Lesson 8:
Generalised Least Squares

Assumptions of linear models

$$y_{ijk} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + e_{ijk}$$

Linear models make many assumptions, including:

1. The model makes biological sense/ physical sense
2. Additivity (terms are added together)
3. Linearity
4. Independence of errors
5. Homoscedasticity – equal variance of errors
6. Normality of errors.

Two situations where GLS is useful

1. Data heteroscedasticity (unclear mean-variance relationships)

2. Autocorrelation between data points (non-independence)

Linear model with ‘good’ errors (IID)

$$\boldsymbol{y} = \boldsymbol{X}\beta + \varepsilon \quad \text{where} \quad \varepsilon \sim \boldsymbol{N}(\boldsymbol{0}, \boldsymbol{V})$$

Where **V** = Cov[$\underline{\varepsilon}$| **X**], the covariance matrix of errors. Ideally:

1. ε_i have even range across predicted y (homoscedasticity)
2. off-diagonal elements of **V** are zero (no autocorrelation)

Y = X * β + ε where **$\underline{\varepsilon}$ ~ N(0, σ^2I)**

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \cdot \\ \cdot \\ y_{n-1} \\ y_n \end{bmatrix} = \begin{array}{c} \begin{array}{ccc} \text{Int} & \text{X1} & \text{X2} \end{array} \\ \begin{bmatrix} 1 & x_{1,1} & x_{2,1} \\ 1 & x_{1,2} & x_{2,2} \\ 1 & x_{1,3} & x_{2,3} \\ & \cdot & \cdot \\ & \cdot & \cdot \\ 1 & x_{1,n-1} & x_{2,n-1} \\ 1 & x_{1,n} & x_{2,n} \end{bmatrix} \end{array} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \cdot \\ \cdot \\ \varepsilon_{n-1} \\ \varepsilon_n \end{bmatrix} \quad \text{where} \quad \boxed{\begin{bmatrix} \varepsilon_{1,1} & 0 & 0 & \dots & 0 & 0 \\ 0 & \varepsilon_{2,2} & 0 & \dots & 0 & 0 \\ 0 & 0 & \varepsilon_{3,3} & \dots & 0 & 0 \\ \cdot & \cdot & \cdot & & \cdot & \cdot \\ \cdot & \cdot & \cdot & & \cdot & \cdot \\ 0 & 0 & 0 & \dots & \varepsilon_{n-1,n-1} & 0 \\ 0 & 0 & 0 & \dots & 0 & \varepsilon_{n,n} \end{bmatrix}}$$

Linear model with heteroscedasticity and/ or error covariance

In reality $\underline{\varepsilon} \sim N(0, \mathbf{V})$, where :

1. ε_i are heteroscedastic
2. Off-diagonal elements are non-zero (covariance)

Y = X * β + ε where **$\underline{\varepsilon}$ ~ N(0, V)**

$$
\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \cdot \\ \cdot \\ y_{n-1} \\ y_n \end{bmatrix}
=
\begin{array}{c}
\begin{array}{ccc} \text{Int} & \text{X1} & \text{X2} \end{array} \\
\begin{bmatrix}
1 & x_{1,1} & x_{2,1} \\
1 & x_{1,2} & x_{2,2} \\
1 & x_{1,3} & x_{2,3} \\
 & \cdot & \cdot \\
 & \cdot & \cdot \\
1 & x_{1,n-1} & x_{2,n-1} \\
1 & x_{1,n} & x_{2,n}
\end{bmatrix}
\end{array}
\begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix}
+
\begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \cdot \\ \cdot \\ \varepsilon_{n-1} \\ \varepsilon_n \end{bmatrix}
\text{ where }
\boxed{\begin{bmatrix}
w_1\varepsilon_{1,1} & \varepsilon_{1,2} & \varepsilon_{1,3} & \cdots & \varepsilon_{1,n-1} & \varepsilon_{1,n} \\
\varepsilon_{2,1} & w_2\varepsilon_{2,2} & \varepsilon_{2,3} & \cdots & \varepsilon_{2,n-1} & \varepsilon_{2.n} \\
\varepsilon_{3,1} & \varepsilon_{3,2} & w_3\varepsilon_{3,3} & \cdots & \varepsilon_{3,n-1} & \varepsilon_{3,n} \\
\cdot & \cdot & \cdot & & \cdot & \cdot \\
\cdot & \cdot & \cdot & & \cdot & \cdot \\
\varepsilon_{n-1,1} & \varepsilon_{n-1,2} & \varepsilon_{n-1,3} & \cdots & & \varepsilon_{n-1,n} \\
\varepsilon_{n,1} & \varepsilon_{n,2} & \varepsilon_{n,3} & \cdots & \varepsilon_{n,n-1} & w_n\varepsilon_{n,n}
\end{bmatrix}}
$$

