# Species as data points

## Outline for today

- The problem with species data
- Phylogenetic signal in ecological traits
- Why phylogeny matters in comparative study
- Phylogenetically independent contrasts (PICs)
- A linear model approach
- A method for discrete data (and issues)

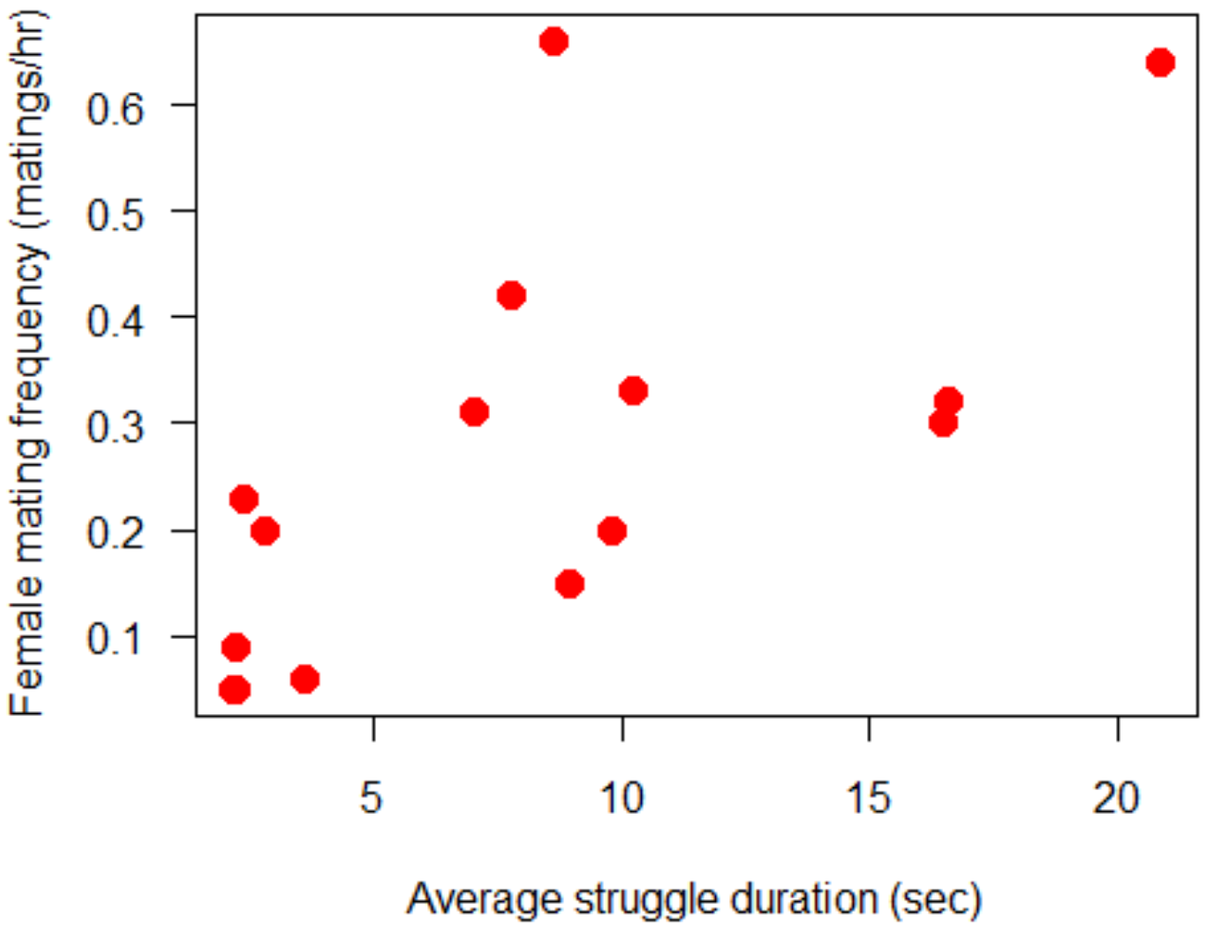
**An example of species data**

Mating behaviors in 15 species of water striders (*Gerris*). Males chase females, who flee by skating away. If a male grasps a female, she initiates a series of leaps, rolls, and summersaults that usually toss him off. Males of some species have clasping genitalia that allow them to stay on longer, but females of these species often have spines or other devices that make it difficult for males to grasp her. Mating takes place after a female stops struggling.

Rowe and Arnqvist (2002) measured average duration of female struggles for each species (the periods of evasive action by females in response to lunges or grasps by males); and average mating frequency of females, under controlled lab conditions.

**An example of species data**

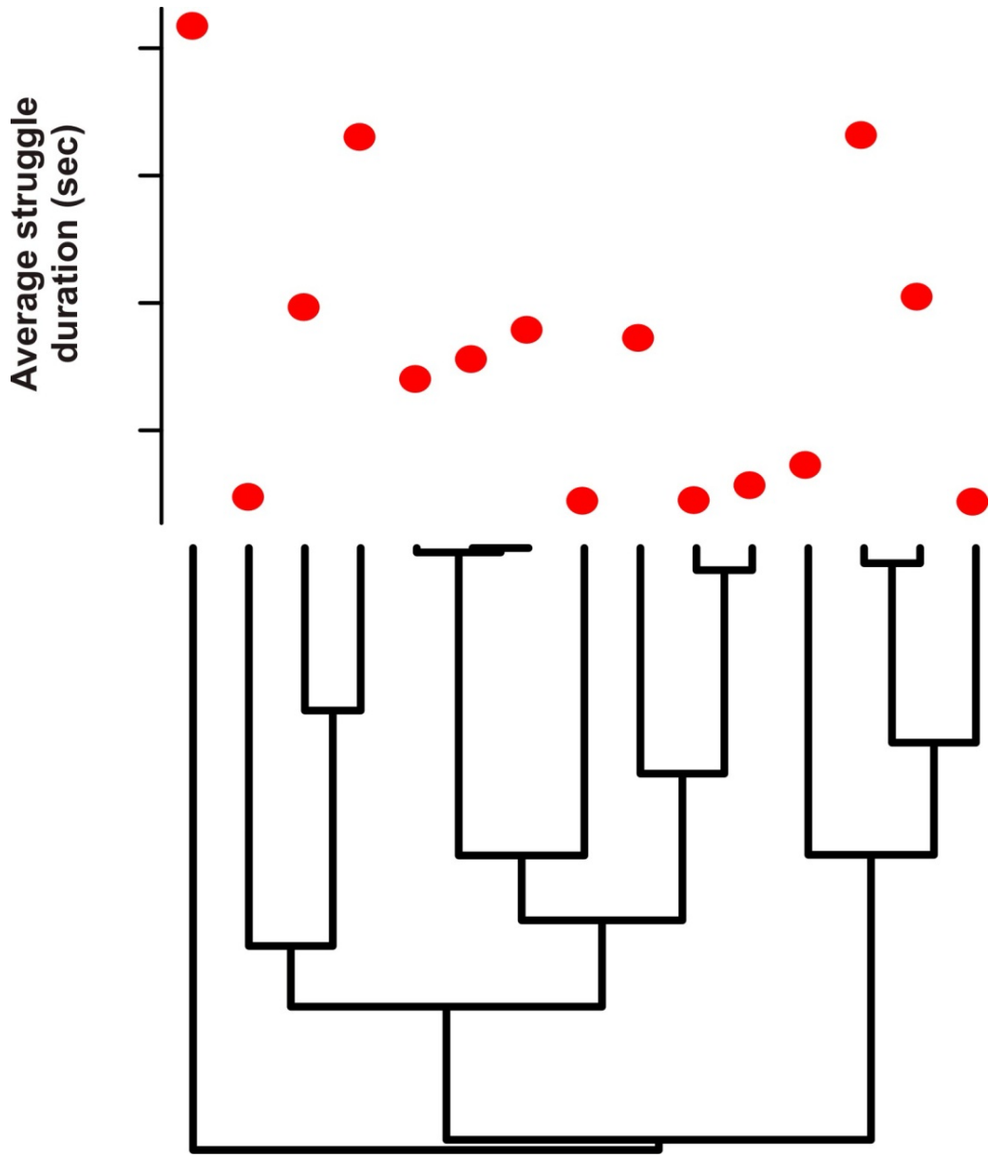
Data reveal a positive association between the two variables.
We would like to estimate the strength of the correlation.



**The problem with species data**

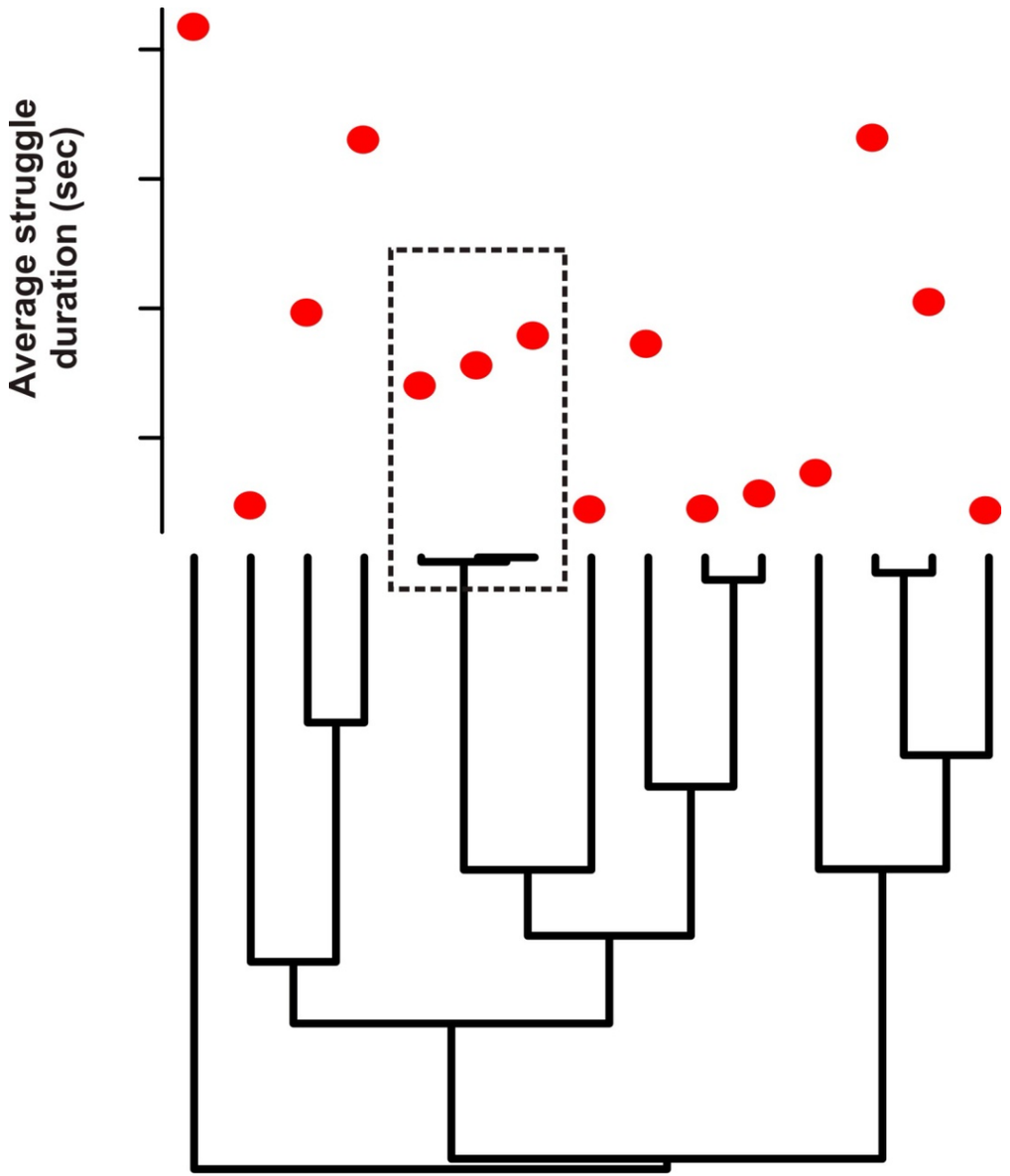
The data points (species) are not independent.

