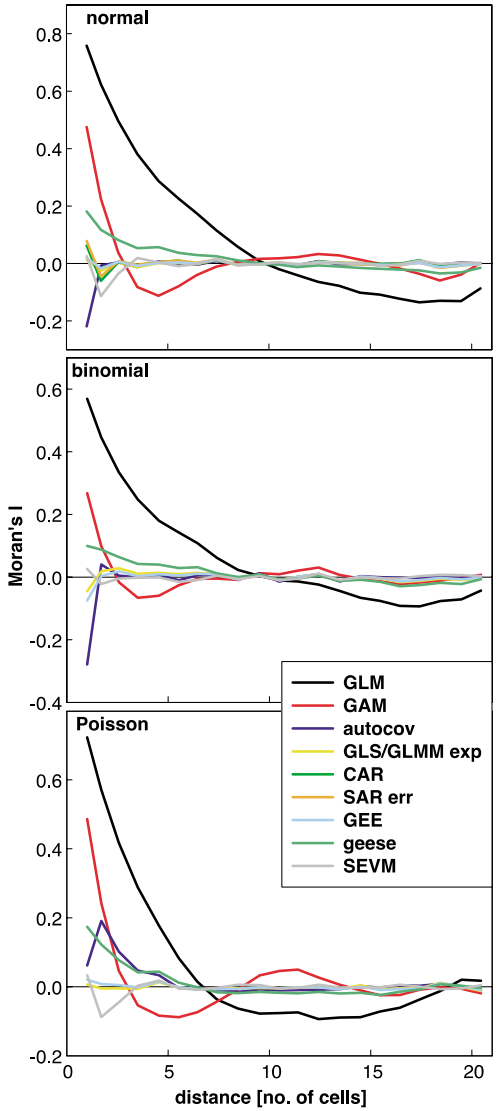
Sokal and Thompson (1987) examined distribution and attributes of understory plant, *Aralia nudicaulis*, (fecundity, density, canopy cover, percent female)

-Significant spatial autocorrelation for most variables (except fecundity), none of which showed the same pattern (i.e. variables exhibited autocorrelation, but not spatial association)



AN=*Aralia* density
CA=canopy cover
F=fecundity
PF=percent female



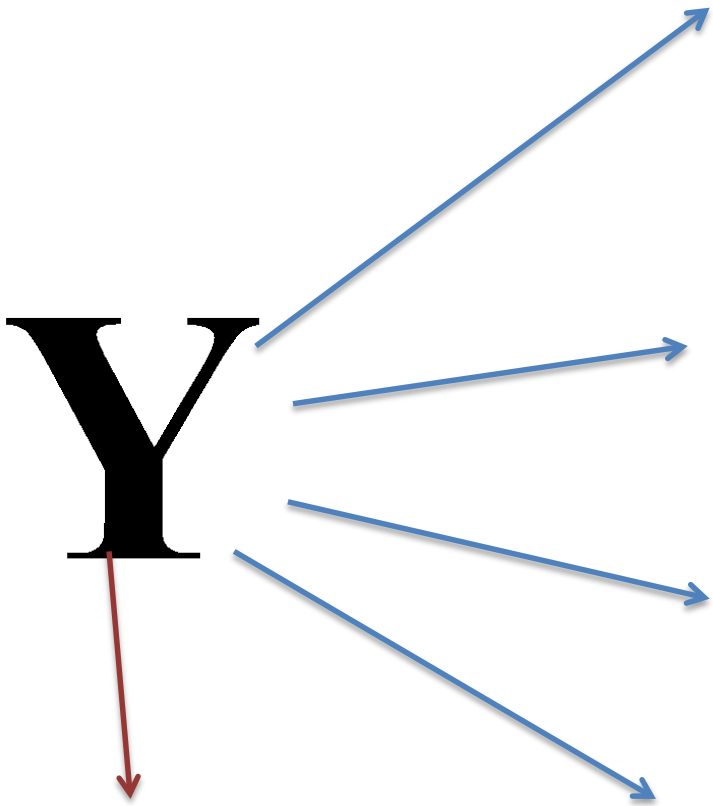


ee Fig. 1 for abbreviations.