Generalised least squares models

- The problem: $\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}$ where $\boldsymbol{\varepsilon} \sim \boldsymbol{N}(\boldsymbol{0}, \boldsymbol{V})$
- OLS solution: $\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} \quad \because \quad \boldsymbol{V} = \boldsymbol{I}\sigma^2$
- The GLS solution: $\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{y}$
- You might recall this from the mixed models lecture.

- We will deal with heteroscedasticity and error covariance separately in the remaining problems

To be clear:

- Heteroscedasticity is about the diagonal elements (homoscedastic residuals should be ~equal in size)

- Error covariance is about the off-diagonal elements (should be close to zero i.e. uncorrelated)

1. Heteroscedascity

$$\mathbf{Y} = \mathbf{X} * \boldsymbol{\beta} + \boldsymbol{\varepsilon} \quad \text{where} \quad \underline{\boldsymbol{\varepsilon}} \sim N(0,\ w\sigma^2 I)$$

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \cdot \\ \cdot \\ y_{n-1} \\ y_n \end{bmatrix} = \begin{array}{c} \begin{array}{ccc} \text{Int} & \text{X1} & \text{X2} \end{array} \\ \begin{bmatrix} 1 & x_{1,1} & x_{2,1} \\ 1 & x_{1,2} & x_{2,2} \\ 1 & x_{1,3} & x_{2,3} \\ & \cdot & \cdot \\ & \cdot & \cdot \\ 1 & x_{1,n-1} & x_{2,n-1} \\ 1 & x_{1,n} & x_{2,n} \end{bmatrix} \end{array} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \cdot \\ \cdot \\ \varepsilon_{n-1} \\ \varepsilon_n \end{bmatrix} \quad \text{where} \quad \boxed{\begin{bmatrix} w_1\varepsilon_{1,1} & 0 & 0 & \dots & 0 & 0 \\ 0 & w_2\varepsilon_{2,2} & 0 & \dots & 0 & 0 \\ 0 & 0 & w_3\varepsilon_{3,3} & \dots & 0 & 0 \\ \cdot & \cdot & \cdot & & \cdot & \cdot \\ \cdot & \cdot & \cdot & & \cdot & \cdot \\ 0 & 0 & 0 & \dots & w_{n-1}\varepsilon_{n-1,n-1} & 0 \\ 0 & 0 & 0 & \dots & 0 & w_n\varepsilon_{n,n} \end{bmatrix}}$$

1. Dealing with heteroscedascity

- Usually we try to define relationships between means and variances, e.g. Poisson, Binomial or Gamma distributions.