Phylogeny of *Gerris*



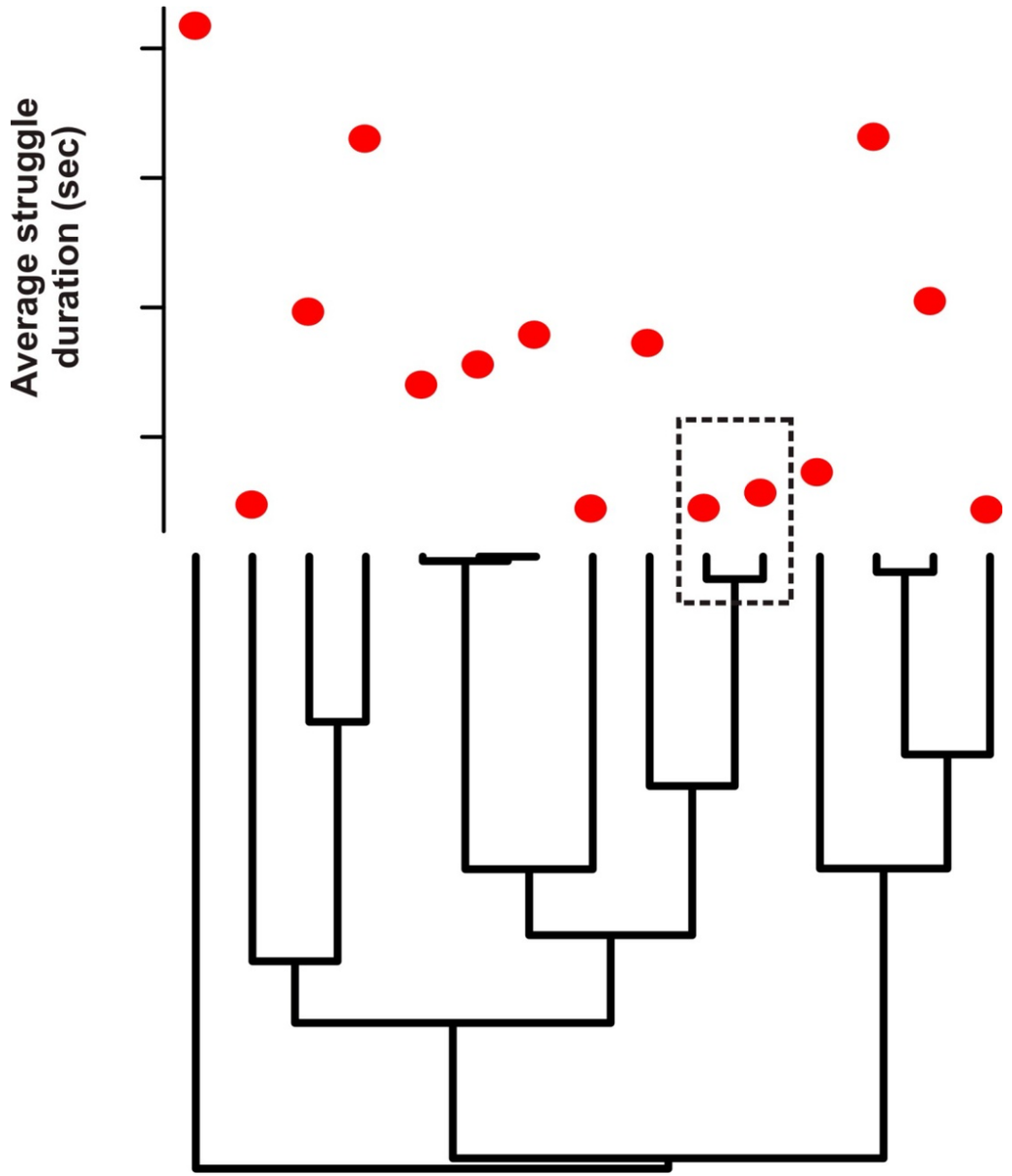
**The problem with species data**

Closely related species tend to have similar trait values.



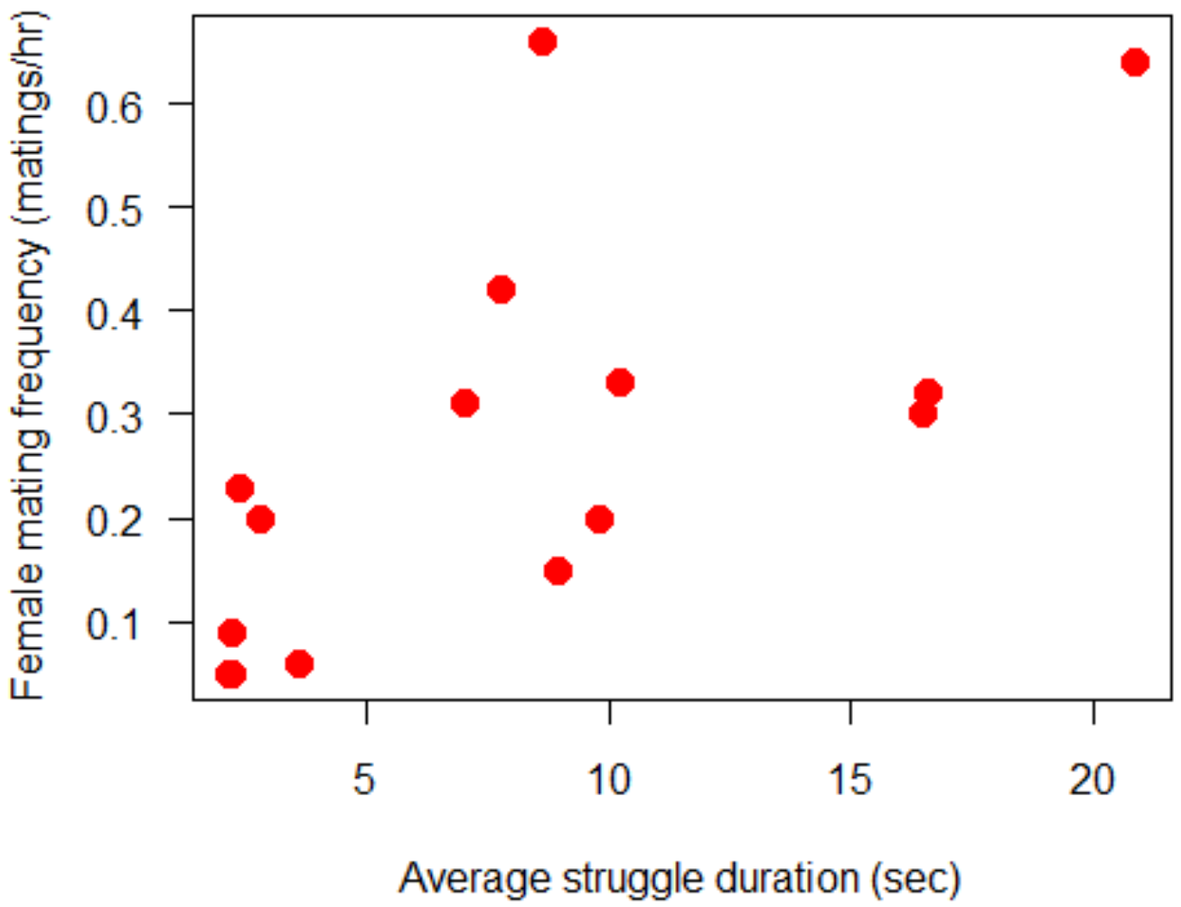
**The problem with species data**

This tendency is called “phylogenetic signal”.



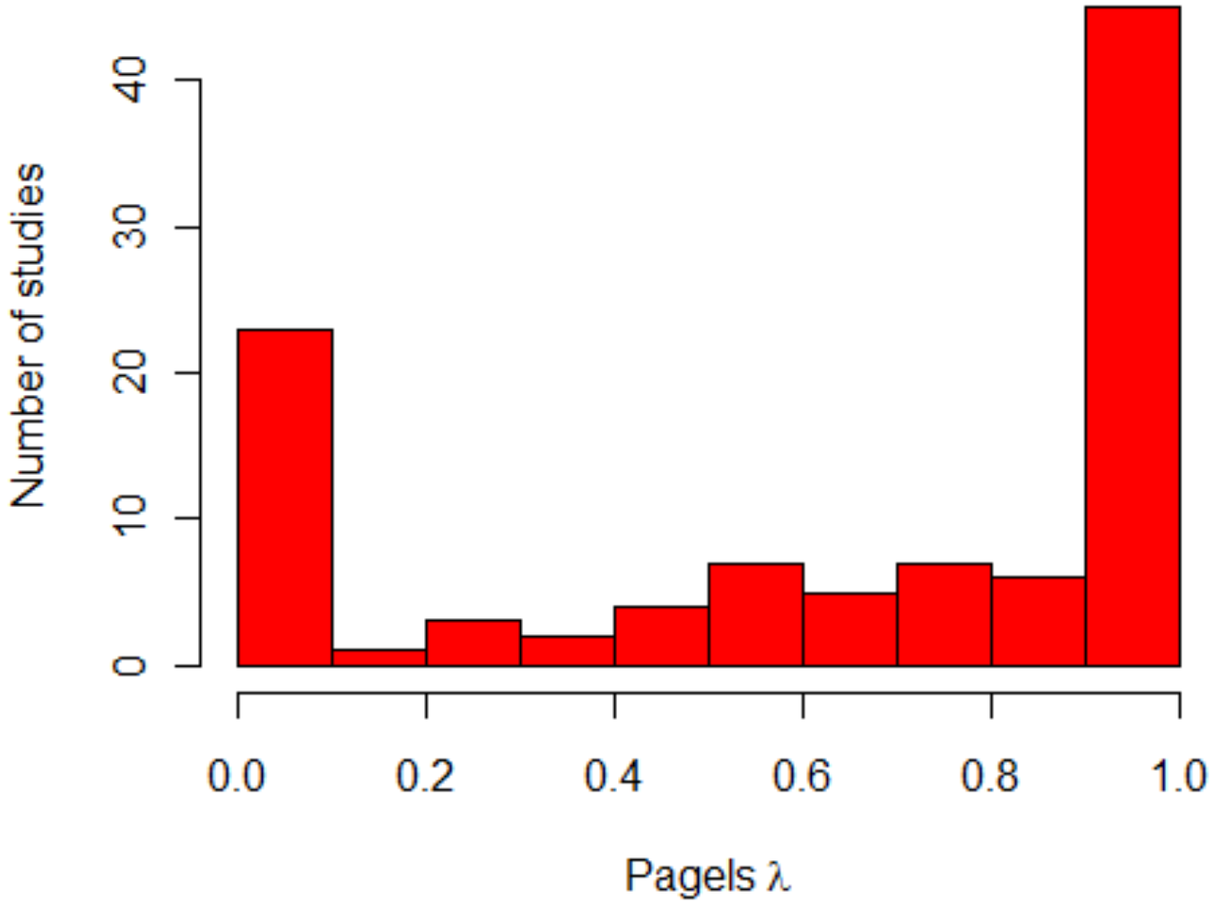
**The problem with species data**

Non-independence of the species data points violates a major assumption of conventional statistical methods for data analysis.



**How prevalent is phylogenetic signal in ecologically relevant traits?**

Pagel's $\lambda$ measures the extent to which closely related species are similar in their trait values (phylogenetic signal). Here is a survey of $\lambda$-values from many studies and traits by Freckleton et al (2002)**:**



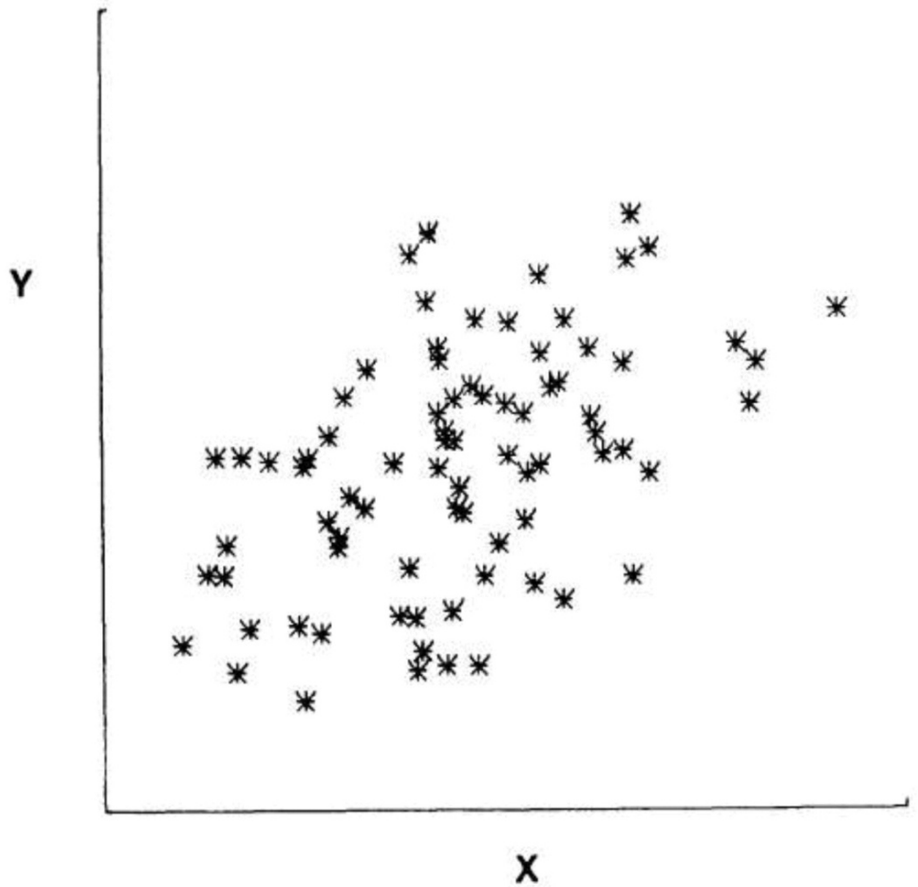
**Why is phylogenetic signal a problem?**

Non-independence leads to wrong calculations of precision (standard errors, confidence intervals). It leads to wrong Type 1 error rates in null hypothesis significance testing.

Example scenario:
Data on two traits
for 40 species

Looks like a strong
correlation between
variables Y and X

Felsenstein (1985) *Am Nat*



**Why is phylogenetic signal a problem?**

Felsenstein's "worst case scenario" for the phylogeny of the 40 species.