Moran's I is generally plotted as a discrete stepwise trend over geographic distance (bins)

Important point! In answering if data are spatially autocorrelated, one must have a **model** to consider. Autocorrelation is a measure of the coordinated residual size and direction of observations found at a certain geographic distance between subjects. Most models won't be bad at large distances. Moran's I allows one to ask if spatial autocorrelation is perhaps intruding at small or intermediate distances.

Dormann et al. (2007) Ecography.



Are the data associated with geography ? (More precisely, are changes found among subjects associated with changes in geography?)

Are the data spatially autocorrelated? (Not the same as associated – more precise as a definition)

If my data have spatial autocorrelation or association with geography, how can I assess ecological hypotheses independent of geographical confounding?

Of the various methods to address the previous concerns, which are suited for what kind of circumstances?

How are my observations distributed in space? Is there a particular pattern that describes their dispersion? I.e., point patterns? (Left for a workshop or course on spatial analysis)

$$\mathbf{Y} = \mathbf{X}\beta + \boldsymbol{\varepsilon}$$

$$\sigma^2\Sigma$$

$$\mathbf{Y} = \mathbf{X}\beta + \varepsilon$$

$$\Sigma^{-1}\mathbf{Y} = \Sigma^{-1}\mathbf{X}\beta + \Sigma^{-1}\varepsilon$$

$$\Sigma^{-1}\sigma^2\Sigma$$

$$\sigma^2\mathbf{I} \sim \mathbf{N}\left(0, \sigma^2\right)$$

Σ = Spatial covariance matrix, based on either an observed or theoretical correlation among subjects (based on geographic distances between them)

- Account for spatial non-independence in S of using GLS & related models

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \qquad \boldsymbol{\varepsilon} = \boldsymbol{\sigma}^2\boldsymbol{\Sigma}$$

$$\boldsymbol{\beta} = \left(\mathbf{X}^t\boldsymbol{\Sigma}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}^t\boldsymbol{\Sigma}^{-1}\mathbf{Y}$$

So what is Σ?

Depends….. It can be variations of the same concept. It is ALWAYS an *n* x *n* matrix with values estimating the expected covariance or correlation-like coefficients between subjects, based on the geographic distance between them.

The coefficients can be directly proportional to distance (linear) or a theoretical decay of expected similarity associated with distance.

- Σ describes covariance among locations, with coefficients proportional to distance (or expected decay similarity as function of distance)

- MANY models proposed to describe spatial non-independence

$$Exponential = \sigma^2 e^{-r/d}$$

$$Guassian = \sigma^2 e^{(-r/d)^2}$$

$$Spherical = \sigma^2 \left(1 - 2/\pi \left(r/d \sqrt{1 - r^2/d^2} + \sin^{-1} r/d \right) \right)$$

See: Dormann et al. (2007) Ecography.

Simultaneous autogressive models (SAR)

$$\mathbf{Y} = \boxed{\rho\mathbf{WY}} + \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

(which is equivalent to $\mathbf{Y} = (\mathbf{I} - \rho\mathbf{W})^{-1}\mathbf{X}\boldsymbol{\beta} + (\mathbf{I} - \rho\mathbf{W})^{-1}\boldsymbol{\varepsilon}$), where ρ is the autoregression parameter, $\mathbf{W}$ the spatial weights matrix, and $\boldsymbol{\beta}$ a vector representing the slopes associated with the predictors in the original predictor matrix $\mathbf{X}$.

Adjust predicted values based on adjusting observed values, by the autoregressive scalar times the spatial weights between subjects – little effect if subjects are far apart in geographic space. **W** can be estimated in various ways.

Recall, **W** is an *n* x *n* matrix

Autoregressive models, ρ = autoregression parameter

See: Dormann et al. (2007) Ecography.