- Sometimes we might need more flexibility if we are uncertain about the mean-variance relationship

- CAVEAT: I find it quite difficult to decide when to apply these methods UNLESS I already know that the mean-variance relationship is systematically skewed and cannot be described by a known distribution

Generalised Least Squares

- GLS extends the linear model to address complex error relationships

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon} \quad \text{where} \quad \boldsymbol{\varepsilon} \sim \boldsymbol{N}(\boldsymbol{0}, \boldsymbol{I}w_{\boldsymbol{i}}\sigma^2)$$

- We aim to apply inverse weights, $1/w_i$, that are the <u>inverse</u> of the relationship $Iw_i\sigma^2$))

- Can model these w_i using the weights argument in the gls() function

The gls() function

- `mod <- gls(y~x, data=dat,`
  `weights=varFixed(form=~1/w²))`

Variance function. Options include:

1) varFixed: weights are linear
2) varPower: weights are power-law
3) varIdent: weights differ by factor level

Express the weights in terms of the covariate

The `lme()` function

- The lme() function allows one to include random effects into the mix as well

- `mod <- lme(y~x, random=~1|group, data=dat, weights=varFixed(form=~1/w^2))`

Code 8.1
Heteroscedastic models

Exercise 8.1

- Return to the plantdamage data.
- Fit a random effects model with lme to test whether damage, light and their interaction affects growth.
- Use growth in untransformed form.
- First, run the model assuming no covariance structure (mod 1) and plot the residuals.
- Inspect the lme model for evidence of heteroscedasticity. Try this command

  Ps: plot(mod, resid(.)~fitted(.)|light, abline=0, layout=c(2, 1))
- Then try varIdent and varExp models (mod 1b, mod 1c) and compare them against your first model using AIC and plot their residuals.
- Second, assume growth is exponentially distributed. Transform it and run the model assuming no co-variance structure (mod 2), and plot the residuals.
- Then try varIdent and varExp models with the transformed growth and compare them using AIC. (mod 2b, mod 2c)
- Finally compare the summaries from mod 1c and mod 2.

