FISH 604
Module 3: Exploratory data analysis

Instructor: Franz Mueter

Lena Point, Rm 315

796-5448

fmueter@alaska.edu

Objectives & Outcomes

You should appreciate…

- … the importance of visually analyzing your data
- .. the variety of methods available to displaying multi-dimensional data

You should know…

- …how to assess (<u>approximate</u>) normality
- … how to detect outliers and what to do in the presence of outliers

You should be able to …

- …quickly and efficiently explore the main features of simple and complex data sets
- … identify and apply appropriate data transformations as needed

Exploratory data analysis

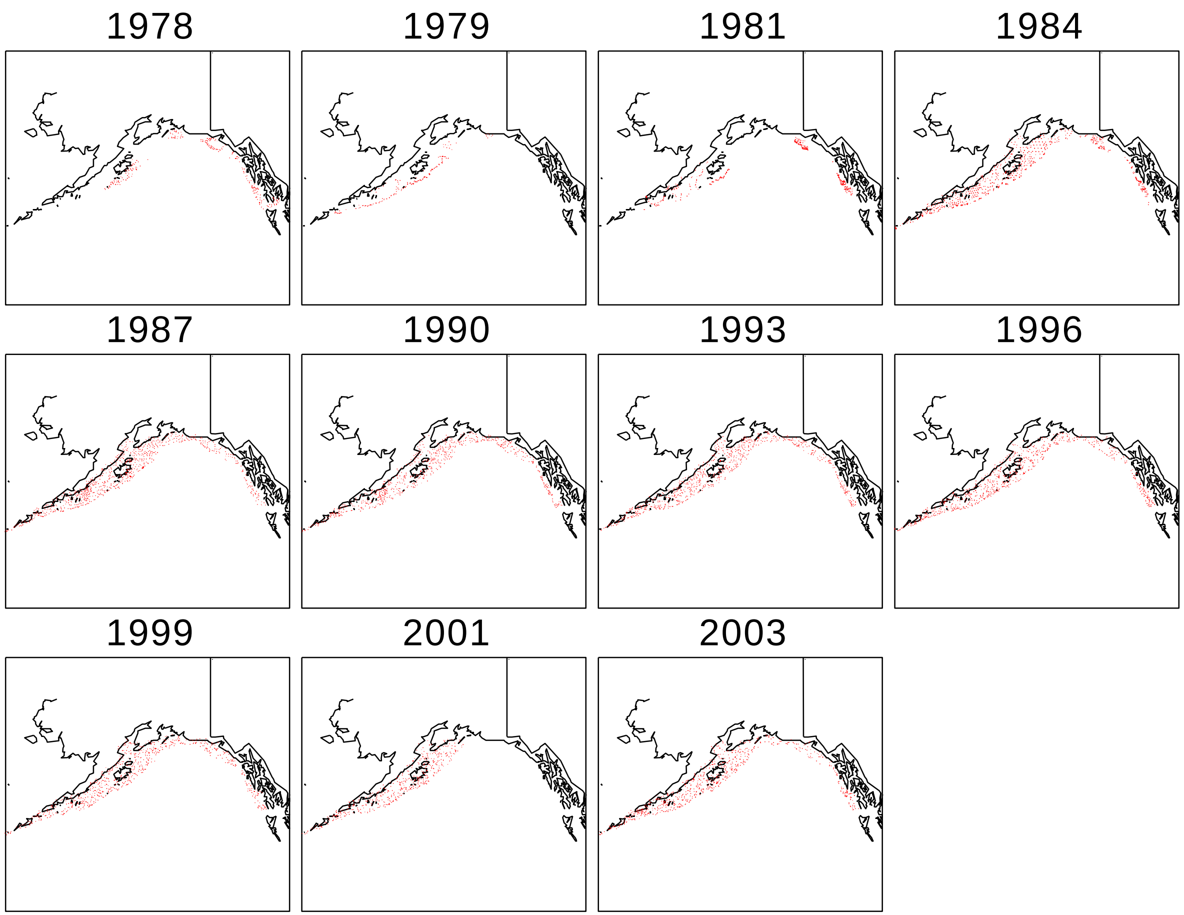
- Visualizing data
- Assessing distributions
- Outliers
- Standardization
- Transformations
- Correlations

Visualizing data

- Explore spatial / temporal structure
- Detect relationships / correlations
  - Scatterplot
  - Scatterplot matrix
  - Parallel coordinates
- Explore grouped data (differences among groups)
  - Co-plots / Trellis graphics / ggplot 'aesthetics' & facets
- Assess distribution
  - Histograms, q-q plots, dotplots, boxplots
- Assess dependence
  - Serial correlation
  - Spatial correlation

Spatial structure

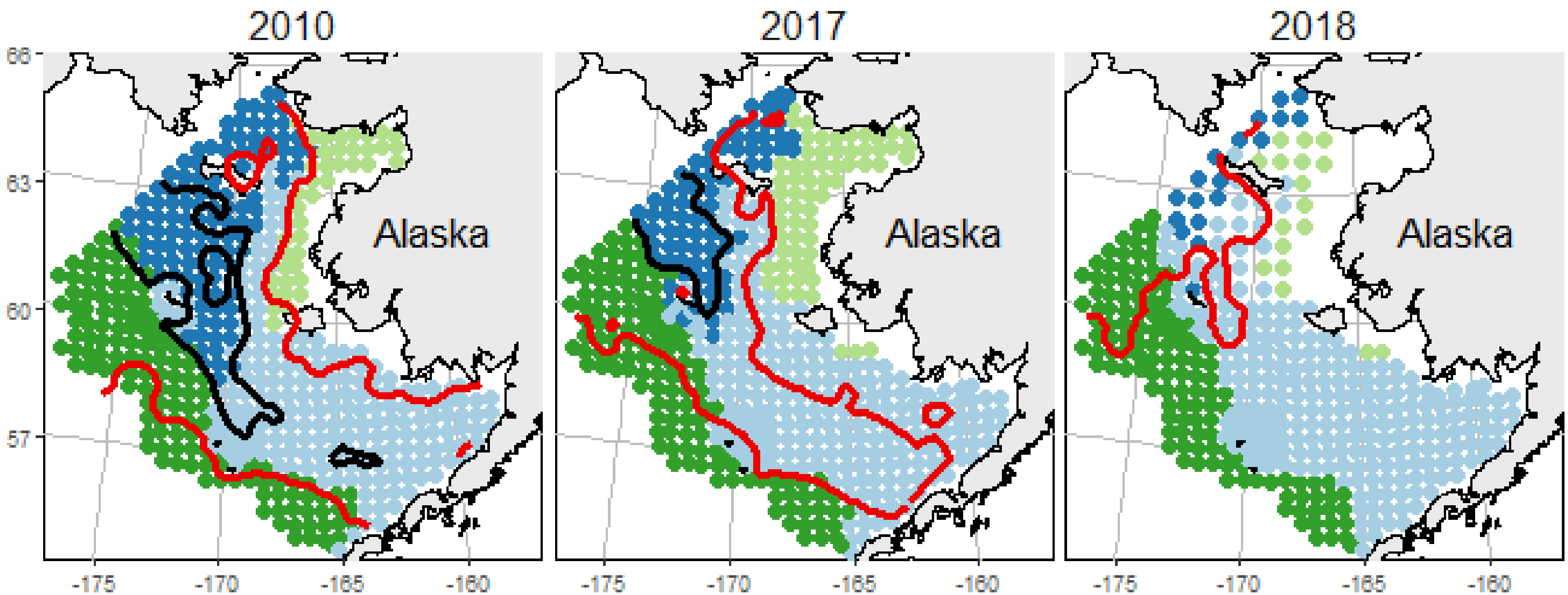
Quick maps for exploring / mapping data locations & attributes



