Case Study H
Classification of Italian Olive Oils

Background

- The definition of olive oil quality (extra virgin, virgin) is expressed as threshold on acidity measured as the proportion of “oleic” fatty acid
- Composition of fatty acids in olive oils can be measured with the aim of identifying the origin of the olive oil
- Here measurements were taken in 1983 from Italian olive oils originating from different regions and areas

Goals of Study

- Distinguish regions of origin from fatty acid measurements (classification task)
- Explore data structures that may appear in the data and influence potential models

Description of Data

- Region
- Area

fatty acid measurements (scaled to 0-100 range each):

- palmitic
- palmitoleic
- stearic
- oleic
- linoleic
- linolenic
- arachidic
- eicosenoic

Analysis

- First look
  - SPLOM and PCP of measurements
  - Barcharts of Area and Region (which are used for brushing in most subsequent tasks)
- Anomalies and structure
  - scatterplot of Linoleic vs Eicosenoic
  - histogram of Arachidic
  - scatterplots Arachidic vs Linolenic, Stearic vs Palmitoleic
- Classification
  - scatterplots Linoleic vs Oleic, Linoleic vs Eicosenoic
  - use cropping in PCP of all measurements to look at region and area subsets plus color brushing

Further Analysis

- Models can be used for the actual classification task
- Tree models deliver partitioning that can be directly visualized in the data space

R code

```
> # load the rpart-library
> library(rpart)
> # create the tree model (exclude Area!)
> t1 <- rpart(Region ˜ . , data = olives[,1:9])
> # plot the tree
> plot(t1); text(t1)
> # create confusion matrix
> tab <- table(predict(t1, type="class"), olives$Region)
> # everything not on the diagonal is an error
> sum(tab) - sum(diag(tab))
[1] 42
```