2. Autocorrelation

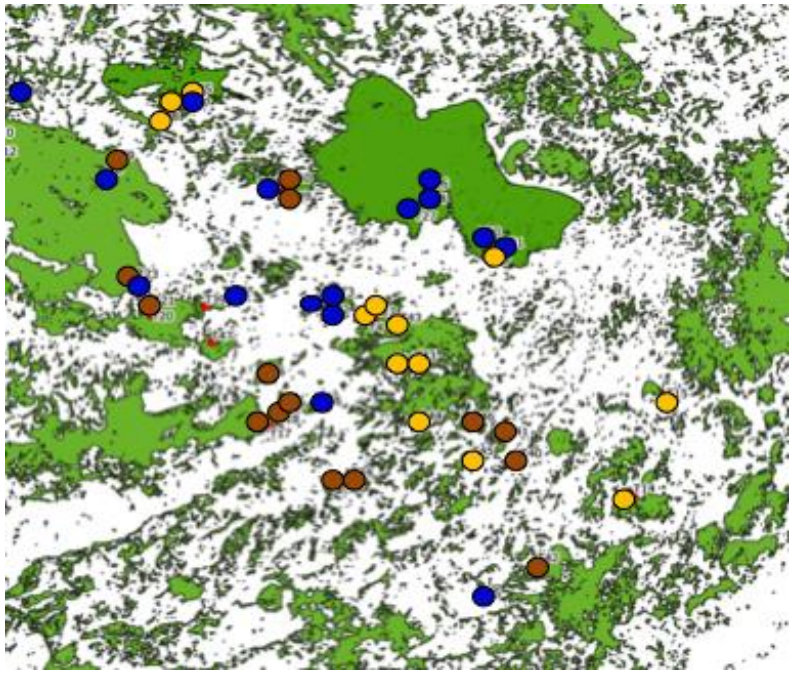
Y = X * β + ε where **ε ~ N(0, V)**

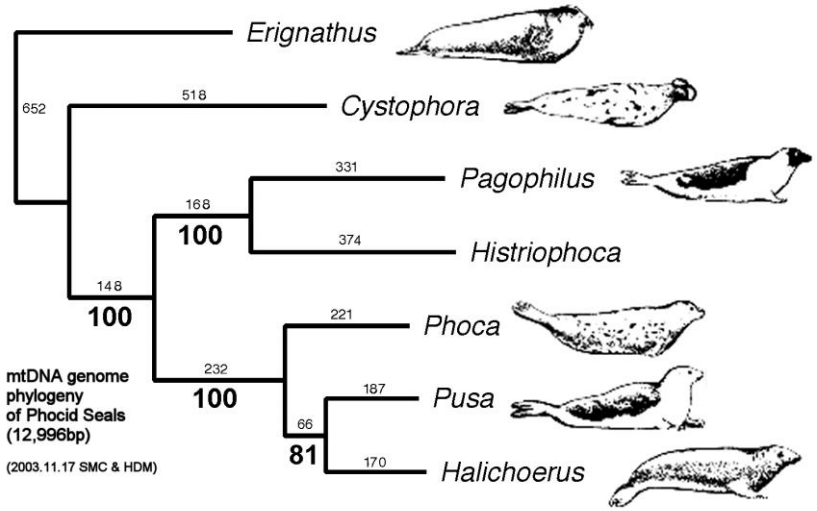
$$
\begin{array}{ccccccc}
 & & \text{Int} & \text{X1} & \text{X2} & & \\
\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ . \\ . \\ y_{n-1} \\ y_n \end{bmatrix} & = & \begin{bmatrix} 1 & x_{1,1} & x_{2,1} \\ 1 & x_{1,2} & x_{2,2} \\ 1 & x_{1,3} & x_{2,3} \\ & . & . \\ & . & . \\ 1 & x_{1,n-1} & x_{2,n-1} \\ 1 & x_{1,n} & x_{2,n} \end{bmatrix} & \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix} & + & \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ . \\ . \\ \varepsilon_{n-1} \\ \varepsilon_n \end{bmatrix} & \text{where}
\end{array}
$$

$$
\begin{bmatrix}
\varepsilon_{1,1} & \varepsilon_{1,2} & \varepsilon_{1,3} & \cdots & \varepsilon_{1,n-1} & \varepsilon_{1,n} \\
\varepsilon_{2,1} & \varepsilon_{2,2} & \varepsilon_{2,3} & \cdots & \varepsilon_{2,n-1} & \varepsilon_{2.n} \\
\varepsilon_{3,1} & \varepsilon_{3,2} & \varepsilon_{3,3} & \cdots & \varepsilon_{3,n-1} & \varepsilon_{3,n} \\
. & . & . & & . & . \\
. & . & . & & . & . \\
\varepsilon_{n-1,1} & \varepsilon_{n-1,2} & \varepsilon_{n-1,3} & \cdots & \varepsilon_{n-1,n-1} & \varepsilon_{n-1,n} \\
\varepsilon_{n,1} & \varepsilon_{n,2} & \varepsilon_{n,3} & \cdots & \varepsilon_{n,n-1} & \varepsilon_{n,n}
\end{bmatrix}
$$

2. Dealing with auto-correlation

- Data points that are close together in space, time or phylogeny are not independent.





2. Dealing with auto-correlation

- Data points that are close together in space, time or phylogeny are not independent.
- Difficult to turn these relationships into discrete blocks (as we did for LMMs).
- Possible to model the auto-correlation directly with gls()and lme() using the correlation argument.

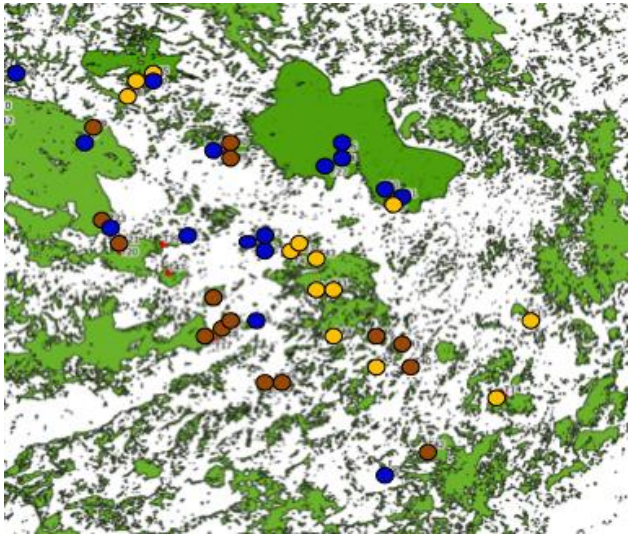
Covariance structures in gls

- All linear models have the form
- $Y = X\beta + \varepsilon$ such that $\varepsilon \sim N(0, V)$

Where sigma is a n x n matrix.