Spatial data are ubiquitous in ecology & environmental science
→ GIS useful but not necessary

Spatial structure

…or fancier maps with appropriate projections

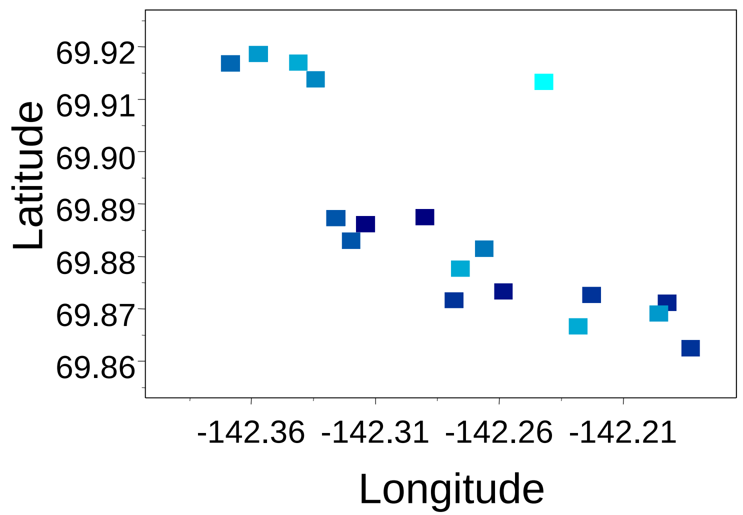


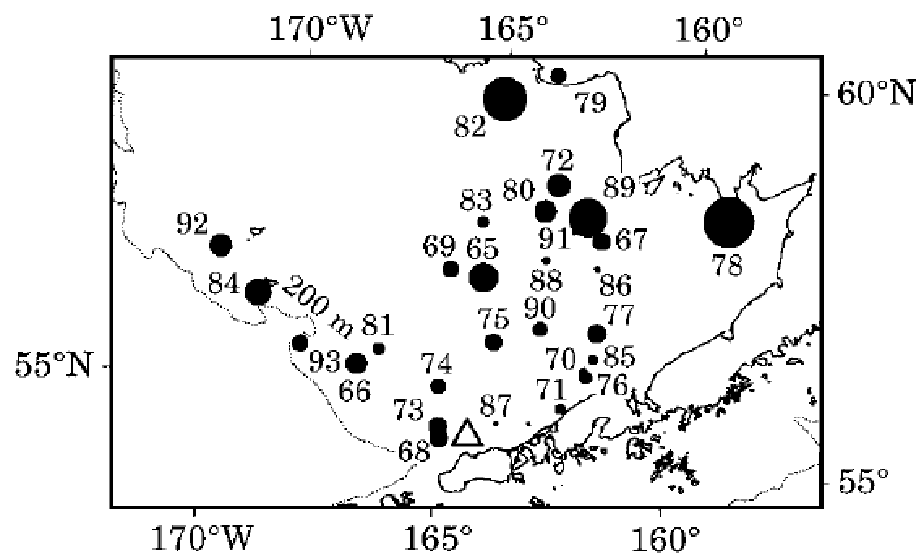
→ Learn how to quickly and efficiently map your data
→ Carefully design graphs and maps for publication

Spatial patterns

- Location
- Magnitude
  - symbol type
  - size
  - color

Iron (Beaufort Sea sediments)





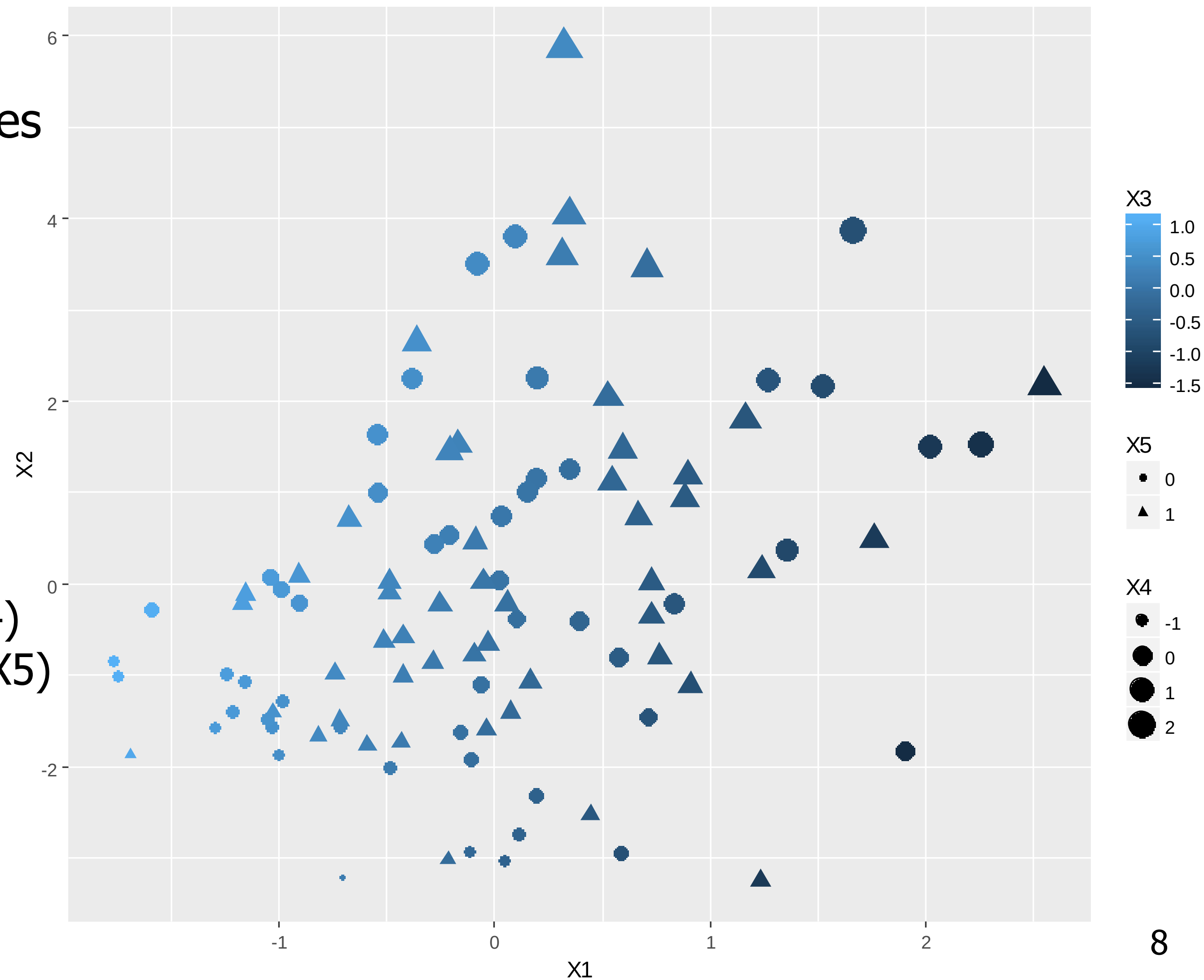
Displaying multiple attributes using 'aesthetics' (ggplot)

MSL FISH 604

Displaying five variables in two dimensions:

- x-axis (X1)
- y-axis (X2)
- color (X3)
- size (X4)
- shape (X5)