Felsenstein (1985) *Am Nat*

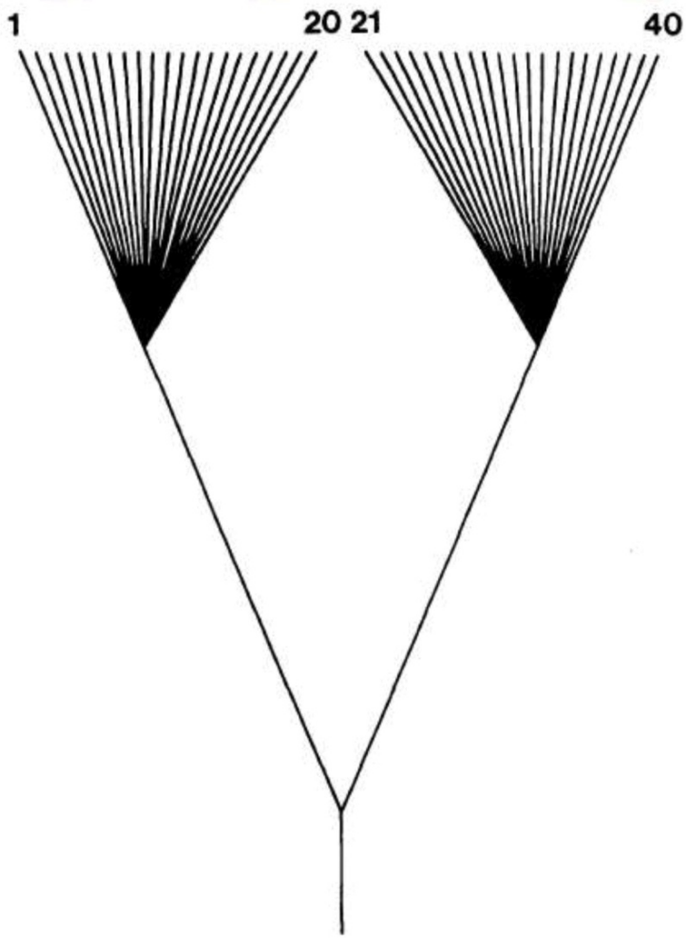



FIG. 5.—A "worst case" phylogeny for 40 species, in which there prove to be 2 groups each of 20 close relatives.

**Why is phylogenetic signal a problem?**

In this case the non-independence is severe, and creates an apparent association between X and Y where there is none.

Felsenstein (1985) *Am Nat*

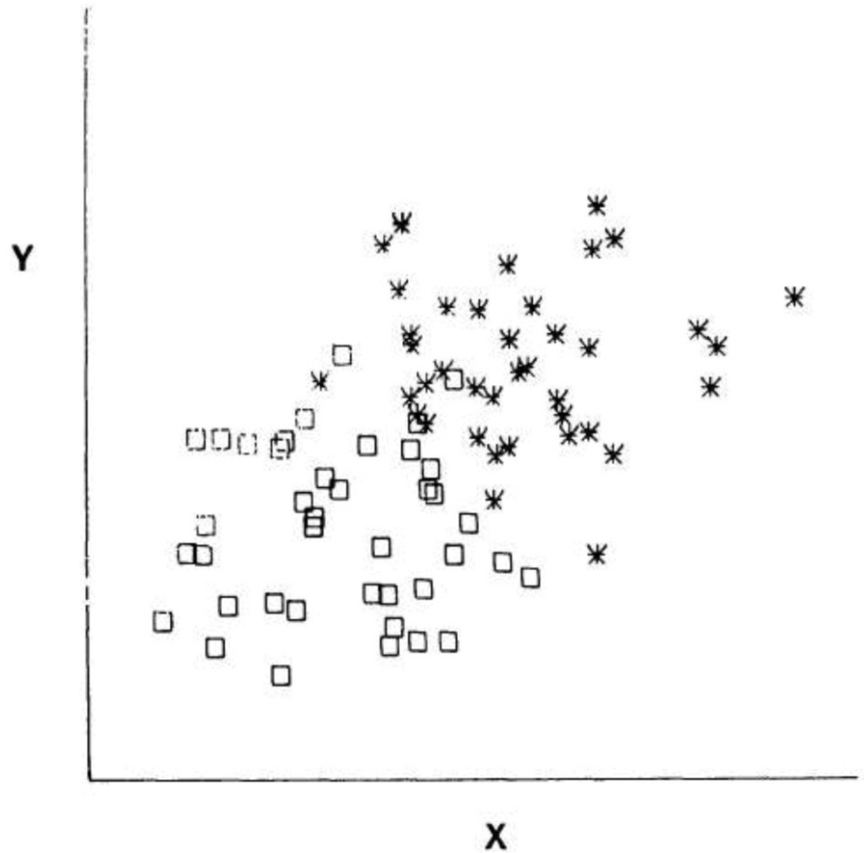
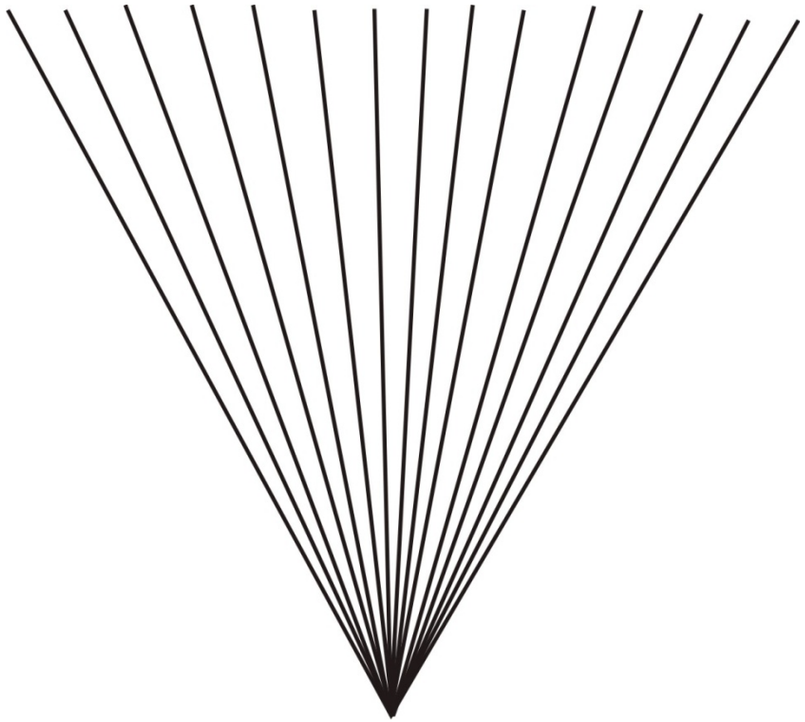
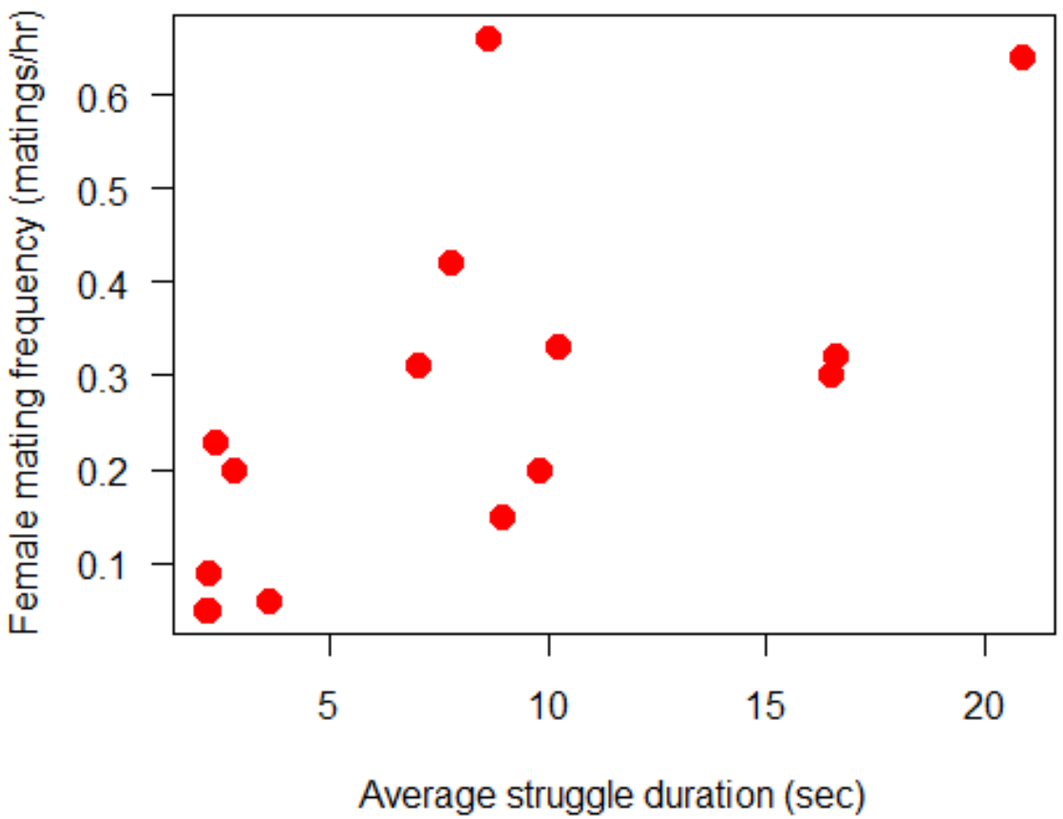



FIG. 7.—The same data set, with the points distinguished to show the members of the 2 monophyletic taxa. It can immediately be seen that the apparently significant relationship of fig. 6 is illusory.

**What we are really assuming when we ignore phylogeny**

That the species are related as in a “star” phylogeny, which leads to no phylogenetic signal.



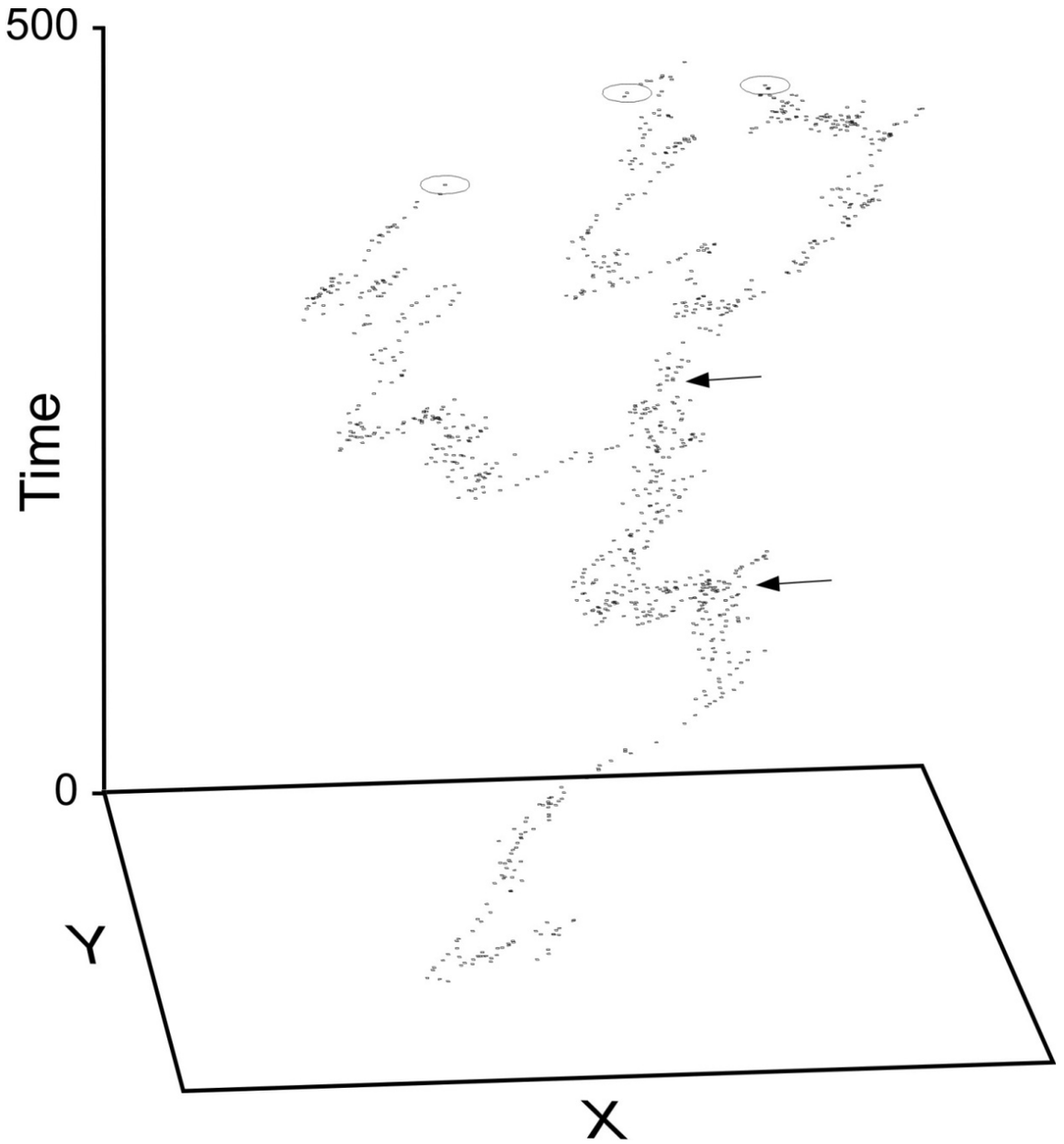


**Felsenstein's (1985) solution**

Method assumes that the evolution of traits is mimicked by a continuous random walk (Brownian motion).