- Each value, from row *i* and column *j* is the covariance of the errors of replicates *i* and *j*
- OLS models assume that $V \sim I\sigma^2$ (in other words: off-diagonal covariances are close to zero)
- In GLS we essentially describe the relationships in V using an assumed covariance structure

Covariance between data



- Consider the correlation between plant communties in space.
- In the figure, it is clear that closer communities are more similar.
- This could be because of geology
- But it could also be because plants have a higher chance of establishing offspring closer to a mother tree
- This relatedness effect falls off as two points move further apart..
- i.e. the correlation between points decreases with distance

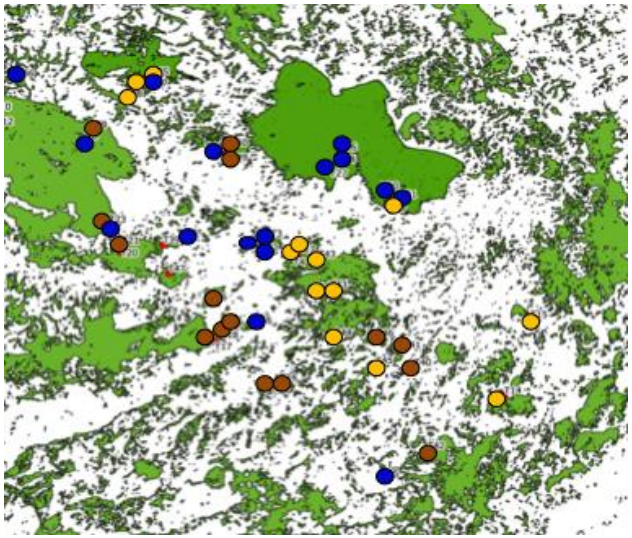
Covariance between data

Take homes:

1. Spatial autocorrelation is due to some process that we may find difficult to diagnose properly

2. The autocorrelation process often has a spatial limit
   e.g. max distance of seed dispersal

3. The autocorrelation-dist relationship has a shape!



Describing covariance between data points: Semivariograms

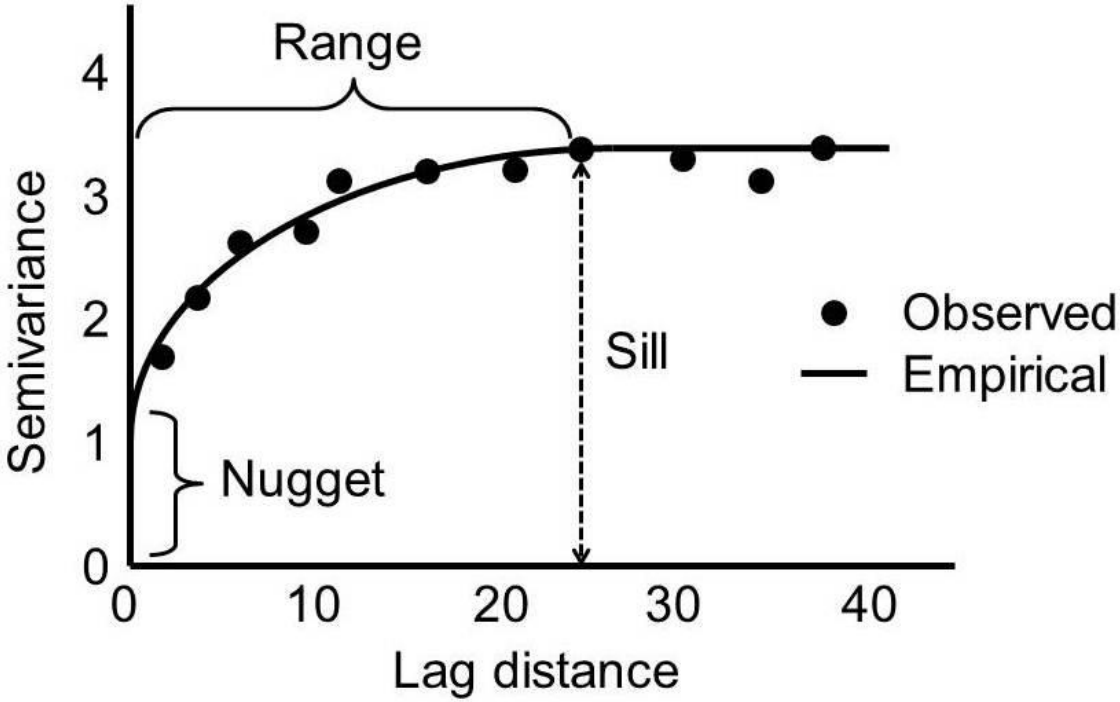
Semi-variograms describe differences (variance) between data points in terms of some distance measure (space, time, DNA).