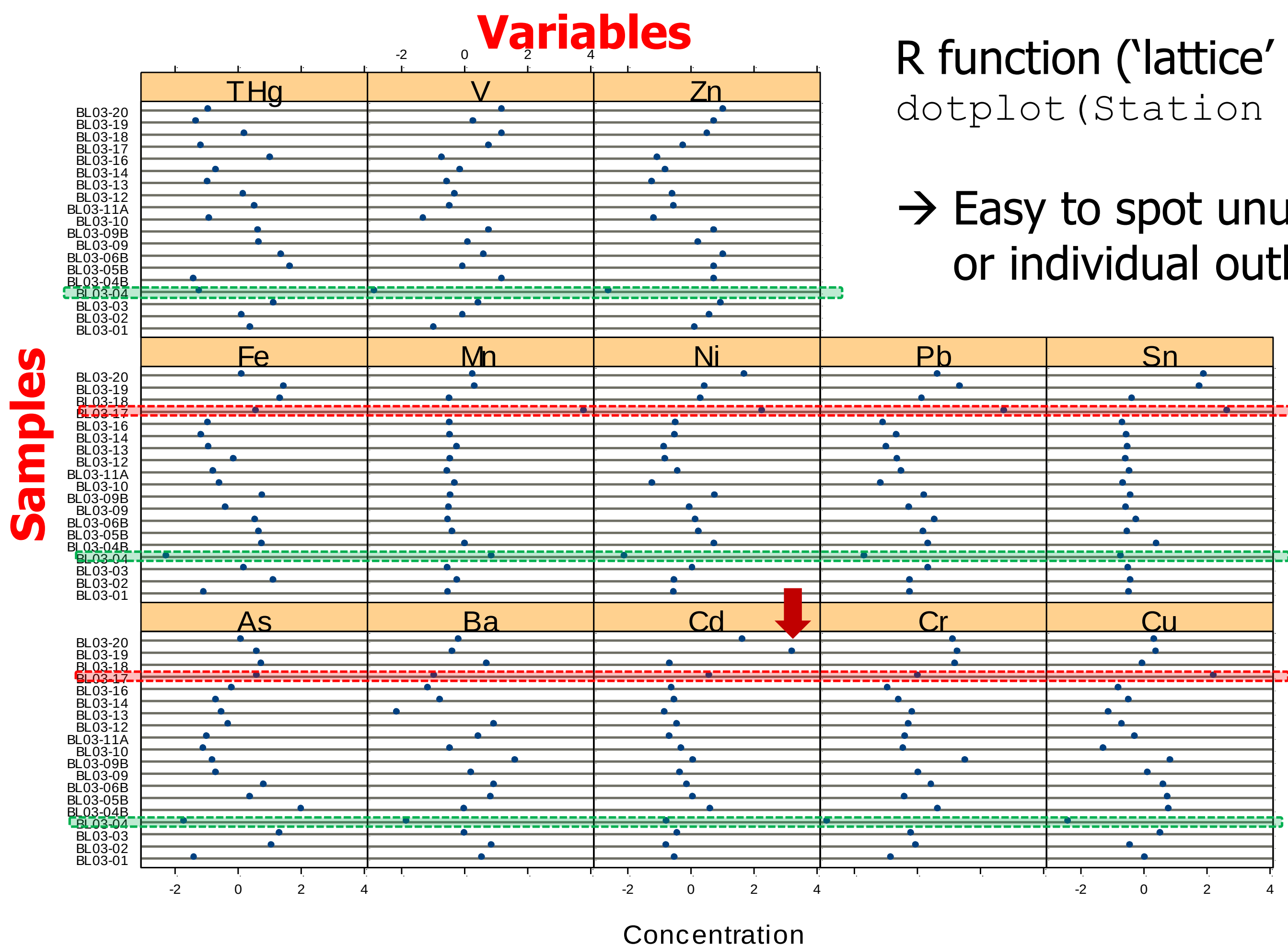
Mix of continuous (X1-X4) & categorical variables (X5)



Dotplots: Show each datapoint (whenever possible)!