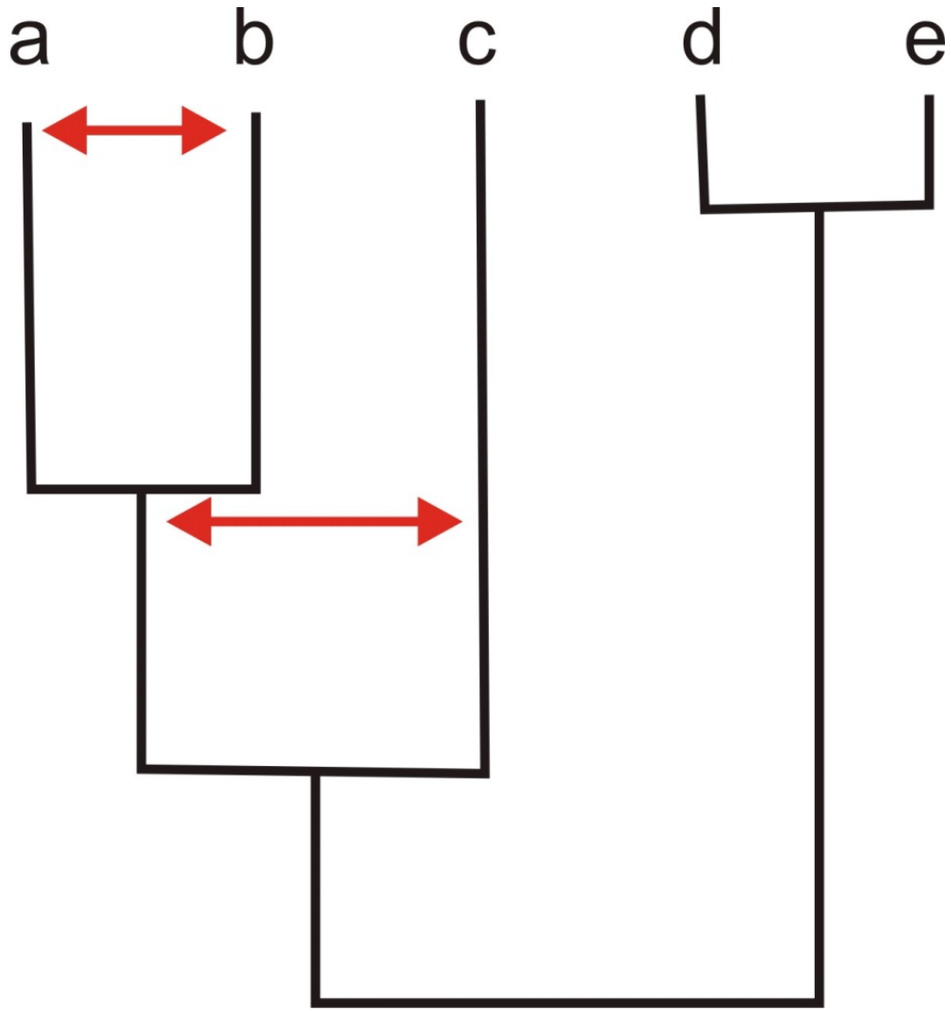
Under Brownian motion, the difference between any two species in a trait has a normal probability distribution with mean 0 and variance proportional to the time since their common ancestor.

Felsenstein (1985) *Am Nat*



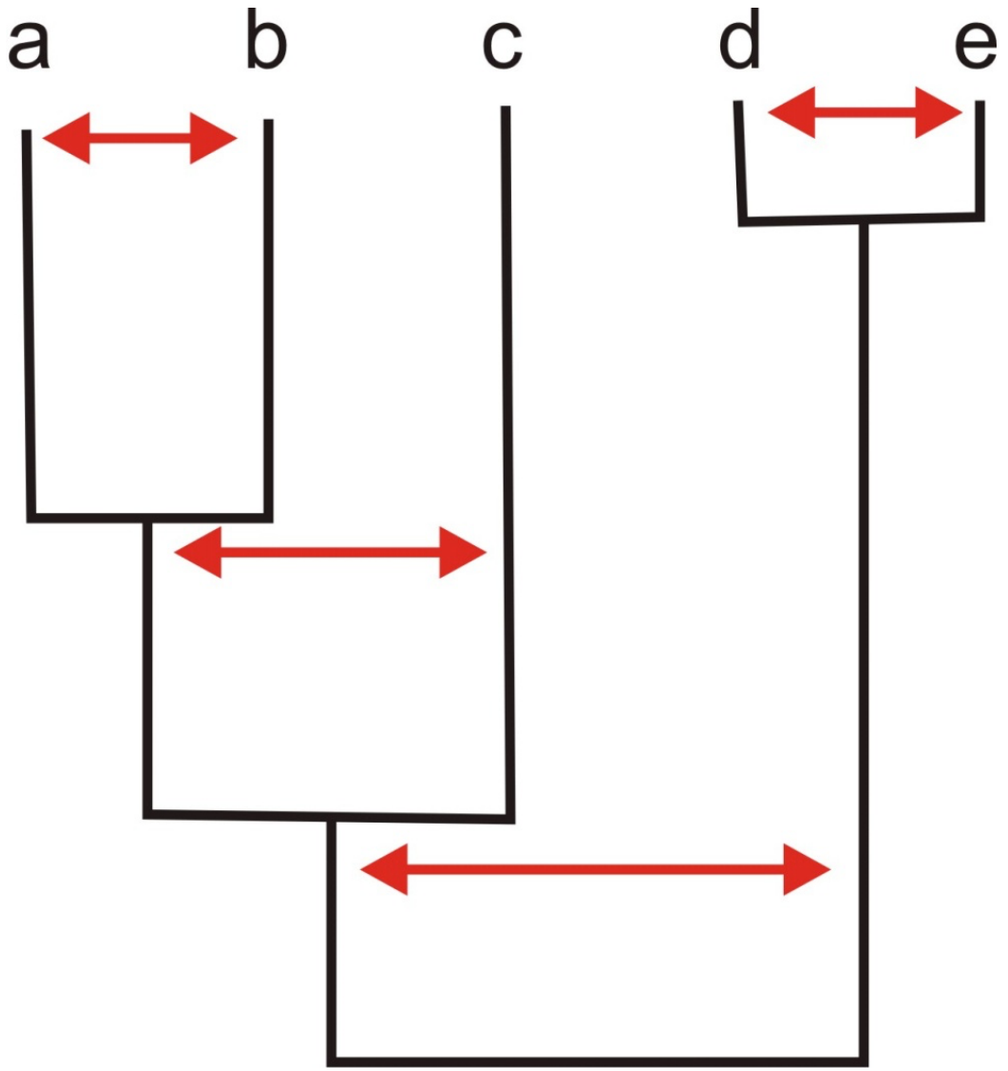
**Felsenstein's method of phylogenetically independent contrasts**

Under Brownian motion, a, b, and c are not independent, but the difference ("contrast") between a and b is independent of the difference between c and (a+b)/2.



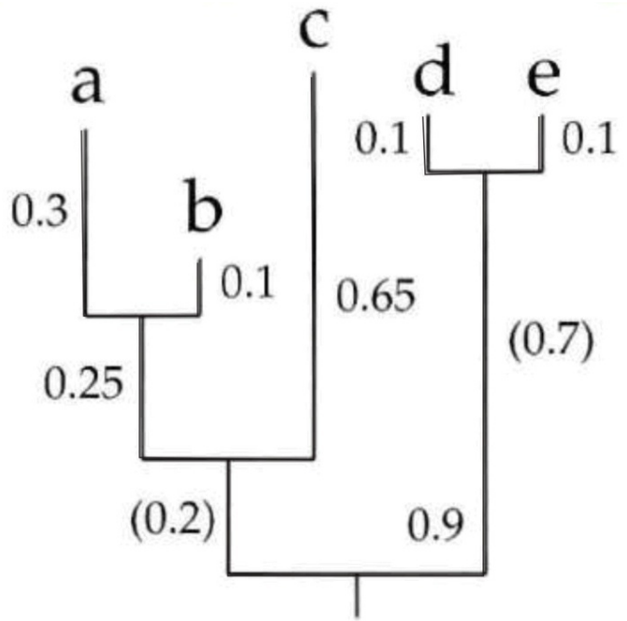
**Phylogenetically independent contrasts**

There are *n* – 1 independent contrasts for *n* species.



**Phylogenetically independent contrasts**

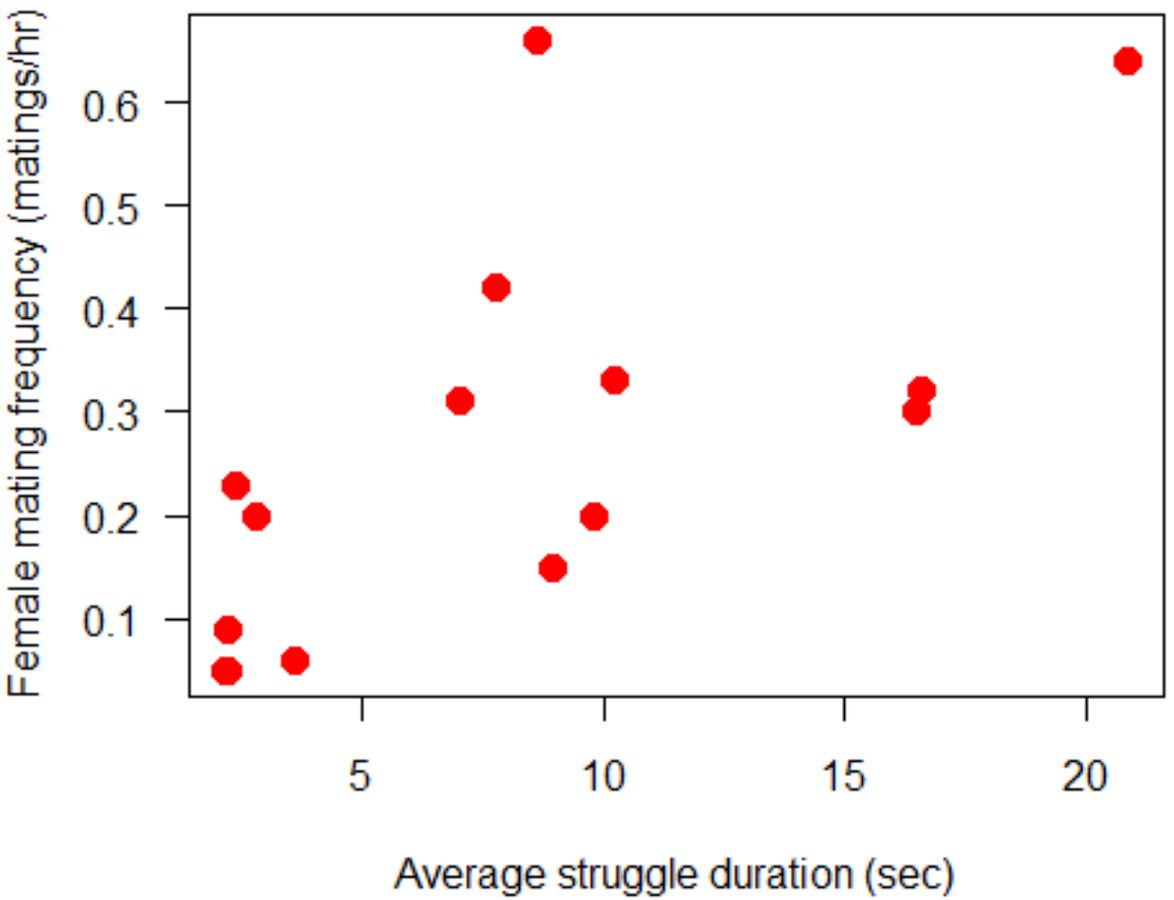
Calculation details. Usually, contrasts are standardized by the square root of the expected variance, which is proportional to branch length.




| Contrast | | | | | | | | | | | Variance proportional to |
|---|---|---|---|---|---|---|---|---|---|---|---|
| $y_1$ | $=$ | $x_a$ | $-$ | $x_b$ | | | | | | | 0.4 |
| $y_2$ | $=$ | $\frac{1}{4}\,x_a$ | $+$ | $\frac{3}{4}\,x_b$ | $-$ | $x_c$ | | | | | 0.975 |
| $y_3$ | $=$ | | | | | | $x_d$ | $-$ | $x_e$ | | 0.2 |
| $y_4$ | $=$ | $\frac{1}{6}\,x_a$ | $+$ | $\frac{1}{2}\,x_b$ | $+$ | $\frac{1}{3}\,x_c$ | $-\frac{1}{2}\,x_d$ | $-$ | $\frac{1}{2}\,x_e$ | | 1.11666 |

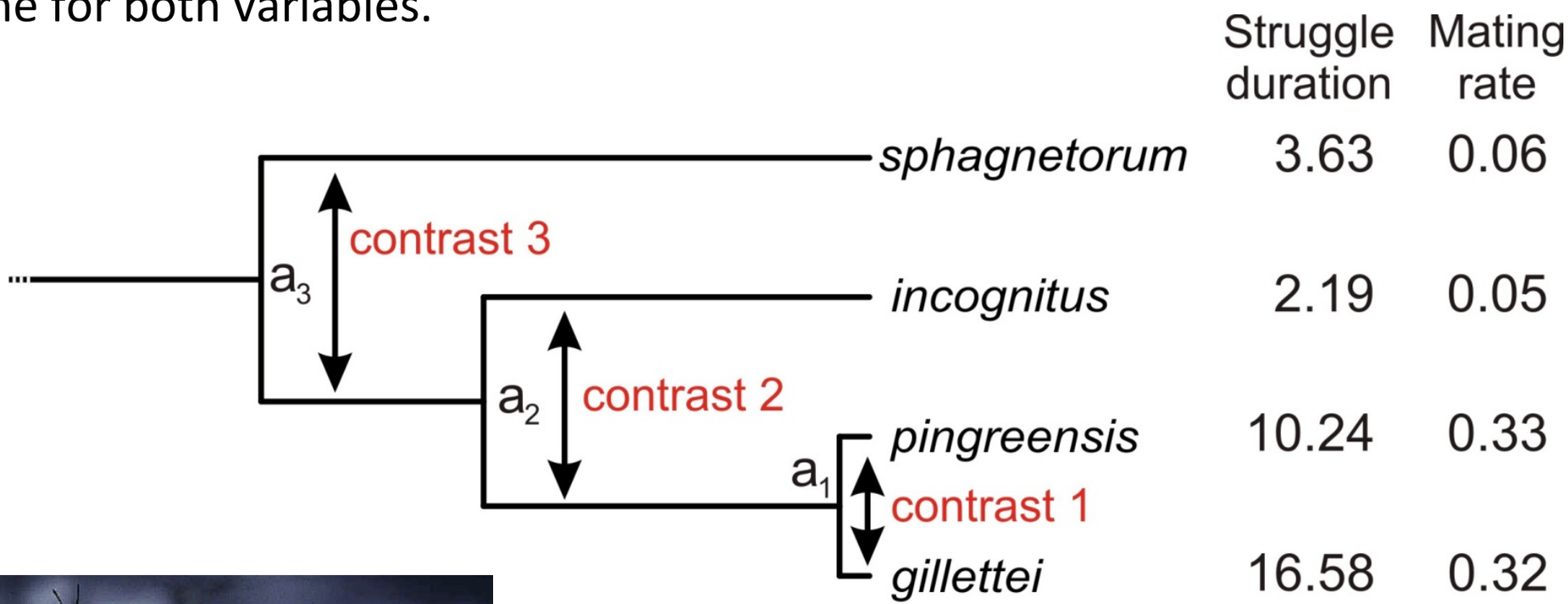
**Phylogenetically independent contrasts**

The idea is to convert the data on both traits to their independent contrasts using the phylogeny of the species. Then calculate the correlation between the independent contrasts of the two traits.