$$\hat{\gamma}(h \pm \delta) := \frac{1}{2|N(h \pm \delta)|} \sum_{(i,j) \in N(h \pm \delta)} |z_i - z_j|^2$$

Range: distance over which points are correlated in space/ time

Sill: average distance between uncorrelated data points = variance

Nugget: some variance at small scale (stochasticity?/ measurement error?)



Fitting a correlation structure

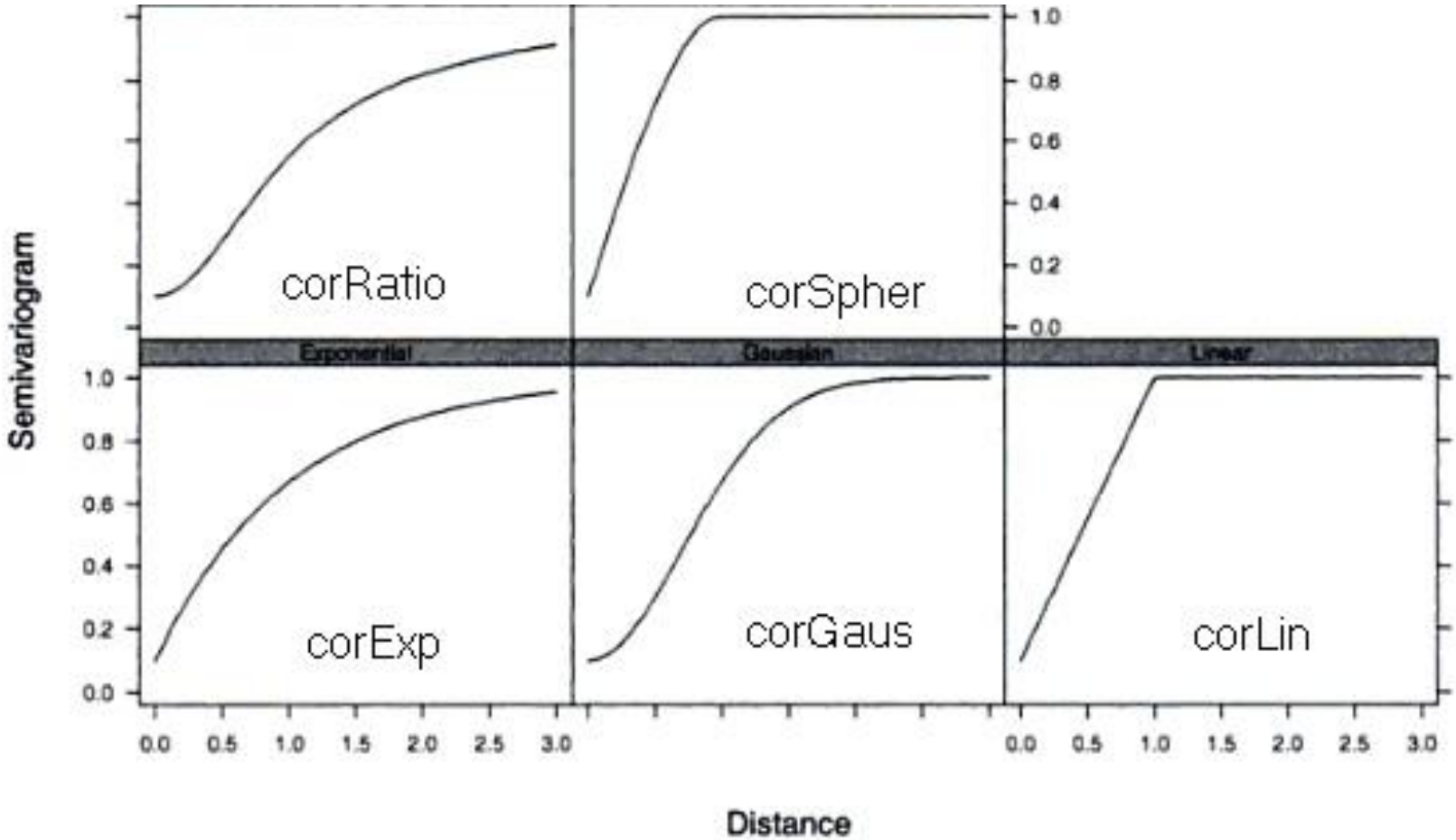
- Distance correlation relationships have pattern;

```
mod  <- gls(y~x, data=dat, correlation=corExp(form=~x+y))
```