Simultaneous autogressive models (SAR)

$$\mathbf{Y} = \rho\mathbf{WY} + \mathbf{X\beta} + \boldsymbol{\varepsilon}$$

(which is equivalent to $\mathbf{Y} = (\mathbf{I} - \rho\mathbf{W})^{-1}\mathbf{X\beta} + (\mathbf{I} - \rho\mathbf{W})^{-1}\boldsymbol{\varepsilon}$), where ρ is the autoregression parameter, $\mathbf{W}$ the spatial weights matrix, and $\boldsymbol{\beta}$ a vector representing the slopes associated with the predictors in the original predictor matrix $\mathbf{X}$.

Second, spatial autocorrelation can affect both response and predictor variables ("lagged-mixed model", SAR mix). Ecologically, this adds a local aggregation component to the spatial effect in the lag-model above. In this case, another term ($\mathbf{WX}\boldsymbol{\gamma}$) must also appear in the model, which describes the regression coefficients ($\boldsymbol{\gamma}$) of the spatially lagged predictors ($\mathbf{WX}$). The SAR lagged-mixed model takes the form

$$\mathbf{Y} = \rho\mathbf{WY} + \mathbf{X\beta} + \mathbf{WX}\boldsymbol{\gamma} + \boldsymbol{\varepsilon}$$

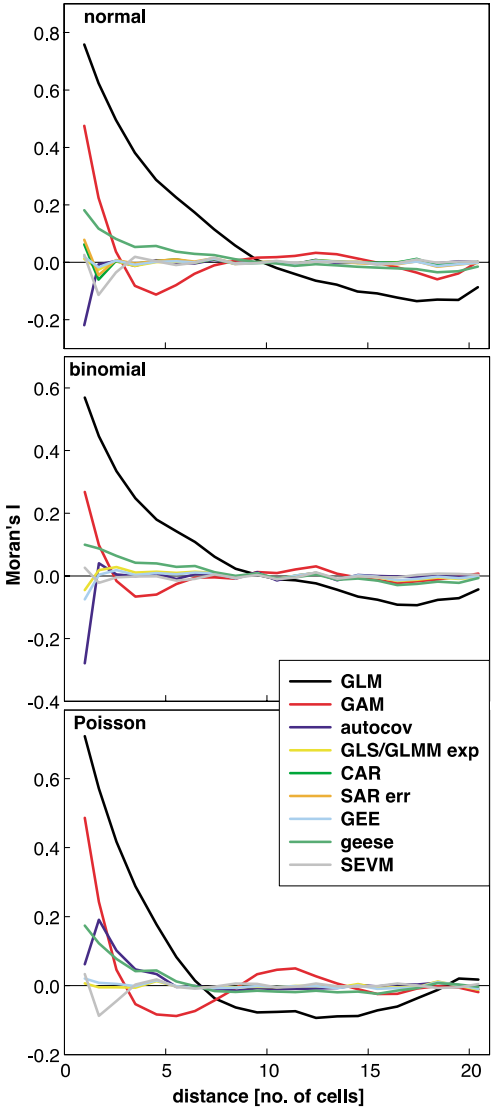
Adjust predicted values based on adjusting observed values and an error edition (because maybe both the explanatory and response variables are auto correlated).

Recall, **W** is an *n* x *n* matrix

See: Dormann et al. (2007) Ecography.

GLM = General Linear Model
GAM = trend-surface General Additive Model (no autocorrelation models – just adds geographic trends to the general linear model)
autocov = Autocovariate model –
GLMM = GLS with mixed effects model (allows for spatial autocorrelation within groups)
CAR = Conditional Autoregressive models
SAR = Sequential Autoregressive models
GEE = Generalized Estimating Equations (like GLM, but allows for different spatial weighting for "clusters" of subjects)
geese = GEE but with greater constraint on spatial correlations based on certain distances
SEVM = PCoA on Gower centered distance matrix, followed by adding eigenvectors as variables to GLM until Moran's I is no longer significant

In all cases, GLS performed about the best!



See: Dormann et al. (2007) Ecography.

ee Fig. 1 for abbreviations.

- Spatial context hugely important to consider in biology

1: How are my observations distributed?

Point Pattern Analysis

2: Are my data associated with geography?

Mantel tests

3: Are my data spatially autocorrelated?

Moran's I

4: Can I account for spatial autocorrelation in my analysis?

GLS is general and very useful model