**Phylogenetically independent contrasts**

A cutaway of the independent contrasts for the water strider mating behavior data. The direction of each contrast is arbitrary, but the contrast direction must be the same for both variables.

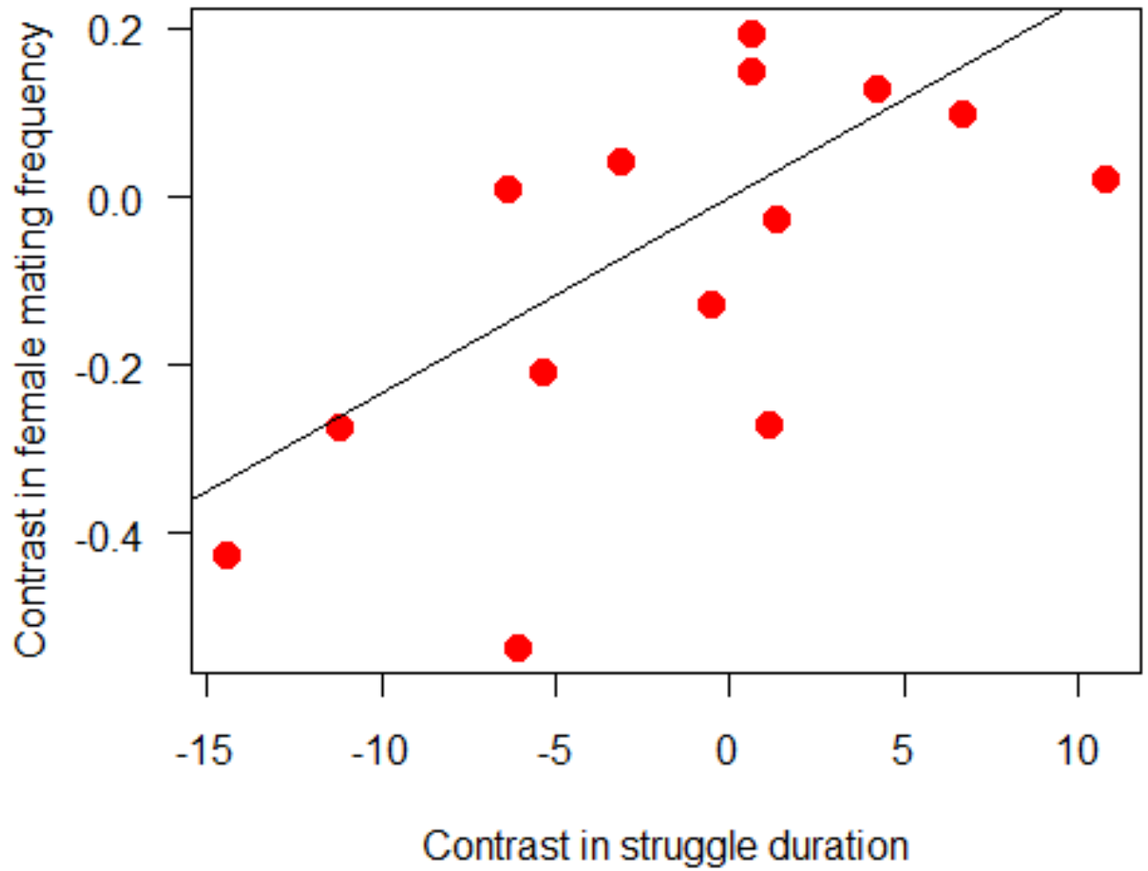



**Phylogenetically independent contrasts**

Because the direction of the contrast is arbitrary, the correlation or regression using independent contrasts is fitted through the origin (0,0).

The ape package in R implements phylogenetically independent contrasts.

Positive correlation confirmed!



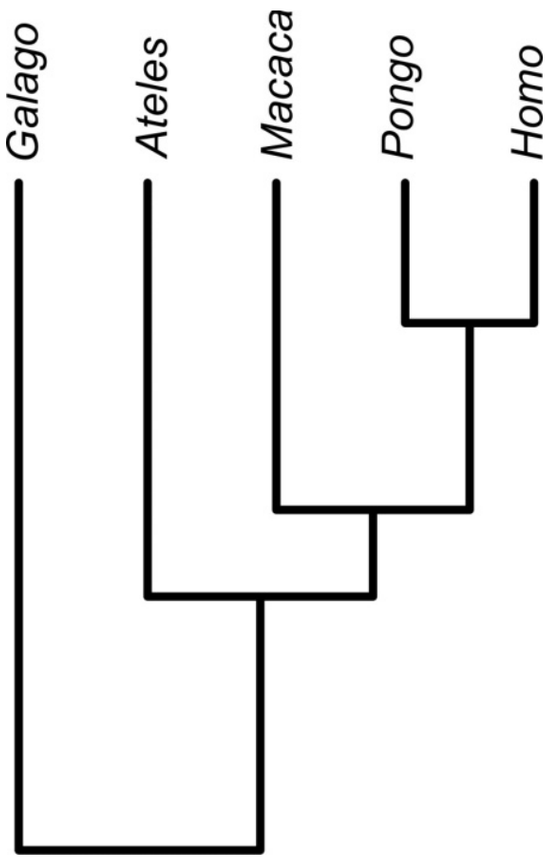
## A linear model approach

General least squares (GLS) is a linear model technique mathematically equivalent to phylogenetically independent contrasts.

GLS allows the residuals to be correlated and have unequal variances. The method incorporates them using a “weight” matrix of expected covariances between species traits.

Using GLS gives access to all the tools of linear models, including model selection methods (AIC, etc).

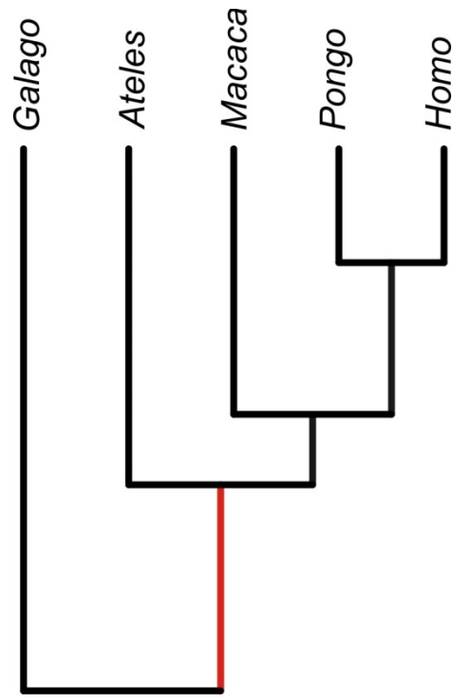
The function `gls()` in the `nlme` package can be used to fit phylogenetic linear models.



## Specifying the covariance matrix between data points

| | *Homo* | *Pongo* | *Macaca* | *Ateles* | *Galago* |
|---|---|---|---|---|---|
| *Homo* | 1.00 | 0.79 | 0.51 | 0.38 | 0 |
| *Pongo* | 0.79 | 1.00 | 0.51 | 0.38 | 0 |
| *Macaca* | 0.51 | 0.51 | 1.00 | 0.38 | 0 |
| *Ateles* | 0.38 | 0.38 | 0.38 | 1.00 | 0 |
| *Galago* | 0.00 | 0.00 | 0.00 | 0.00 | 1 |

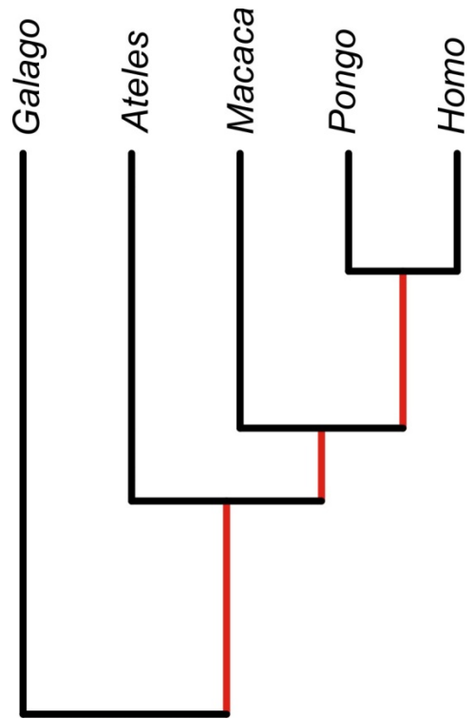
To analyze, we must know what the variances and correlations are between species. Under Brownian motion, the expected covariance between two species is the proportion of total history, from root to tip, that they share.



## Specifying the covariance matrix between data points

| | *Homo* | *Pongo* | *Macaca* | *Ateles* | *Galago* |
|---|---|---|---|---|---|
| *Homo* | 1.00 | 0.79 | 0.51 | 0.38 | 0 |
| *Pongo* | 0.79 | 1.00 | 0.51 | 0.38 | 0 |
| *Macaca* | 0.51 | 0.51 | 1.00 | 0.38 | 0 |
| *Ateles* | 0.38 | 0.38 | 0.38 | 1.00 | 0 |
| *Galago* | 0.00 | 0.00 | 0.00 | 0.00 | 1 |

These expected covariances between pairs of data points (species) are used as "weights" in the linear model fitting. A pair of data points (species) that share most of their phylogenetic history end up being down-weighted in the analysis. In effect, each of them is counted as only a fraction of a data point.



**Assumptions of the method**

- Evolution in each trait mimics a continuous random walk in time (Brownian motion).
- The rate of evolution is constant through time and along all branches of the phylogeny.
- Speciation and extinction are unrelated to trait values.

These assumptions are difficult to verify.

Branch lengths of phylogenies can be transformed to improve agreement with Brownian motion assumption.

If the assumptions are not met, then in extreme cases using independent contrasts might be worse than simply treating the species data as though they were independent (Harvey and Rambaut 2000).

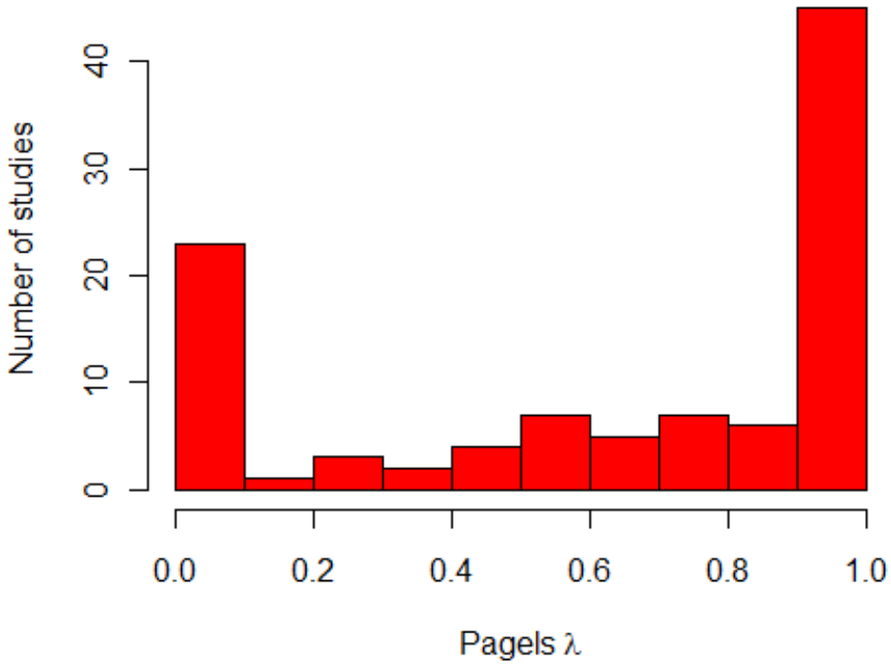
**Assumptions of the method**

The GLS linear model approach makes it easy to transform branch lengths of the tree to better meet the assumption of Brownian motion.

Under Brownian motion, Pagel's phylogenetic signal λ = 1.

If phylogenetic signal λ is less than one, each of the non-diagonal elements of the phylogenetic matrix can be multiplied by the estimated λ. This allows us to fit a model in which phylogenetic signal in the data is weaker than expected under simple Brownian motion.

The ape package in R can finds the "best" estimate of λ for a given data set using maximum likelihood. We'll try this in the workshop



**Discrete species data**

Patterson and Givnish (2002) found that lily species flowering in the low light environment of the forest understory, such as the blue bead lily (*Clintonia borealis*), tend to have small and inconspicuous flowers whitish or greenish in color.

Lilies that live in sunny, open habitats, or that live in deciduous woods but flower before the tree leaves come out, such as the Turk's-cap lily (*Lilium superbum*), tend to have large, showy flowers.