- Dependency decreases as an exponential function of distance.

- Other options:
  - corSpher
  - corGaus
  - corLin
  - corRatio

Correlation structures for shapes



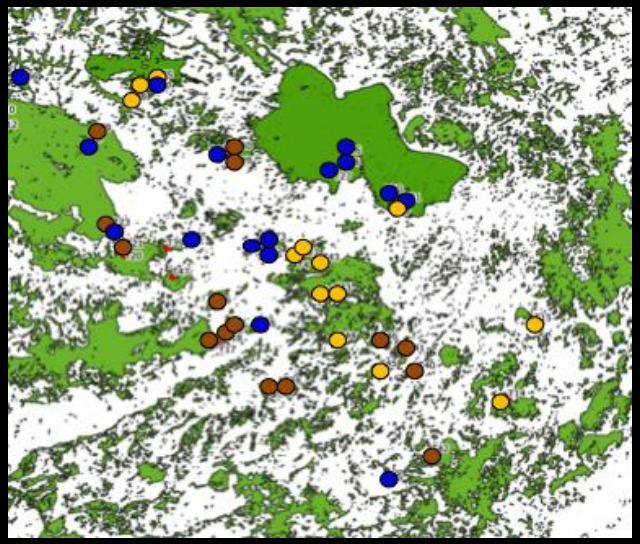
Pinheiro and Bates 2004, p. 233

Approach

- 1) fit a basic model without a correlation structure,
- 2) examine a variogram of the residuals (i.e., their correlation with respect to distance),
- 3) choose an appropriate corStruct class based on the variogram shape (maybe test a few..), and
- 4) test whether addition of the correlation term significantly improves model fit via AICs or LRTs (**anova** function).
- Use **Variogram** function in **nlme** to diagnose the type:

```
plot(Variogram(gls1,form=~lon+lat))
```

Code 8.2
Spatial correlation structures



Exercise 8.2

- The data in the file trees.csv come from an analysis of tree growth in an Austrian forest.
- The response of Relative growth rate (RGR) to light was compared between two species (beech and spruce)
- Determine if there is an interaction between species and light in their effects on RGR
- Are the residuals independent?
- Modify the models to deal with any spatial dependence