MSL FISH 604

Metal concentrations, Beaufort Lagoon

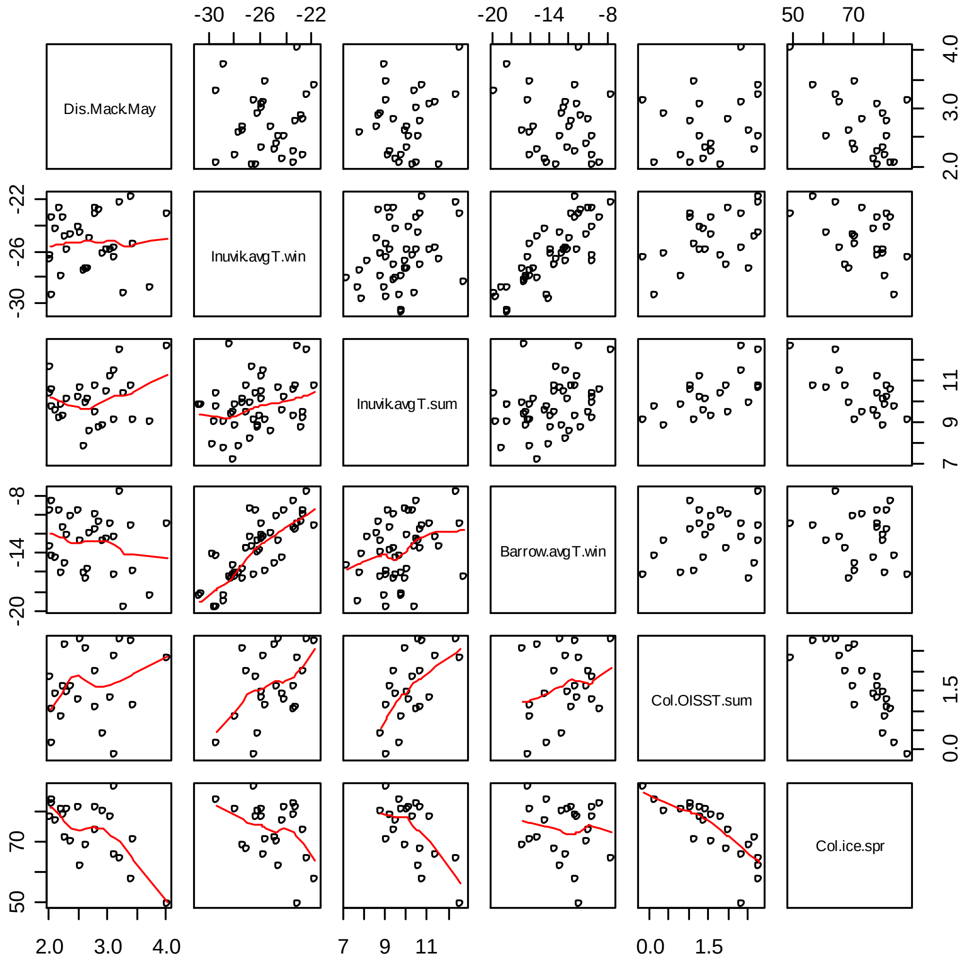


R function ('lattice' package):

```
dotplot(Station ~ Conc | Metal)
```

→ Easy to spot unusual samples or individual outliers

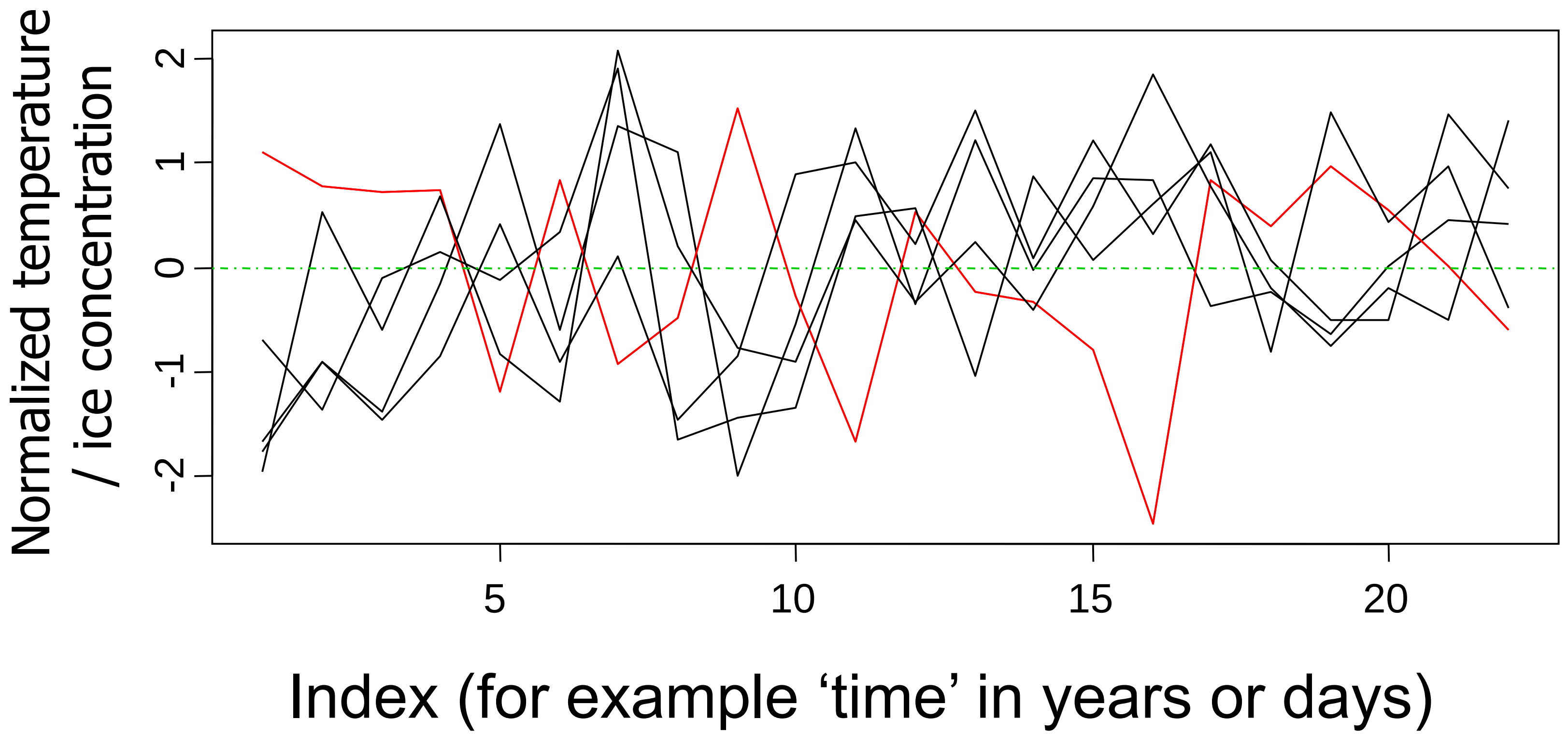
Relationships among variables: Scatterplot matrix



<u>R function:</u>

```
pairs(data, lower.panel=panel.smooth)
```

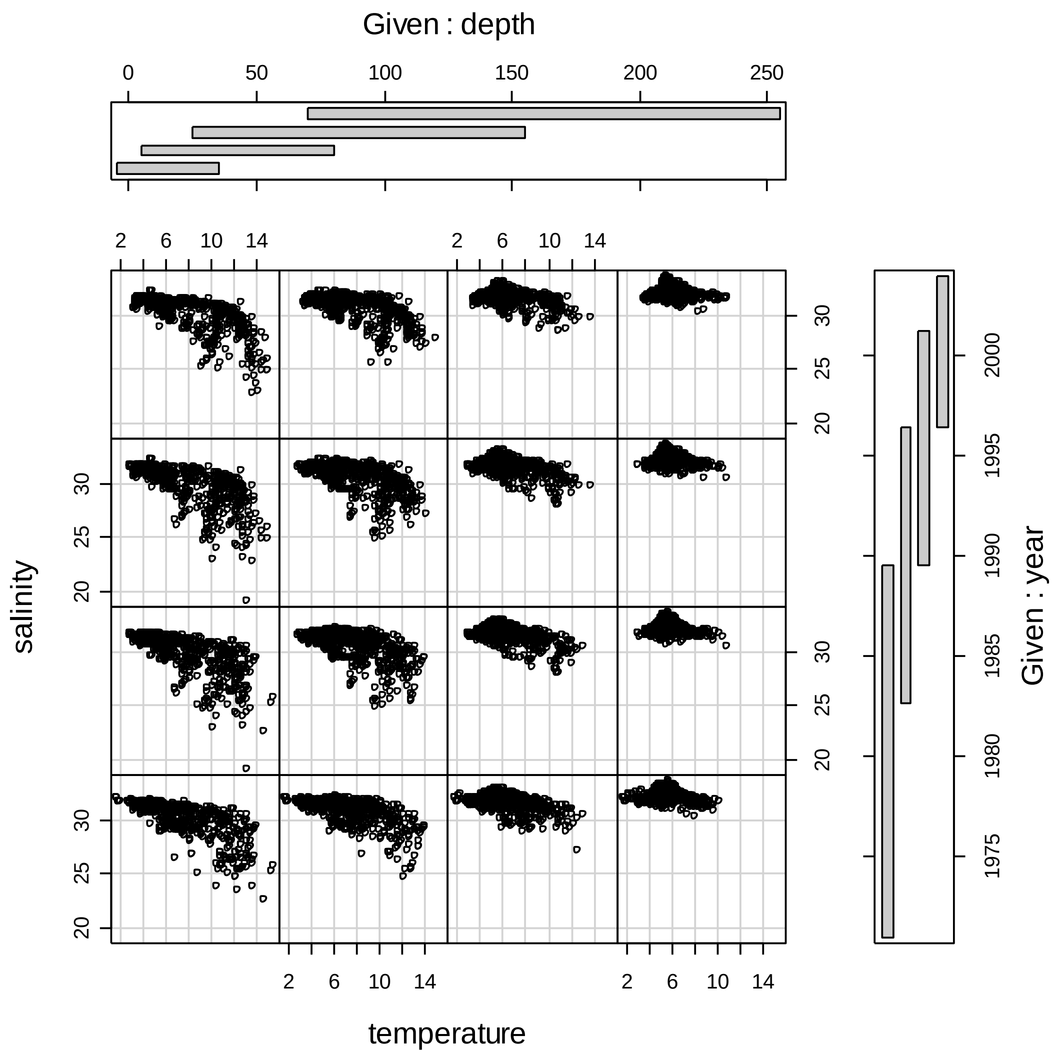
Relationships among variables: Parallel coordinates