**Discrete species data**

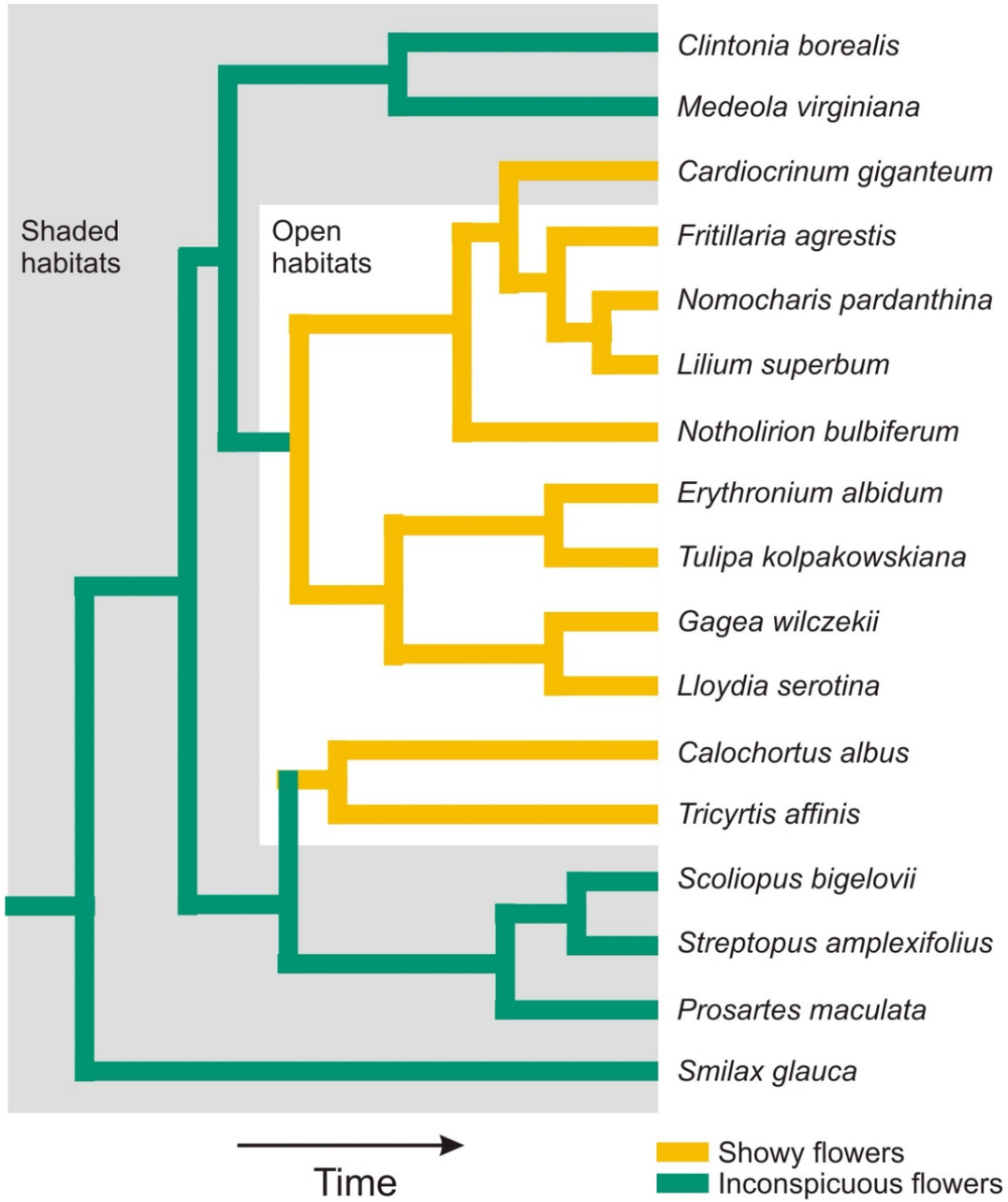
Data from 17 lily species indicated an almost perfect association between habitat and flower type. All ten species flowering in open habitats had large and showy flowers. Six of the seven species flowering in shaded habitats had relatively small and inconspicuous flowers. This seemed like a strong association.

| | **Open habitat** | **Shaded habitat** |
|---|---|---|
| **Showy flowers** | 10 | 0 |
| **Inconspicuous flowers** | 1 | 6 |

**Discrete species data**

But the phylogeny of the group reveals the same problem as in the water strider example: closely related species tend to be similar.

Even though there are 17 species, there might have been as few as three transitions between habitats in the past, leaving fewer effective data points than first assumed.



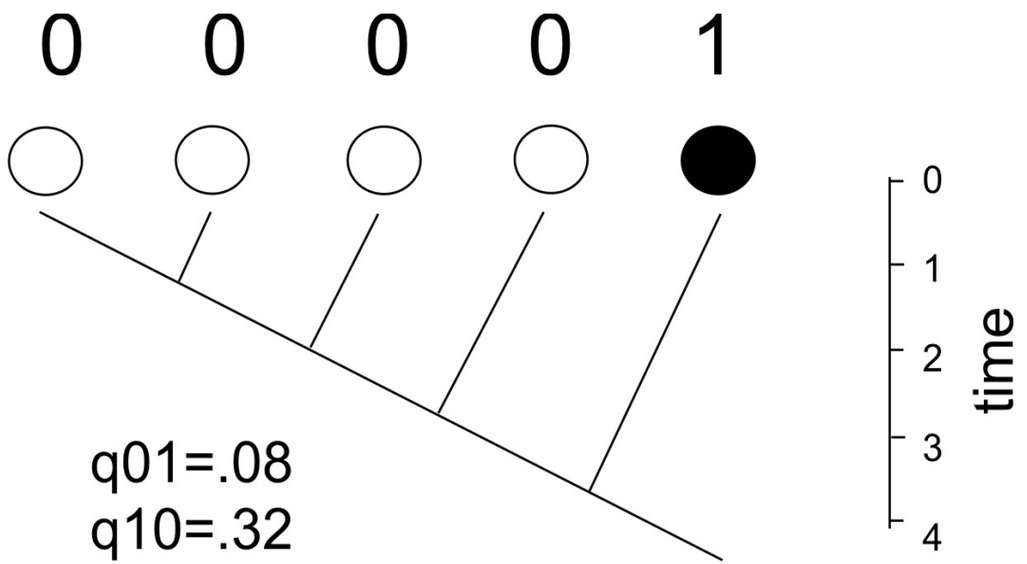
**Discrete species data**

Pagel (1994) developed a maximum likelihood method for analyzing discrete characters. The method assumes that evolution in each trait mimics a discrete random walk in time (Markov process).

It estimates the transition rates *q* between states through time on a phylogeny.

It uses likelihood to estimate and test how transitions between states in one trait (e.g., flower conspicuousness) depend on the character states of a second trait (e.g., habitat).

The method is implemented in the corHMM package in R.



**Discrete species data**

Maddison, W. and R. Fitzjohn. 2015. *The unsolved challenge to phylogenetic correlation tests for categorical characters*. Syst. Biol. 64:127–136.

"*… Pagel's test is susceptible to yielding significant results from the effects of a single change in one of the characters, …. Other tests suffer the same problem, which we will call "within-clade pseudoreplication*".

Possible solution:

Beaulieu, J. M., & B. C. O'Meara. 2014. *Hidden Markov models for studying the evolution of binary morphological characters*. In Modern phylogenetic comparative methods and their application in evolutionary biology (pp. 395-408). Springer, Berlin, Heidelberg.

**Is phylogenetically independent contrasts/GLS also susceptible?**

Uyeda, J. C., R. Zenil-Ferguson, and M. W. Pennell. 2018. *Rethinking phylogenetic comparative methods*. Syst. Biol 67: 1091-1109.

*“...phylogenetically independent contrasts can be misled by a single extraordinary event...”*

Method development continues apace.

**Workshop on phylogenetic comparative methods**

This Thursday!

# R: an embarrassment of riches

**CRAN Task View: Analysis of Spatial Data**

**Maintainer:** Roger Bivand, Jakub Nowosad
**Contact:** Roger.Bivand at nhh.no
**Version:** 2022-04-06
**URL:** https://CRAN.R-project.org/view=Spatial
**Source:** https://github.com/cran-task-views/Spatial/
**Contributions:** Suggestions and improvements for this task view are very welcome and can be made through issues or pull requests on GitHub or via e-mail to the maintainer address. For further details see the Contributing guide.
**Citation:** Roger Bivand, Jakub Nowosad (2022). CRAN Task View: Analysis of Spatial Data. Version 2022-04-06. URL https://CRAN.R-project.org/view=Spatial.
**Installation:** The packages from this task view can be installed automatically using the ctv package. For example, `ctv::install.views("Spatial", coreOnly = TRUE)` installs all the core packages or `ctv::update.views("Spatial")` installs all packages that are not yet installed and up-to-date. See the CRAN Task View Initiative for more details.

Base R includes many functions that can be used for reading, visualising, and analysing spatial data. The focus in this view is on "geographical" spatial data, where observations can be identified with geographical locations, and where additional information about these locations may be retrieved if the location is recorded with care.
Base R functions are complemented by contributed packages provided as source packages, and as ready-to-run binary packages for Windows and macOS (Intel 64-bit and Apple Silicon arm64 architectures). Information about source installs of packages using software external to R may be found at the end of this page. This task view covers the current status of contributed packages available from CRAN.

The contributed packages address two broad areas: moving spatial data into and out of R including coordinate transformation, and analysing spatial data in R. Because the contributed packages constitute an evolving ecosystem, there are several points of entry for users looking for help and information. Two informal organisations curate websites: r-spatial with a hyphen, and rspatial without. R-spatial is more generally geo-informatics based, grew from the legacy sp package and now clearly aligned with the modern sf and stars packages. Rspatial has grown from the raster package, now moving towards the modern terra package. It is also worth noting the wealth of online book projects, which may be helpful for users seeking an introduction, including Geocomputation with R.
Specific questions or issues may be raised where `packageDescription(<pkg>)$BugReports` returns an URL for bug reports or issues (where `<pkg>` is the name of the package as a string), or directly with package maintainers by email. Use may also be made of the R-SIG-Geo mailing-list after subscription, or of stackoverflow with appropriate tags, or of stackexchange. Using the `#rspatial` tag on twitter may also be worth trying, or browsing traffic using that tag (among others).
The packages in this view can be roughly structured into the following topics. If you think that some package is missing from the list, please e-mail the maintainer or submit an issue or pull request in the GitHub repository linked above.

## Classes for spatial data and metadata

Many of the packages for handling and analysing spatial data use shared classes to reduce duplication of effort. Up until 2016, the sp package provided shared classes for spatial vector and raster data, but the representations used preceded more modern and efficient international standards for spatial vector data. From the release of sf, these modern vector representations are to be preferred. For spatial raster data, the representations proposed in stars and terra suit overlapping but slightly different requirements. Conversion between objects of classes defined by sf, stars, terra and the legacy sp packages are available, and are described in Conversions between different spatial classes in R.
Complementary initiatives are ongoing to support better handling of geographic metadata in R.

### Spatial data - general

- sf is a CRAN package for spatial vector data, and is being actively developed here: sf, providing Simple Features for R, in compliance with the OGC Simple Feature standard. The development of the package was supported by the R Consortium.

It provides simple features access for vector data, and as such is a modern implementation and standardization of parts of the legacy sp package. sf is documented in an R Journal article. sf uses the PROJ, GEOS and GDAL external software libraries, which must be available for source installs together with other external software libraries that they in turn depend on.

- stars is being actively developed here: stars, and was supported by the R Consortium; it provides for spatiotemporal data in the form of dense arrays. It supercedes the spacetime package, which extended the shared classes defined in sp for spatio-temporal data (see Spatio-Temporal Data in R ). stars uses PROJ and GDAL through sf.
- The vapour package offers low-level access to GDAL functionality for R packages.
- The spatstat contains classes suited to the analysis of point patterns, and may be coerced to and from `"sf"`, `"stars"` and other spatial classes.
- The rcosmo package provides simple access to spherical and HEALPix data. It extends standard dataframes for HEALPix-type data.
- inlmisc has followed on from Grid2Polygons and converts a spatial object from class SpatialGridDataFrame to SpatialPolygonsDataFrame among many other possibilities for legacy sp classes.

### Raster data

- terra is a re-implementation of raster functionality, linking directly to PROJ, GDAL and GEOS, and introducing new S4 classes for raster and vector data. See the manual and tutorials to get started. terra is very similar to the raster package; but terra is simpler, better, and faster.
- stars provides for spatiotemporal data in the form of dense arrays, with space and time being array dimensions. Examples include socio-economic or demographic data, environmental variables monitored at fixed stations, time series of satellite images with multiple spectral bands, spatial simulations, and climate model results.
- The gdalcubes package also provides classes for data cubes, including proxy data cubes, it links to PROJ, GDAL and NetCDF4.

### Geographic metadata

- geometa provides classes and methods to write geographic metadata following the ISO and OGC metadata standards (ISO 19115, 19110, 19119) and export it as XML (ISO 19139) for later publication into metadata catalogues. Reverserly, geometa provides a way to read ISO 19139 metadata into R. The package extends sf to provide GML (ISO 19136) representation of geometries. geometa is under active development on Github: geometa.
- ncdf4 provides read and write functions for handling metadata (CF conventions) in the self-described NetCDF format.