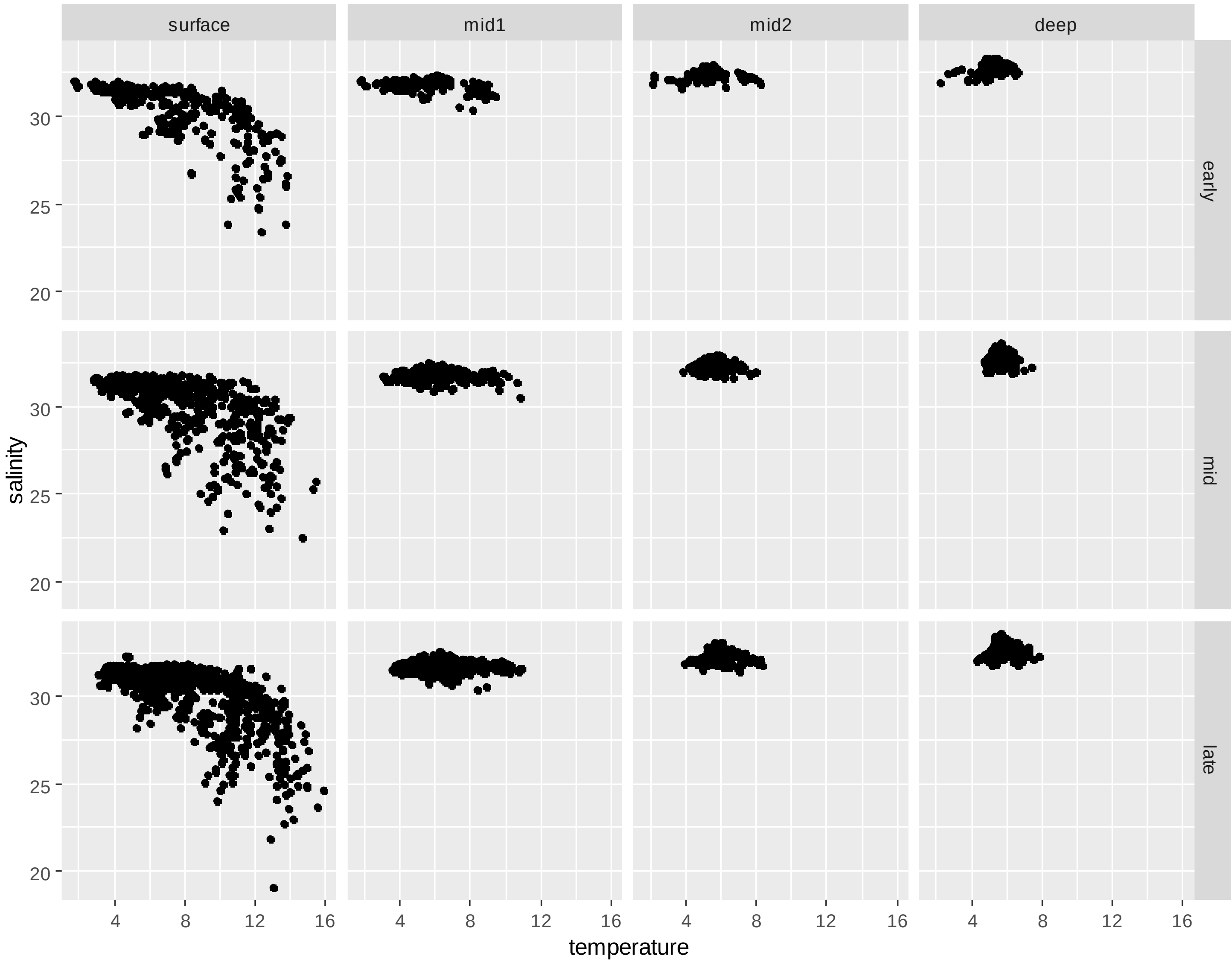
Multivariate exploration (e.g. "brushing") (e.g. using GGOBI)

Exploring grouped data: Co-plots

- GAK 1 data
- T-S plots by
  - depth and
  - time period

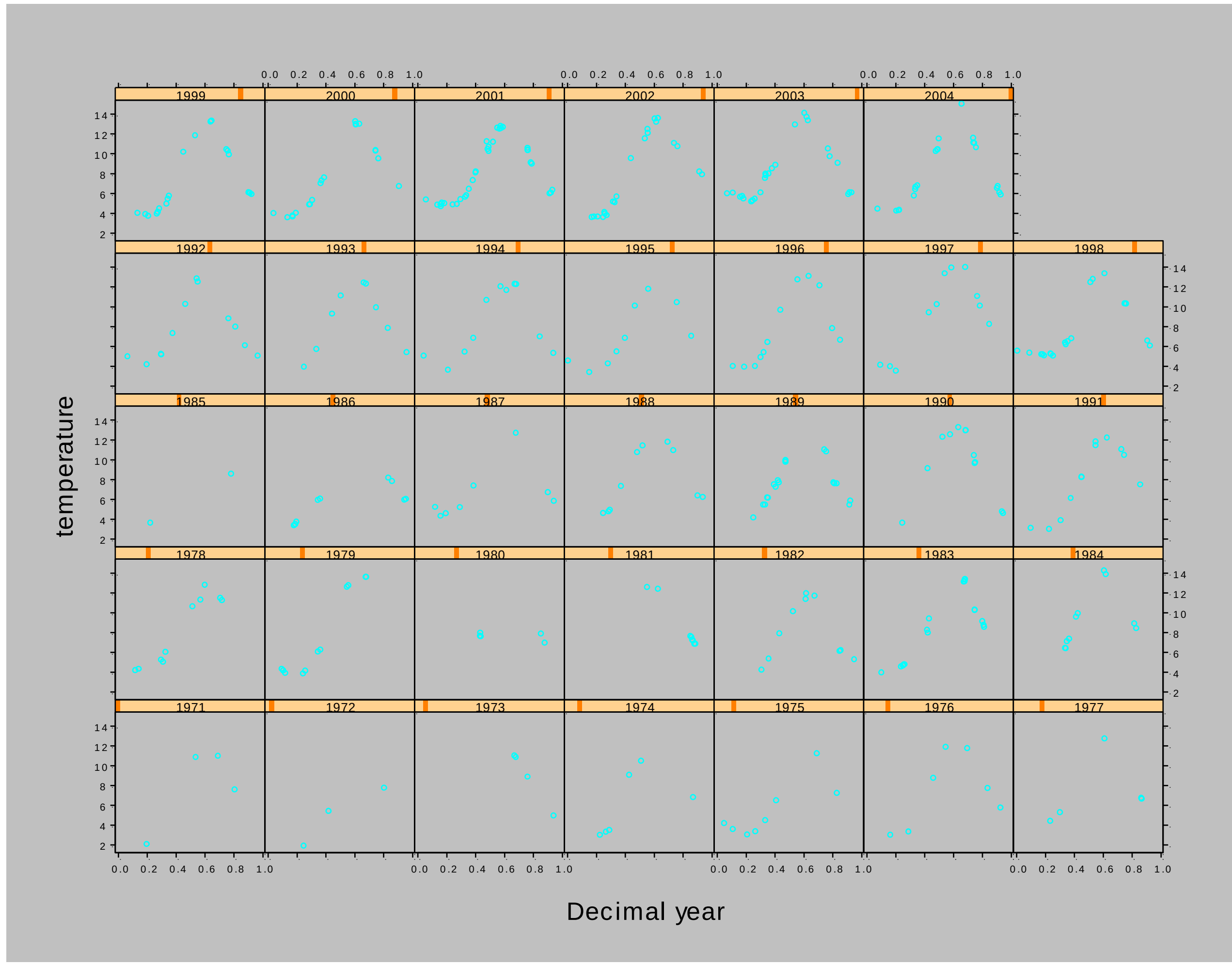


Exploring grouped data: ggplot (non-overlapping groups)



Exploring grouped data: Trellis graphics

Ocean temperature data collected irregularly over multiple years

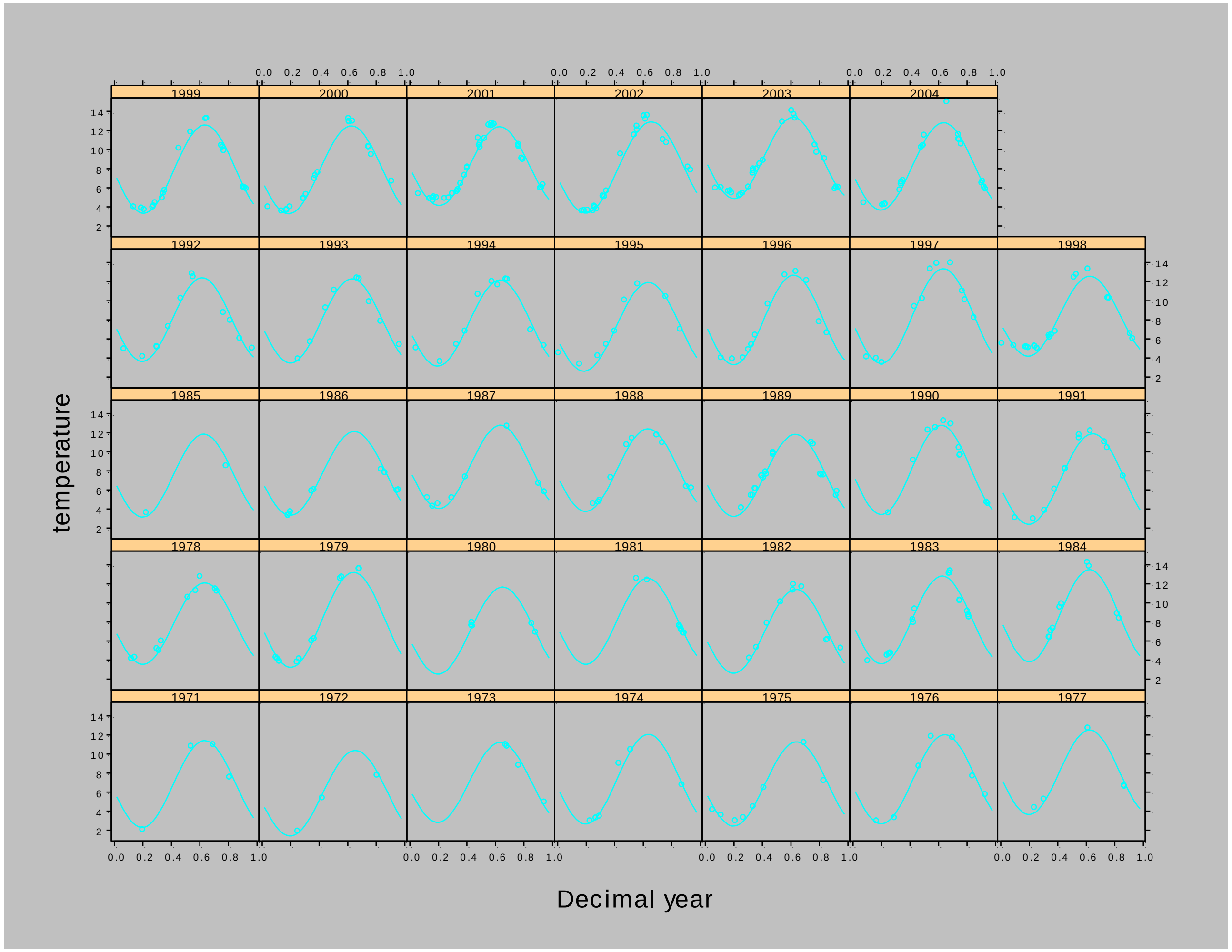


R code:
xyplot(temperature ~ dec.year | Year, data = GAK1, subset = depth==0)

Visualizing model fits

Fitted non-linear mixed-effects model

→ Learn how to visualize simple and complex model fits



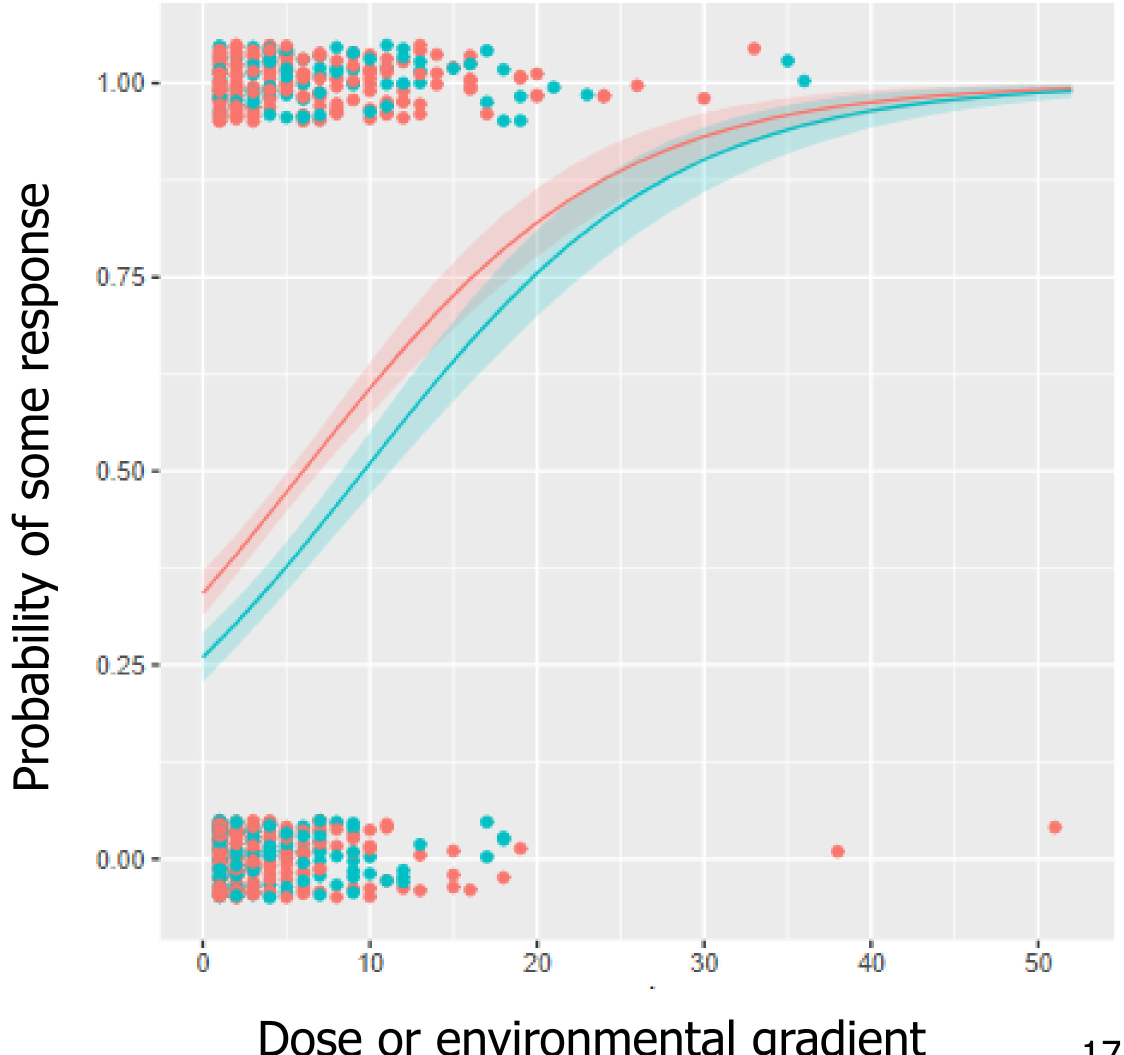
R code (requires library nlme):

```
plot(augPred(fitted.object))
```

Visualizing model fits

`ggplot()` is great for visualizing model fits for many "standard" models, but be aware of what you are doing!