## Reading and writing spatial data

Spatial data is most often represented by one of two data models, vector or raster, and both models have many of their own file formats. GDAL (Geospatial Data Abstraction Library) is a (non-R) library that provides a unified way to read and write hundreds of spatial data formats. Formats supported by GDAL include both OGC standard data formats (e.g., GeoPackage) and proprietary formats (e.g., ESRI Shapefile). GDAL is used by a large number of GIS software and also many R packages, such as sf, terra, and vapour. This allows us to read and write spatial data in R from and to various spatial file formats. Important note: CRAN offers binary versions of packages sf, terra, and vapour for Windows and macOS, that contain specific GDAL version with a subset of possible data source drivers. If other drivers are needed, you need to either use other conversion utilities or install these packages from the source against a version of GDAL with the required drivers.
In the past, rgdal and raster (through rgdal) were recommended for reading and writing of spatial data in R. However, due to the retirement of [rgdal](../packages/rgdal/index.html) by the end of 2023 new projects should not use it, and existing projects should implement migration to the packages mentioned in the previous paragraph.

### Reading and writing spatial data - data formats

Other packages provide facilities to read and write spatial data, dealing with open standard formats or proprietary formats.
*Open formats*

- *Well-Known Text (WKT) / Well-Known Binary (WKB):* These standards are part of the OGC Simple Feature specification. Both WKT/WKB formats are supported by the sf package that implements the whole OGC Simple Feature specification in R. Additionally, wk and wkutils may be used to parse well-known binary and well-known text representation of geometries to and from R-native formats.
- *GeoJSON:* An rOpenSci blog entry describes a GeoJSON-centred approach to reading GeoJSON and WKT data. The entry lists geojson, and geojsonio, among others. The GeoJSON format can also be read and write with sf, terra, and vapour. wellknown makes possible conversions from WKT to GeoJSON and GeoJSON to WKT.
- *Geographic Markup Language (GML):* GML format can be read and written with sf. Additional GML native reader and writer is provided by geometa model with bindings to the sf classes, for extension of geographic metadata with GML data and metadata elements (GML 3.2.1 and 3.3) and interfacing OGC web-services in ows4R package.
- *NetCDF files:* NetCDF files can be read and write with ncdf4 or RNetCDF. Additionally, both terra and stars have capabilities for reading and writing NetCDF files.
- *LAS / LAX:* These file formats are designed to work with lidar point cloud data and can be read/write with lidR or rLiDAR.

*Proprietary Data Formats*

- *ESRI formats:* Many of spatial data saved into ESRI file formats can be read with GDAL, and thus also with sf, terra, and vapour. Additionally, shapefiles reads and writes ESRI ArcGIS/ArcView shapefiles.
  Additionally, maps (with mapdata and mapproj) provides a legacy tool to access to the same kinds of geographical databases as S.
- *Others:* The gmt package gives a simple interface between GMT map-making software and R.

## Reading and writing spatial data - GIS Software connectors

- *PostGIS:* The rpostgis package provides additional functions to the RPostgreSQL package to interface R with a 'PostGIS'-enabled database, as well as convenient wrappers to common 'PostgreSQL' queries. It is documented in an R

Journal article. postGIStools package provides functions to convert geometry and 'hstore' data types from 'PostgreSQL' into standard R objects, as well as to simplify the import of R data frames (including spatial data frames) into 'PostgreSQL'. sf also provides an R interface to PostGIS, for both reading and writing, through GDAL.

- *GRASS GIS:* Integration with version 7.* of the leading open source GIS, GRASS GIS, is provided in CRAN package rgrass7. For GRASS 6.*, use spgrass6.
- *SAGA GIS:* RSAGA and Rsagacmd offer shell-based wrapper for SAGA GIS commands.
- *Quantum GIS (QGIS):* QGIS2 was supported by RQGIS (RQGIS). QGIS3 (version >= 3.16) is supported by qgisprocess, which establishes an interface between R and QGIS, i.e., it allows the user to access QGIS functionalities from the R console. It achieves this by using the qgis_process command-line utility.
- *WhiteboxTools:* whitebox is an R frontend for the WhiteboxTools software.
- *ArcGIS:* RPyGeo is a wrapper for Python access to the ArcGIS GeoProcessor. The ESRI company also offers their own package (r-bridge) that allows transferring data from ArcGIS to R.
- Various GIS Software, including Orfeo ToolBox and SAGA GIS, can also be connected to R using link2GI.

## Interfaces to Spatial Web-Services

Some R packages focused on providing interfaces to web-services and web tools in support of spatial data management. Here follows a first tentative (non-exhaustive) list:

- ows4R is a package that intends to provide an R interface to OGC standard Web-Services. It is in active development at ows4R and currently support interfaces to the Web Feature Service (WFS) for vector data access, with binding to the sf package, and the Catalogue Service (CSW) for geographic metadata discovery and management (including transactions), with binding to the geometa package.
- geosapi is an R client for the GeoServer REST API, an open source implementation used widely for serving spatial data.
- geonapi provides an interface to the GeoNetwork legacy API, an open source catalogue for managing geographic metadata.

- rgee is an Earth Engine client library for R. All of the 'Earth Engine' API classes, modules, and functions are made available. Additional functions implemented include importing (exporting) of Earth Engine spatial objects, extraction of time series, interactive map display, assets management interface, and metadata display.

## Specific geospatial data sources of interest

- rnaturalearth package facilitates interaction with Natural Earth map data. It includes functions to download a wealth of Natural Earth vector and raster data, including cultural (e.g., country boundaries, airports, roads, railroads) and physical (e.g., coastline, lakes, glaciated areas) datasets.
- Historical country boundaries (1886-today) can be obtained from the cshapes.
- marmap package is designed for downloading, plotting, and manipulating bathymetric and topographic data in R. It allows to query the ETOPO1 bathymetry and topography database hosted by the NOAA, use simple latitude-longitude-depth data in ascii format, and take advantage of the advanced plotting tools available in R to build publication-quality bathymetric maps (see the PLOS paper).
- tidycensus provides access to US Census Bureau data in a tidy format, including the option to bind the data spatially on import.
- tigris provides access to cartographic elements provided by the US Census Bureau TIGER, including cartographic boundaries, roads, and water.
- rgbif package is used to access Global Biodiversity Information Facility (GBIF) occurrence data
- geonames is an interface to the www.geonames.org service.
- osmdata is an R package for accessing relatively small datasets from OpenStreetMap (OSM), delivered via the Overpass API. osmextract matches, downloads, converts, and reads OpenStreetMap data covering large areas, obtained from Geofabrik and other providers.
- OpenStreetMap gives access to open street map raster images, and osmar provides an infrastructure to access OpenStreetMap data from different sources, to work with the data in a common R manner, and to convert data into available infrastructure provided by existing R packages.

- giscoR provides access to spatial elements provided by GISCO - Eurostat, including boundary files of countries, NUTS regions, municipalities, and other spatial objects.
- chilemapas provides access to spatial data of political and administrative divisions of Chile.
- The former `geobr` package provided easy access to official spatial data sets of Brazil for multiple geographies and years.
- geouy loads and process geographic information for Uruguay.
- RCzechia downloads spatial boundary files of administrative regions and other spatial objects of the Czech Republic.
- rgugik allows to search and retrieve data from Polish Head Office of Geodesy and Cartography ("GUGiK").
- mapSpain downloads spatial boundary files of administrative regions and other spatial objects of Spain.

## Handling spatial data

### Data processing - general

- sf provides an interface to spatial geometry functions using the GEOS and S2 libraries.
- stars contains tools for manipulating raster and vector data cubes.
- terra package introduces many GIS methods for spatial vector and raster data.
- The gdalUtils and gdalUtilities packages provide wrappers for the Geospatial Data Abstraction Library (GDAL) Utilities.
- rmapshaper is a wrapper around the 'mapshaper' 'JavaScript' library to perform topologically-aware polygon simplification and other operations such as clipping, erasing, dissolving, and converting 'multi-part' to 'single-part' geometries.
- gdistance, provides functions to calculate distances and routes on geographic grids. geosphere permits computations of distance and area to be carried out on spatial data in geographical coordinates. cshapes package provides functions for calculating distance matrices (see Mapping and Measuring Country Shapes).
- magclass offers a data class for increased interoperability working with spatial-temporal data together with corresponding functions and methods (conversions, basic calculations and basic data manipulation).
- The rcosmo package offers various tools for geometric transformations, computations, and statistical analysis of spherical data.
- The trip package extends spatial classes to permit the accessing and manipulating of spatial data for animal tracking.

### Data cleaning

- sf has a built-in functions st_is_valid to check whether an sf geometry is valid and st_make_valid to fix invalid geometry (from GEOS 3.8).
- lwgeom may also be used to facilitate handling and reporting of topology errors and geometry validity issues in sf objects.

### Data processing - specific

- The landsat package with accompanying JSS paper provides tools for exploring and developing correction tools for remote sensing data.
- The areal package can be used to interpolate overlapping but incongruent polygons, also known as areal weighted interpolation.
- The qualmap package can be used to digitize qualitative GIS data.

### Spatial sampling

- spsurvey provides a range of sampling functions.
- Spbsampling allows selecting probability samples well spread over the population of interest, in any dimension and using any distance function.
- spatialsample is a member of the tidymodel family of packages and contains functions and classes for spatial resampling to use with the rsample.
- MBHdesign provides spatially survey balanced designs using the quasi-random number method.
- SpotSampling contains three methods for spatial and temporal sampling.

## Visualizing spatial data

### Base visualization packages

- Packages such as sf, stars, terra, and rasterVis provide basic visualization methods through the generic plot function.

- classInt package provides functions for choosing class intervals for thematic cartography.
- rcosmo package provides several tools to interactively visualize HEALPix data, in particular, to plot data in arbitrary spherical windows.
- Currently, the grDevices package (included with the R installation) contains a large number of color palettes that can be accessed with the hcl.colors and palette.colors functions; see also New features in this blog. Some of these color palettes can be also retrieved using separate packages, such as RColorBrewer, viridis, or rcartocolor.

### Thematic cartography packages

- tmap package accepts most spatial data classes and provides a modern basis for thematic mapping using a Grammar of Graphics syntax. It also allows for interactive spatial data mapping.
- mapsf package allows various cartographic representations such as proportional symbols, choropleth, or typology maps; it accepts sf (sf) and SpatRaster (terra) objects
- ggplot2 package has a built-in support for sf objects with the geom_sf function and additional support for stars object is available through the geom_stars function available in the stars package. Its spatial visualization capabilities can be further extended with ggspatial, which adds support for more spatial classes (including classes from the raster package), allows adding north arrows and scale bars, etc.
- The mapmisc package is a minimal, light-weight set of tools for producing nice-looking maps in R, with support for map projections.
- Additional processing and mapping functions are available in PBSmapping package; PBSmodelling provides modelling support. In addition, GEOmap provides mapping facilities directed to meet the needs of geologists and uses the geomapdata package.

### Packages based on web-mapping frameworks

- mapview and leaflet packages provide methods to view spatial objects interactively, usually on a web mapping base. Additionally, tmap has a view mode that allows for interactive spatial data mapping.

- mapdeck package provides a mechanism to plot interactive maps through javascript libraries ‘Mapbox GL’ and ‘Deck.gl’.
- RgoogleMaps package for accessing Google Maps(TM) may be useful if the user wishes to place a map backdrop behind other displays.
- ggmap may be used for spatial visualization with Google Maps and OpenStreetMap; ggsn provides north arrows and scales for such maps.
- mapedit provides an R shiny widget based on leaflet for editing or creating sf geometries.

### Building Cartograms

- cartogram package allows for constructions of a continuous area cartogram by a rubber sheet distortion algorithm, non-contiguous area cartograms, and non-overlapping circles cartogram.
- geogrid package turns polygons into rectangular or hexagonal cartograms.
- micromap package provides linked micromaps using ggplot2.
- recmap package provides rectangular cartograms with rectangle sizes reflecting for example population.
- geogrid turns spatial polygons into regular or hexagonal grids. statebins provides a simple binning approach to US states.

## Analyzing spatial data

The division of spatial statistics into three partly overlapping areas: point pattern analysis, geostatistics and the analysis of areal/lattice data, is widely accepted. However, areal data analysis can be split into disease mapping and spatial regression (also partly overlapping). In addition, ecological analyses often approach spatial data in particular ways, giving rise to a specific topical cluster of packages. All of these approaches to analysing spatial data treat the spatial relationships between observations as a way of exploring and making use of important sources of information about the observations over and above what is known when assuming that they are independent of each other.

### Point pattern analysis

Point pattern analysis examines the distance relationships between observed points, where the set of observations is expected to encompass all such entities in the study area.

- spatstat is a family of R packages for analysing spatial point pattern data (and other kinds of spatial data). It has extensive capabilities for exploratory analysis, statistical modelling, simulation and statistical inference. It allows freedom in defining the region(s) of interest, and makes extensions to marked processes and spatial covariates. Its strengths are model-fitting and simulation, and it has a useful homepage; it is actively developed. It is the only package that will enable the user to fit inhomogeneous point process models with interpoint interactions.
- The splancs package allows point data to be analysed within a polygonal region of interest, and covers many methods, including 2D kernel densities.
- The spatial package is a recommended package shipped with base R, and contains several core functions, including an implementation of Khat by its author, Prof. Ripley.
- The spatgraphs package provides graphs, graph visualisation and graph based summaries to be used with spatial point pattern analysis.
- The smacpod package provides various statistical methods for analyzing case-control point data. The methods available closely follow those in chapter 6 of Applied Spatial Statistics for Public Health Data by Waller and Gotway (2004).
- ecespa provides wrappers, functions and data for spatial point pattern analysis, used in the book on Spatial Ecology of the ECESPA/AEET. The functions for binning points on grids in
- ads may also be of interest. The ads package performs first- and second-order multi-scale analyses derived from Ripley's K-function.
- The dbmss package allows simple computation of a full set of spatial statistic functions of distance, including classical ones (Ripley's K and others) and more recent ones used by spatial economists (Duranton and Overman's Kd, Marcon and Puech's M). It relies on spatstat for core calculation.

### Geostatistics

Geostatistics uses a model fitted using the distances between observations to interpolate values observed at point to unobserved points

- The gstat package provides a wide range of functions for univariate and multivariate geostatistics, also for larger datasets.
- geoR contains functions for model-based geostatistics.