Example: Logistic regression fit with two groups



MSL FISH 604

Assessing distributions

Visually

- histograms / density plots / boxplots
- quantile-quantile plots (q-q plots)

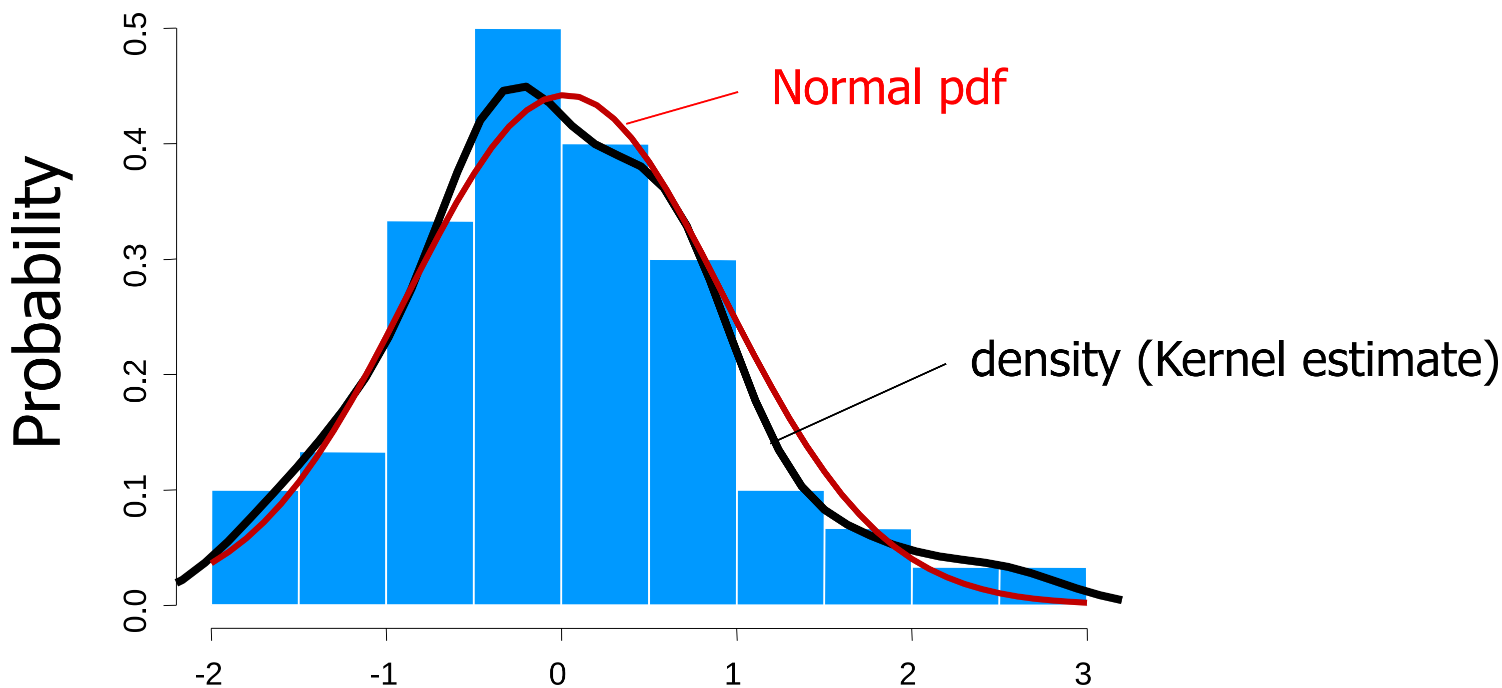
Statistical

- tests for normality
- testing against any distribution
  - Kolmogorov-Smirnov test

R functions

```
hist(); histogram(); density(); boxplot()
qqplot(); qqnorm(); qqline()
shapiro.test(); ks.test()
```

Histogram with density



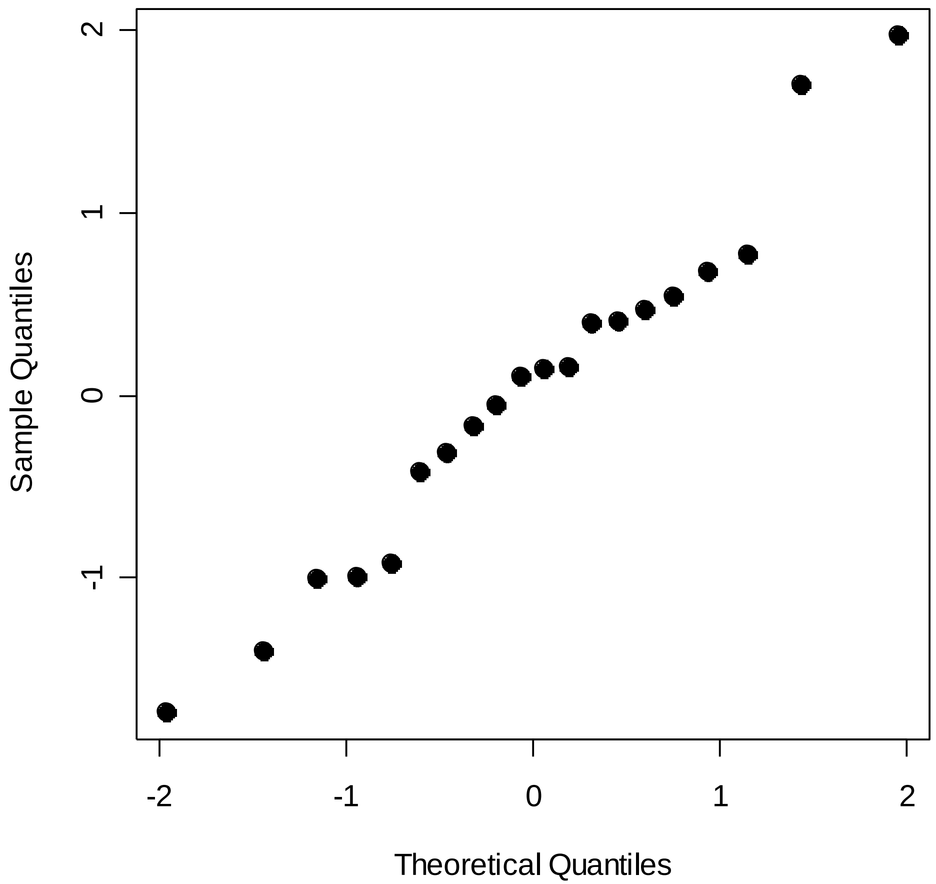
Algorithms to determine bin size:

Scott (1979): $$\text{"optimal" bin width} = 3.49 \cdot \sigma \cdot N^{-1/3}$$

Freedman and Diaconis (1981): $$\text{"robust" bin width} = 2 \cdot IQR \cdot N^{-1/3}$$

quantile-quantile plots

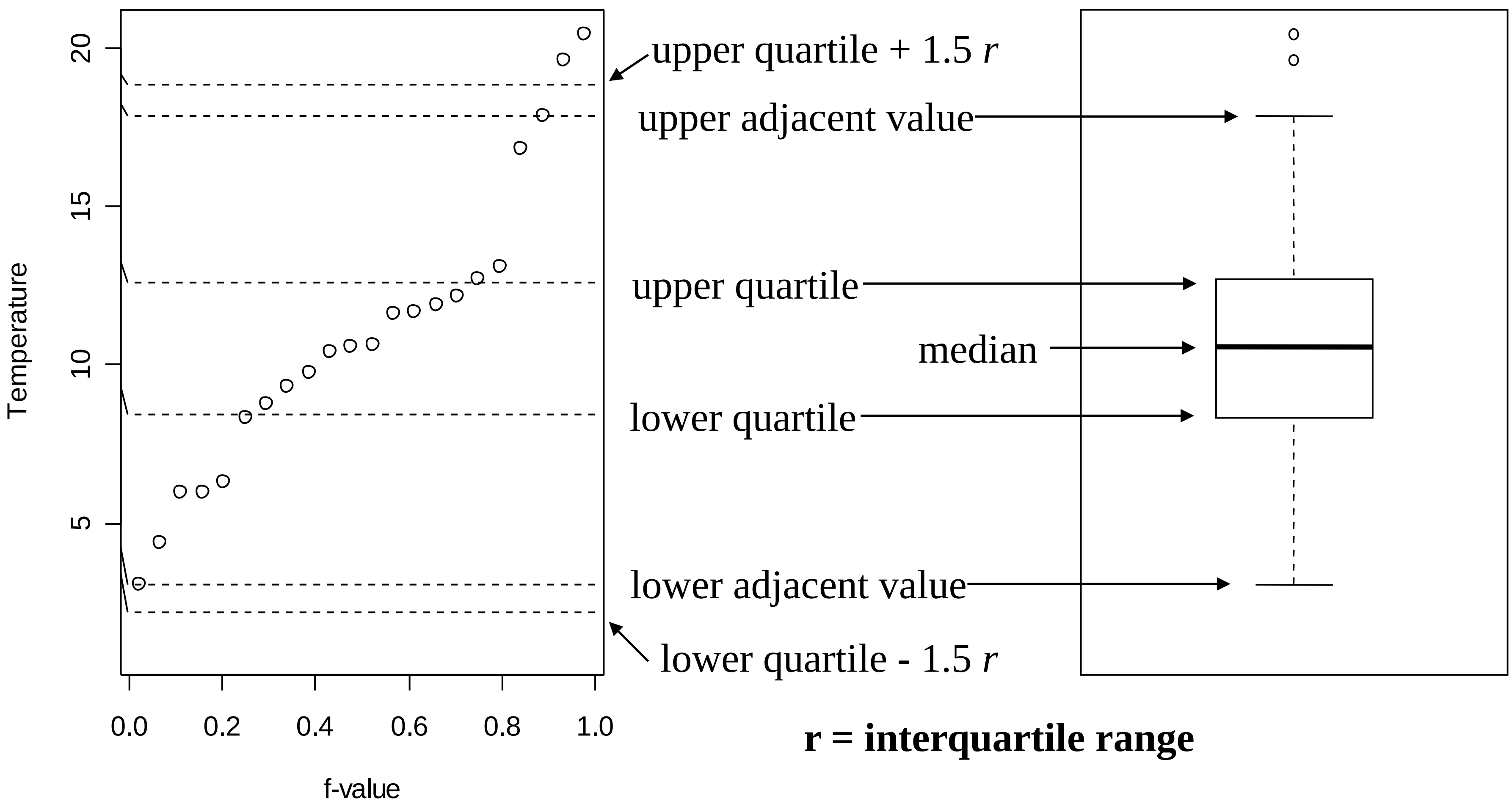
Normal q-q plot



- Sample quantiles
  - Sort data: $x_{(i)}$
  - $f_i = (i\text{-}0.5)/n$
  - $x_{(i)}$ is the f_i quantile of the data
- Theoretical quantiles
  - $q_{\mu,\sigma}(f_i)$ is normal f_i quantile
  - $q_{\mu,\sigma}(f) = \mu + \sigma\, q_{0,1}(f)$
- Normal q-q plot
  - $x_{(i)}$ vs. $q_{0,1}(f_i)$

→ usually sufficient as a diagnostic check of normality assumption!

Boxplot



r = interquartile range

→ Extremely effective for visualizing / comparing many distributions

→ Be aware of <u>what</u> boxes & whiskers show and <u>explain it</u> in figure captions

Tests for normality

- Shapiro-Wilk test
- Kolmogorov-Smirnov test
- Many others!!!

R functions

```
shapiro.test(); ks.test()
```

→ Remember that we only care about normality of residuals from a linear model, NOT normality of the response variable (or the independent variables)

→ Residuals from other models (e.g. GLM) often transformed to ‘approximate’ normality for visual assessment only

Shapiro-Wilk test

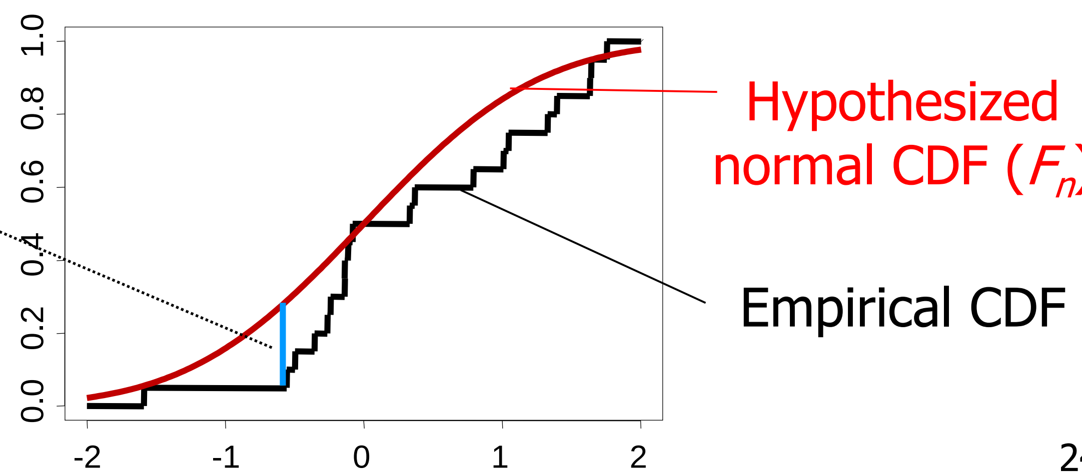
- Tests the null hypothesis that a sample $x_1, \ldots, x_n$ came from a normally distributed population
- Test statistic: $$W = \frac{\left(\sum_{i=1}^{n} a_i x_{(i)}\right)^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

  where $x_{(i)}$ is i^{th} order statistic
  a_i are constants that depend on expected values and covariance of order statistics
- Null hypothesis is rejected if W is too small

Kolmogorov-Smirnov test

- Null hypoptheses:
  - Two samples have the same underlying probability distributions
  - A single sample comes from a hypothesized distribution (e.g. normal)
- Test is based on maximum distance between two cumulative distribution functions:

$$D = \max\left(F_n - \frac{i-1}{N}, \frac{i}{N} - F_n \right)$$



MSL FISH 604

Reading assignment

- Zuur et al. (2007). **Analyzing Ecological Data**. Springer. *Chapter 4: Exploration* (see pdf posted on Canvas)

MSL FISH 604

Further reading

Graphical analysis

- Cleveland, W.S., 1993. Visualizing Data. AT&T Bell Laboratories, Murray Hill, NJ.
- Wilkinson, L. 1999. The Grammar of Graphics. Springer, New York
- Yau, Nathan 2013. Data Points: Visualization that means something. Wiley.

Multivariate data

- Cook, D., Swayne, D.F., and Buja, A. 2007. Interactive and Dynamic Graphics for Data Analysis: With R and GGobi. Springer, New York.

General exploratory analyses

- Zar, J.H., 1984. Biostatistical Analysis. Prentice-Hall, Englewood Cliffs, NJ.
- Barnett V, 2004. Environmental Statistics: methods and applications, John Wiley & Sons, Chichester, England.