- Variogram diagnostics may be carried out with vardiag.
- Automated interpolation using gstat is available in automap.
- This family of packages is supplemented by intamap with procedures for automated interpolation.
- A similar wide range of functions is to be found in the fields package, extended by LatticeKrig for large spatial datasets and autoFRK.
- The spatial package is shipped with base R, and contains several core geostatistical functions.
- The spBayes package fits Gaussian univariate and multivariate models with MCMC.
- ramps is a different Bayesian geostatistical modelling package.
- The geospt package contains some geostatistical and radial basis functions, including prediction and cross validation. Besides, it includes functions for the design of optimal spatial sampling networks based on geostatistical modelling.
- The rcosmo package offers various geostatistics methods for spherical data: descriptive statistics, entropy based methods, covariance-variogram methods, etc. Most of rcosmo features were developed for Cosmic Microwave Background data, but they can also be used for any spherical data.
- The FRK package is a tool for spatial/spatio-temporal modelling and prediction with large datasets. The approach, discussed in Cressie and Johannesson (2008), decomposes the field, and hence the covariance function, using a fixed set of n basis functions, where n is typically much smaller than the number of data points (or polygons) m.
- The RandomFields package provides functions for the simulation and analysis of random fields, and variogram model descriptions can be passed between geoR, gstat and this package.
- SpatialExtremes proposes several approaches for spatial extremes modelling using RandomFields.
- In addition, CompRandFld, constrainedKriging and geospt provide alternative approaches to geostatistical modelling.
- The spTimer package is able to fit, spatially predict and temporally forecast large amounts of space-time data using [1] Bayesian Gaussian Process (GP) Models, [2] Bayesian Auto-Regressive (AR) Models, and [3] Bayesian Gaussian Predictive Processes (GPP) based AR Models.
- The rtop package provides functions for the geostatistical interpolation of data with irregular spatial support such as runoff related data or data from administrative units.

- The georob package provides functions for fitting linear models with spatially correlated errors by robust and Gaussian Restricted Maximum Likelihood and for computing robust and customary point and block kriging predictions, along with utility functions for cross-validation and for unbiased back-transformation of kriging predictions of log-transformed data.
- The SpatialTools package has an emphasis on kriging, and provides functions for prediction and simulation. It is extended by ExceedanceTools, which provides tools for constructing confidence regions for exceedance regions and contour lines.
- The gear package implements common geostatistical methods in a clean, straightforward, efficient manner, and is said to be a quasi reboot of SpatialTools.
- The sperrorest package implements spatial error estimation and permutation-based spatial variable importance using different spatial cross-validation and spatial block bootstrap methods, used by mlr3spatiotempcv.
- The sgeostat package is also available. Within the same general topical area are the deldir package for triangulation and the interp package for spline interpolation; the MBA package provides scattered data interpolation with multilevel B-splines.
- In addition, there are the spatialCovariance package, which supports the computation of spatial covariance matrices for data on rectangles, the regress package building in part on spatialCovariance, and the tgp package.
- The Stem package provides for the estimation of the parameters of a spatio-temporal model using the EM algorithm, and the estimation of the parameter standard errors using a spatio-temporal parametric bootstrap.
- FieldSim is another random fields simulations package.
- The SSN is for geostatistical modeling for data on stream networks, including models based on in-stream distance. Models are created using moving average constructions. Spatial linear models, including covariates, can be fit with ML or REML. Mapping and other graphical functions are included.
- The ipdw provides functions to interpolate georeferenced point data via Inverse Path Distance Weighting. Useful for coastal marine applications where barriers in the landscape preclude interpolation with Euclidean distances.
- RSurvey may be used as a processing program for spatially distributed data, and is capable of error corrections and data visualisation.

## Disease mapping and areal data analysis

Both point pattern analysis and geostatistics enter into disease mapping, which is concerned with representing public health information over space and time in a communicative and responsible way. Estimation is important to present calculated rates that are comparable both in terms of levels and uncertainty.

- DCluster is a package for the detection of spatial clusters of diseases. It is complemented by DClusterm for model-based cluster detection, and by rflexscan and FlexScan, two implementations of flexible scan statistics.
- DCluster extends and depends on the spdep package, which provides basic functions for building neighbour lists and spatial weights.
- spdep also provides global and local tests for spatial autocorrelation, including join-count tests, Moran's I, Geary's C, Getis-Ord G and others.
- rgeoda is a wrapper for GeoDa and provides efficient alternatives for calculating global and local tests for spatial autocorrelation.
- Some functions for fitting spatial regression models, such as SAR and CAR models are in spatialreg, see below.
- The SpatialEpi package provides implementations of cluster detection and disease mapping functions, including Bayesian cluster detection, and supports strata.
- The smerc package provides statistical methods for the analysis of data areal data, with a focus on cluster detection.
- The diseasemapping *(archived)* package offers the formatting of population and case data, calculation of Standardized Incidence Ratios, and fitting the BYM model using INLA.
- A Markov Random Field `"mrf"` effect may be added to models in the mgcv package shipped with base R, providing flexible modelling tools in a recommended package.
- The hglm package also provides SAR and CAR model fitting approaches.
- Regionalization of polygon objects is provided by AMOEBA: a function to calculate spatial clusters using the Getis-Ord local statistic. It searches for irregular clusters (ecotopes) on a map, as does `skater()` in spdep.

- The seg, divseg and OasisR packages provide functions for measuring spatial segregation; OasisR includes Monte Carlo simulations to test the indices.
- The lctools package provides researchers and educators with easy-to-learn user friendly tools for calculating key spatial statistics and to apply simple as well as advanced methods of spatial analysis in real data. These include: Local Pearson and Geographically Weighted Pearson Correlation Coefficients, Spatial Inequality Measures (Gini, Spatial Gini, LQ, Focal LQ), Spatial Autocorrelation (Global and Local Moran's I), several Geographically Weighted Regression techniques and other Spatial Analysis tools (other geographically weighted statistics). This package also contains functions for measuring the significance of each statistic calculated, mainly based on Monte Carlo simulations. The
- sparr package provides another approach to relative risks.
- The CARBayes package implements Bayesian hierarchical spatial areal unit models. In such models, the spatial correlation is modelled by a set of random effects, which are assigned a conditional autoregressive (CAR) prior distribution. Examples of the models included are the BYM model as well as a recently developed localised spatial smoothing model.
- The spaMM package fits spatial GLMMs, using the Matern correlation function as the basic model for spatial random effects.
- The PReMiuM package is for profile regression, which is a Dirichlet process Bayesian clustering model; it provides a spatial CAR term that can be included in the fixed effects (which are global, ie. non-cluster specific, parameters) to account for any spatial correlation in the residuals.
- Spatial survival analysis is provided by the spBayesSurv package: Bayesian Modeling and Analysis of Spatially Correlated Survival Data.
- The spselect package provides modelling functions based on forward stepwise regression, incremental forward stagewise regression, least angle regression (LARS), and lasso models for selecting the spatial scale of covariates in regression models.
- Spatial microsimulation is offered by rakeR, sms, synthACS, and NetLogoR permits the building and running of spatially explicit agent-based models.

## Spatial regression

Many packages providing functions for fitting spatial regression models have already been given as they are used in disease mapping. In this subsection, more attention is given to the subset of methods used in spatial econometrics, and so complements general econometric methods covered in the Econometrics Task View.

- The choice of function for spatial regression will depend on the support available. If the data are characterised by point support and the spatial process is continuous, geostatistical methods may be used, or functions in the nlme package.
- If the support is areal, and the spatial process is not being treated as continuous, functions provided in the spatialreg package may be used. This package can also be seen as providing spatial econometrics functions. spdep provides the full range of local indicators of spatial association, such as local Moran's I and diagnostic tools for fitted linear models, including Lagrange Multiplier tests. Spatial regression models that can be fitted using maximum likelihood and Bayesian MCMC methods in spatialreg include spatial lag models, spatial error models, two parameter models, their Durbin variants and SLX models. For larger data sets, sparse matrix techniques can be used for maximum likelihood fits. In spatialreg, the `ME` and `SpatialFiltering` functions provide Moran Eigenvector model fitting, as do more modern functions in the spmoran package.
- When using the generalized method of moments (GMM), sphet can be used to accommodate both autocorrelation and heteroskedasticity, also with instrumental variables.
- The splm package provides methods for fitting spatial panel data by maximum likelihood and GM.
- The spsur package provides functions to test and estimate spatial seemingly unrelated regression models (spatial SUR) by maximum likelihood and three-stage least squares.
- The two small packages S2sls and spanel provide alternative implementations without most of the facilities of splm.
- The former `HSAR` package provides Hierarchical Spatial Autoregressive Models (HSAR), based on a Bayesian Markov Chain Monte Carlo (MCMC) algorithm.
- spatialprobit makes possible Bayesian estimation of the spatial autoregressive probit model (SAR probit model).
- The ProbitSpatial package provides methods for fitting Binomial spatial probit models to larger data sets; spatial autoregressive (SAR) and spatial error (SEM) probit models are included.

- The starma package provides functions to identify, estimate and diagnose a Space-Time AutoRegressive Moving Average (STARMA) model.
- The spgwr package contains an implementation of geographically weighted regression methods for exploring possible spatial non-stationarity. The gwrr package fits geographically weighted regression (GWR) models and has tools to diagnose and remediate collinearity in the GWR models. It also fits geographically weighted ridge regression (GWRR) and geographically weighted lasso (GWL) models. The GWmodel package contains functions for computing geographically weighted (GW) models. Specifically, basic, robust, local ridge, heteroskedastic, mixed, multiscale, generalised and space-time GWR; GW summary statistics, GW PCA and GW discriminant analysis; associated tests and diagnostics; and options for a range of distance metrics.

## Ecological analysis

There are many packages for analysing ecological and environmental data. They include:

- ade4 for exploratory and Euclidean methods in the environmental sciences, the adehabitat family of packages for the analysis of habitat selection by animals (adehabitatHR, adehabitatHS, adehabitatLT, and adehabitatMA)
- pastecs for the regulation, decomposition and analysis of space-time series
- vegan for ordination methods and other useful functions for community and vegetation ecologists, and many other functions in other contributed packages. One such is tripEstimation, basing on the classes provided by trip.
- ncf provides a range of spatial nonparametric covariance functions.
- The spind package provides functions for spatial methods based on generalized estimating equations (GEE) and wavelet-revised methods (WRM), functions for scaling by wavelet multiresolution regression (WMRR), conducting multi-model inference, and stepwise model selection.
- The siplab package is a platform for experimenting with spatially explicit individual-based vegetation models.
- ModelMap builds on other packages to create models using underlying GIS data.
- The SpatialPosition computes spatial position models: Stewart potentials, Reilly catchment areas, Huff catchment areas.
- The Watersheds package provides methods for watersheds aggregation and spatial drainage network analysis.

- The ngspatial package provides tools for analyzing spatial data, especially non-Gaussian areal data. It supports the sparse spatial generalized linear mixed model of Hughes and Haran (2013) and the centered autologistic model of Caragea and Kaiser (2009).
- landscapemetrics package calculates landscape metrics for categorical landscape patterns. It can be used as a drop-in replacement for FRAGSTATS , as it offers a reproducible workflow for landscape analysis in a single environment. It also provides several visualization functions, e.g. to show all labeled patches or the core area of all patches.

The Environmetrics Task View contains a much more complete survey of relevant functions and packages.

## Installing packages linking to PROJ, GDAL or GEOS

Installation of packages like sf and terra which use external software libraries such as PROJ, GDAL or GEOS requires care. For most users on platforms such as Windows or macOS who are not themselves package developers, it is always better to avoid what are known as source installs, because CRAN binary packages include all of the external software required.
Because `getOption("pkgType")` on these platforms is usually `"both"`, you may be asked to choose to install a source package if it is more recent than the latest binary.
Please do not be tempted to choose a source install for sf or terra or similar; the binary package will be generated within a day or two. To avoid being asked, you may see from `?options` under options provided by the utils package that the default behaviour of your installation of R may be controlled by setting
options `install.packages.check.source` and `install.packages.compile.from.source`, or by setting environment variable `R_COMPILE_AND_INSTALL_PACKAGES`, see also this helpful comment.
If you are a developer using Windows or macOS or installing from `github`, the same static-linked binary external software libraries, header files, etc. as those used in building CRAN binary packages are available from: Windows 4.0 and 4.1 downloaded on-the-fly, Windows 4.2 forthcoming rtools42 and macOS both architectures. These external software libraries have been built using the same compile and link settings as R itself, so avoid the risk of possible errors caused by mismatched binaries.

If you are a user (or developer) on systems where `getOption("pkgType")` is `"source"`, you will need to ensure that the external software is available when installing source packages. Advice for some such systems may be found here. The most common reason for failure is having multiple versions of external software installed on your platform.

## CRAN packages

## Related links

- R-SIG-Geo mailing list
- r-spatial
- rspatial
- Geocomputation with R

## Other resources

- CRAN Task View: Econometrics
- CRAN Task View: Environmetrics
- GitHub Project: geometa
- GitHub Project: ows4R
- GitHub Project: qgisprocess
- GitHub Project: r-bridge
- GitHub Project: RQGIS
- GitHub Project: sf
- GitHub Project: stars

**Use R!**

This course was an introduction to more advanced methods in data analysis in ecology and evolution, how they work, and how you can avoid *some* of the most common misinterpretations and perils.

These methods will likely be useful to your future work. Hopefully you have a basis to go further as needed.

The R tips web site and the workshops will remain online and available for the foreseeable future. I'll do my best to keep it up to date. Revisit and refresh your memories as needed.

Lots of people use R for data analysis here, so there is help all around. Start a data analysis group!

Bye!